SIAM J. CONTROL AND OPTIMIZATION
Vol. 33, No. 1, pp. 1–28, January 1995

© 1995 Society for Industrial and Applied Mathematics
001

EXPONENTIAL STABILITY OF COUPLED BEAMS WITH DISSIPATIVE JOINTS: A FREQUENCY DOMAIN APPROACH*

RICHARD REBARBER[†]

Abstract. Two examples of coupled Euler–Bernoulli beams with a dissipative joint are considered. The joint placements that lead to exponential stability for these systems are characterized. The technique used shows input-output stability of a related controlled, observed system, and then shows that in these examples, input-output stability implies exponential stability. In the first example, the energy dissipation arises from a discontinuity in the shear at the joint. In the second example, the energy dissipation arises from a discontinuity in the bending moment at the joint. The analysis of this system involves a complete spectrum analysis of the zero dynamics of the associated controlled, observed system.

Key words. exponential stability, beam equation, transfer function, input-output stability, dissipative joints

AMS subject classifications. 93, 93C20, 93D15

1. Introduction.

In this paper we discuss the exponential stability of two beams coupled in a dissipative joint. The problem addressed here has been posed by Chen et al. [3], and systems of this form have been discussed in detail by Chen et al. [5]. In [5] asymptotic formulas for the eigenvalues of such systems are given, and the eigenvalue placement suggests that some of these systems are exponentially stable. Further spectrum analysis for coupled beams has been done by Conrad [6]. It is well known, however, that distributed parameter systems do not necessarily have spectrum-determined growth. There are examples in Zabczyk [25] and Huang [11], among others, where the resolvent set of a generator A of a C_0-semigroup $S(t)$ contains $\mathbf{C}_\alpha^+ = \{z \in \mathbf{C} \,|\, \mathrm{Re}(z) > \alpha\}$ for $\alpha < 0$, but $S(t)$ has positive exponential growth. Although there are several classes of systems that do have spectrum-determined growth, there is no reason to expect that the boundary feedback systems considered in [5] have this property.

A useful way to show that $S(t)$ is exponentially stable has been given in [11] (or independently, Prüss [14]). If $S(t)$ is a bounded semigroup, if $R(\lambda, A)$, and if the resolvent of A satisfies

$$\sup_{\omega \in \mathbf{R}} \|R(i\omega, A)\| < M \tag{1.1}$$

for some M, then $S(t)$ is exponentially stable. This condition has been used to prove exponential stability for mechanical systems by Chen et al. [4], Liu [12], and Liu, Huang, and Chen [13].

For the systems under consideration in this paper, we show that the resolvent is bounded on the imaginary axis by first showing that a transfer function for a related controlled, observed system is bounded on the imaginary axis. This shows that the related system is *input-output stable*, and we then apply a result in Rebarber [16] to show that in this case input-output stability implies exponential stability. We describe this approach in detail in §2.

We assume that one of the beams has spatial extent from $s = 0$ to $s = s_1$ and that the other beam has extent from $s = s_1$ to $s = 1$. In [5] the eigenvalues are only computed for the joint in the middle of the span, but the results in this paper are for general s_1. We assume that both are uniform Euler–Bernoulli beams with the same mass density per unit length m and the same flexural rigidity EI. We normalize so that $EI/m = 1$. This last assumption is not necessary for our work, but it makes the calculations a bit simpler. Let $w(s, t)$ be the displacement of

* Received by the editors November 16, 1992; accepted for publication (in revised form) July 16, 1993. This work was partially funded by National Science Foundation grant DMS-9206986, the Faculty of Applied Mathematics of the University of Twente, the Netherlands, and the University of Exeter, United Kingdom.

[†] Department of Mathematics and Statistics, University of Nebraska-Lincoln, Lincoln, Nebraska 68502.

the coupled beams at position $s \in [0,1]$ and time $t \in [0,\infty]$. The notation $\dot{w}(t,s)$ denotes the derivative of $w(s,t)$ with respect to time, and D denotes the spatial differentiation operator. Then w satisfies the following Euler–Bernoulli equation in both beams:

$$\ddot{w}(s,t) + D^4 w(s,t) = 0, \qquad s \in (0,s_1) \cup (s_1,1). \tag{1.2}$$

The energy of this system is given by

$$E(w(\cdot,t)) = \int_0^1 \{\dot{w}^2(s,t) + (D^2 w(s,t))^2\} ds.$$

We consider two sets of boundary and joint conditions. In both cases the energy of the system can be shown to dissipate, that is $(d/dt)E(t) \leq 0$; this is done in [5] when the beams are of equal length, and the argument can be easily modified for general s_1. The goal of this paper is to characterize those joint placements for which the energy decay is in fact exponential. We say that the system (1.2) with various boundary conditions is exponentially stable if there exists $M > 0$ and $\alpha > 0$ such that $|E(w(\cdot,t))| \leq Me^{-\alpha t}|E(w(\cdot,0))|$.

Let $w(s_1^-,t)$ be the limit of $w(s,t)$ as $x \longrightarrow s_1$ from the left and $w(s_1^+,t)$ be the limit of $w(s,t)$ as $x \longrightarrow s_1$ from the right. The two cases we consider are as follows.

Case 1. In this case the end at $s = 0$ is simply supported, and at $s = 1$ there is a shear hinge end. The dissipative joint condition at s_1 is a rigid support joint where the discontinuity in the shear is proportional to the velocity at s_1. See [3] for a discussion of the joint and end conditions for coupled beams. The boundary and joint conditions are then given by

$$\begin{aligned} w(0,t) &= D^2 w(0,t) = 0, \\ Dw(1,t) &= D^3 w(1,t) = 0, \\ w(s_1^-,t) &= w(s_1^+,t), \\ Dw(s_1^-,t) &= Dw(s_1^+,t), \\ D^2 w(s_1^-,t) &= D^2 w(s_1^+,t), \end{aligned} \tag{1.3}$$

$$D^3 w(s_1^+,t) - D^3 w(s_1^-,t) = -k\dot{w}(s_1,t), \tag{1.4}$$

where $k > 0$. In (1.4) we are assuming that $\dot{w}(s_1^-,t) = \dot{w}(s_1^+,t)$, which follows from the third equation in (1.3) when $\dot{w}(\cdot,t)$ is in $H^1[0,1]$.

The following theorem states our conclusions about the exponential stability of (1.2), (1.3), (1.4).

THEOREM 1.1. *The system described by the coupled beam equations* (1.2), (1.3), (1.4) *is exponentially stable if and only if s_1 is a rational number with coprime factorization*

$$s_1 = \frac{p}{q}, \quad \textit{where } p \textit{ is odd.} \tag{1.5}$$

Remark 1.2. The proof of this theorem, which is given in §2, is easily modified if the end conditions (the first two equations in (1.3)) are changed. However, the proof of Theorem 1.1 is very dependent on the form of the feedback (1.4). This is because the system in Case 1 is *regular* in the sense given in Weiss [22]–[24]. This will be made precise in §2. The system in Case 2 below is quite a bit more difficult to handle, because the natural associated controlled, observed system is not regular, or even well posed.

Case 2. Both ends are simply supported in this case. The dissipative joint condition at s_1 is an angle guide (see [3]), where the discontinuity in the bending moment is proportional to the angular velocity at s_1. The boundary and joint conditions are then given by

$$\begin{aligned} w(0,t) &= D^2w(0,t) = 0,\\ w(1,t) &= D^2w(1,t) = 0,\\ w(s_1^-,t) &= w(s_1^+,t),\\ Dw(s_1^-,t) &= Dw(s_1^+,t),\\ D^3w(s_1^-,t) &= D^3w(s_1^+,t), \end{aligned} \tag{1.6}$$

$$D^2w(s_1^-,t) - D^2w(s_1^+,t) = -\hat{k}\, D\dot{w}(s_1,t), \tag{1.7}$$

where $\hat{k} > 0$. In (1.7) we are assuming that $D\dot{w}(s_1^-,t) = D\dot{w}(s_1^+,t)$, which follows from (1.6) if $\dot{w}(\cdot,t)$ is in $H^2[0,1]$.

The following theorem states our conclusions about the exponential stability of (1.2), (1.6), (1.7).

THEOREM 1.3. *The system described by the coupled beam equations* (1.2), (1.6), (1.7) *is exponentially stable if and only if* s_1 *is a rational number with coprime factorization*

$$s_1 = \frac{p}{q}, \quad \text{where } q \text{ is odd.} \tag{1.8}$$

Conditions (1.5) and (1.8) indicate that exponential stability of coupled beams is highly nonrobust with respect to the placement of the joint. Conditions analogous to (1.5) and (1.8) for coupled wave equations can be found in Theorems 4.3 and 4.4 in [12]. Section 2 is devoted to the proof of Theorem 1.1, whereas Theorem 1.3 is proved in §3.

Remark 1.4. To prove Theorem 1.3 we follow the same general approach as in the proof of Theorem 1.1. However, before we apply this approach, we need to do a careful eigenvalue-eigenvector analysis of a "nonstandard" generator, defined in §3. This approach should work if the end conditions (the first two equations in (1.6)) are changed, but the eigenvalue-eigenvector analysis might be more difficult.

2. Proof of Theorem 1.1. We need to put (1.2), (1.3), (1.4) into a standard state-space form. It is important to be very precise about the state-space formulation, because the proof of exponential stability is systems theoretic in nature, rather than being motivated by classical partial differential equations techniques such as multiplier methods. For any $n \in \mathbf{Z}^+$, let $H^n[0,1]$ be the space of all functions f such that $f, Df, D^2f, \ldots, D^{n-1}f$ are absolutely continuous and in $L^2[0,1]$, and D^nf is in $L^2[0,1]$. Let the state space be

$$X = \{[x_1,x_2]^T \in H^2[0,1] \oplus L^2[0,1] \mid x_1(0) = Dx_1(1) = 0\}.$$

Then X is a Hilbert space with the inner product

$$\langle [x_1,x_2]^T, [y_1,y_2]^T \rangle = \int_0^1 \{D^2x_1(s)\overline{D^2y_1(s)} + x_2(s)\overline{y_2(s)}\}ds. \tag{2.1}$$

Let the operator A be defined on X by

$$A\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} O & I \\ -D^4 & 0 \end{bmatrix}\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \tag{2.2}$$

with domain

$$\begin{aligned} \mathcal{D}(A) = \{[x_1,x_2]^T \in H^4[0,1] \oplus H^2[0,1] \mid x_1(0) = Dx_1(1)\\ = D^2x_1(0) = D^3x_1(1) = x_2(0) = Dx_2(1) = 0\}. \end{aligned}$$

For any $k \in \mathbf{Z}^+$, let

$$\omega_k = -\pi/2 + \pi k, \quad \lambda_{\pm k} = \pm i\omega_k^2, \quad \phi_k(s) = \sin(\omega_k s). \tag{2.3}$$

Let the index set I henceforth deote $\{\dots, -n, -(n-1), \dots, -1, 1, \dots, n, n+1, \dots\}$. It is easy to see that $\{\lambda_k\}_{k\in I}$ are the eigenvalues of A with associated eigenvectors

$$\Phi_{\pm k} = \frac{1}{\sqrt{2}} \begin{bmatrix} \phi_k/\lambda_{\pm k} \\ \phi_k \end{bmatrix}.$$

It is also easy to see tht A^*, the adjoint of A, is equal to $-A$ (in other words A is skew adjoint) and that $\{\Phi_k\}_{k\in I}$ is an orthonormal basis of X.

We need to identify the appropriate input operator for the control system (1.2), (1.3),

$$D^3 w(s_1^+, t) - D^3 w(s_1^-, t) = u(t), \tag{2.4}$$

where $u \in L^2_{\mathrm{loc}}[0, \infty)$. We follow the approach introduced by Ho and Russell in [9]. To that end let $\hat{A}$ be the extension of A with domain

$$\begin{aligned} \mathcal{D}(\hat{A}) = \{&[x_1, x_2]^T \in X \mid x_1|_{[0,s_1)} \in H^4[0, s_1),\ x_2|_{(s_1,1]} \in H^4(s_1, 1], \\ &x_2 \in H^2[0,1],\ D^2x_1(0) = D^3x_1(1) = x_2(0) = Dx_2(1) = 0, \\ &D^2x_1(s_1^-) = D^2x_1(s_1^+)\}. \end{aligned} \tag{2.5}$$

Let $x = [x_1, x_2]^T \in \mathcal{D}(\hat{A})$ and $v = [v_1, v_2]^T \in \mathcal{D}(A^*) = \mathcal{D}(A)$. Using integration by parts we find that

$$\langle \hat{A}x, v\rangle = \langle x, -Av\rangle + [D^3x_1(s_1^+) - D^3x_1(s_1^-)]\overline{v_2(s_1)}. \tag{2.6}$$

Let

$$B = \begin{bmatrix} 0 \\ \delta(\cdot - s_1) \end{bmatrix}.$$

Then B is not an element of X, but it is easy to check that B is in the larger space $X_{-1} = (\mathcal{D}(A^*))'$, the dual space of $\mathcal{D}(A)$. The notation X_{-1} is used in Weiss [21] and is motivated by the fact that if X is a Hilbert space, X_{-1} is the closure of X in the norm $\|(\lambda - A)^{-1}x\|$ for any λ in the resolvent set of A.

We also need to define another extension of A. Let $\tilde{A} : X \to X_{-1}$ be defined by

$$\langle \tilde{A}x, v\rangle = \langle x, A^*v\rangle \quad \text{for all } v \in \mathcal{D}(A^*), \tag{2.7}$$

with $\mathcal{D}(\tilde{A}) = X$. Because $\tilde{A}$ is a standard extension for A, we refer to both A and its extension $\tilde{A}$ as A whenever no confusion will arise. It follows from (2.6) that for $x \in \mathcal{D}(\hat{A})$,

$$\hat{A}x = \tilde{A}x + [D^3x_1(s_1^+) - D^3x_1(s_1^-)]B \tag{2.8}$$

in X_{-1}.

If $w(s,t)$ satisfies (1.2), (1.3), and $x(t) = [w(\cdot,t), \dot{w}(\cdot,t)]^T$, then $\dot{x}(t) = \hat{A}x(t)$. Now suppose that we further insist that $w(s,t)$ satisfy (2.4). Then it follows from (2.8) that $\hat{A}x(t) = \tilde{A}x(t) + Bu(t)$ in X_{-1}. From the above discussion, we see that our state-space equation for (1.2), (1.3), (2.4) is

$$\dot{x}(t) = Ax(t) + Bu(t), \tag{2.9}$$

where we look for generalized solutions $x(t) \in X$ such that (2.0) is true in X_{-1}.

Consider now the observation for (1.2), (1.3), (2.4) given by

$$y(t) = \dot{w}(s_1, t). \tag{2.10}$$

The observed system (1.2), (1.3), (1.4) with $k = 0$, and (2.10) is then equivalent to $\dot{x}(t) = Ax(t)$ with

$$y(t) = B^* x(t), \tag{2.11}$$

where

$$B^* \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = x_2(s_1), \qquad \mathcal{D}(B^*) = \mathcal{D}(A).$$

As is typical for colocated sensing and control, the output operator is the dual of the input operator.

For some purposes it is convenient to represent (2.9), (2.11) as a diagonal system in the basis $\{\Phi_k\}_{k\in I}$. In this basis A can be represented as the diagonal matrix $\text{diag}[\lambda_k]_{k\in I}$, and B can be represented by the column vector $[b_k]^T_{k\in I}$, where

$$b_k = B^*\Phi_k = (1/\sqrt{2})\ \sin((\pi/2 + \pi k)s_1). \tag{2.12}$$

In a general setting, letting U be a Hilbert space, we say that $B \in \mathcal{B}(U, X_{-1})$ is an *admissible* input operator for $S(t)$ if there exists t_1, $\alpha > 0$ such that for every $u \in L^2[(0, t_1); U]$,

$$\left\| \int_0^{t_1} S(t-\tau)Bu(\tau)d\tau \right\|_X \leq \alpha \|u\|_{L^2[(0,t_1);U]}.$$

Because $\{b_k\}_{k\in I}$ is a bounded sequence, it is easy to use the Carleson measure theorem, as in Ho and Russell [9] or Weiss [20], to show that in our case $B \in \mathcal{B}(\mathbf{R}, X_{-1})$ is an admissible input operator for $S(t)$.

In a general setting, letting Y be a Hilbert space, we say that $C \in \mathcal{B}(\mathcal{D}(A), Y)$ is an admissible observation operator for $S(t)$ if there exists $\alpha > 0$ such that for every $x \in \mathcal{D}(A)$,

$$\|CS(t)x\|_Y \leq \alpha\|x\|_X.$$

In our case, letting $Y = \mathbf{R}$, it follows by duality that B^* is an admissible observation operator for $S(t)$.

We are of course interested in the feedback system (1.2), (1.3), (1.4). This system is equivalent to $\dot{x}(t) = A_k x(t)$, where A_k is given by the matrix (2.2) with domain

$$\mathcal{D}(A_k) = \{[x_1, x_2]^T \in \mathcal{D}(\hat{A}) \mid D^3x_1(s_1^+) - D^3x_1(s_1^-) = -kx_2(s_1)\}$$

(cf. (2.5)).

It follows from the Lumer–Phillips theorem and an easy generalization of [5, eq. (1.4)] that A_k is a dissipative operator that generates a C_0-semigroup of contractions $S_k(t)$ on X. If we can show that $S_k(t)$ is exponentially stable, then (1.2), (1.3), (1.4) is exponentially stable in the sense discussed in §1. For our subsequent development we need that $A_k = A - kBB_L^*$, where B_L^* is the Lebesque extension of B^* defined by

$$B_L^* x = \lim_{\tau\to 0} \frac{1}{\tau} \int_0^t B^*S(t)x\,dt, \tag{2.13}$$

defined on the set $\mathcal{D}(B_L^*)$ of all $x \in X$ such that the limit on the right side exists in X (see Weiss [19], [23]). The proof of this appears in Lemma 2.6.

We now describe the approach we use to show that A_k generates an exponentially stable semigroup. Because A_k is dissipative, it suffices to show that $R(\lambda, A_k)$ is bounded on the imaginary axis as in (1.1). To do this, we consider the controlled, observed system for A_k given by (1.2), (1.3), and (2.10) with

$$D^3w(s_1^+,t) - D^3w(s_1^-,t) + k\dot{w}(s_1,t) = u(t). \tag{2.14}$$

It is easy to show that (1.2), (1.3), (2.14) is equivalent to

$$\dot{x}(t) = A_k x(t) + Bu(t) \tag{2.15}$$

in the same way that (2.9) was obtained.

In a general setting we denote the controlled, observed system

$$\dot{x}(t) = Ax(t) + Bu(t),$$

$$y(t) = Cx(t)$$

by (C, A, B). As this system is written now, it may be that the solution $x(t)$ of the first equation is not in $\mathcal{D}(C)$, so we need to put conditions on the system so that the mapping from the input to the output is continuous.

DEFINITION 2.1. *Suppose A is the generator of a C_0-semigroup $S(t)$, $B \in \mathcal{B}(U, X_{-1})$, and $C \in \mathcal{B}(\mathcal{D}(A), U)$. Then (C, A, B) is regular if*

(1) *B is an admissible input operator for $S(t)$,*
(2) *C is an admissible observation operator for $S(t)$,*
(3) *the range of $R(\lambda, A)B$ is in $\mathcal{D}(C_L)$ (see (2.13)) for some $\lambda \in \rho(A)$,*
(4) *$C_L R(\lambda, A)B$ is bounded in the half plane $\mathbf{C}_\alpha^+$ for some $\alpha \in \mathbf{R}$.*

For a regular system it would be more precise to replace the observation $y(t) = Cx(t)$ with $y(t) = C_L x(t)$, because the solution of the controlled solution $x(t)$ is in the domain of C_L (see [23]).

Recall that the transfer function $\mathbf{H}$ of a controlled, observed system is that function for which $\hat{y}(\lambda) = \mathbf{H}(\lambda)\hat{u}(\lambda)$, where $\hat{y}$ denotes the Laplace transform of y. Let H_α^∞ be the space of all bounded and analytic complex functions on $\mathbf{C}_\alpha^+$. H_α^∞ is a Banach space with the supremum norm. For a regular system (C, A, B), it is shown in [23] that the transfer function is given by $\mathbf{H}(\lambda) = C_L R(\lambda, A)B$ and $\mathbf{H}(\lambda)$ has a strong limit of $D = 0$ as $\lambda \to \infty$ along the positive real axis.

This definition is a special case of the definition of regular system given in Weiss [23] because we are assuming here that the feedthrough operator D is 0. In this paper, if a system is given by the abstract triple (C, A, B), then we assume that the feedthrough is 0. However, when the system is given by a partial differential equation with boundary control and observation, if we wish to describe it by the abstract triple, we need to verify that the feedthrough $\lim_{\lambda \to +\infty, \lambda \in \mathbf{R}} \mathbf{H}(\lambda) = D$ is zero. Therefore, we do not describe the system (1.2), (1.3), (2.10), (2.14) as an abstract triple until we verify that the feedthrough is indeed 0.

We would like to consider the closed-loop system obtained by closing the loop in $\dot{x}(t) = Ax(t) + Bu(t)$, $y(t) = Cx(t)$ with $y(t) = Ju(t)$, where $J \in \mathcal{B}(U)$. When B and C are both unbounded, the closed-loop system is not necessarily well posed, and it is not guaranteed that the closed-loop operator $A + BJC$ generates a strongly continuous semigroup. Therefore, we need to be precise about when such a feedback leads to a well-posed system. In Weiss [24] the following definition of admissible feedback is introduced. The notation H_∞^∞ is used to denote the set $\cup_{\alpha \in R} H^{\infty}_{\alpha}$, or, more precisely, the vector space of equivalence classes of elements in this set, where two functions are in the same equivalence class if one is a restriction of the other.

DEFINITION 2.2. *$J \in \mathcal{B}(U)$ is an admissible feedback for* $\mathbf{H}$, *or equivalently, for (C, A, B), if $(I - J\mathbf{H})$ is invertible in H^{∞}_{∞}.*

It is shown in [24, Thm. 6.6] that if J is an admissible feedback for a regular system (C, A, B), then the closed-loop generator obtained by letting $u(t) = Jy(t)$ is $A + BJC_L$, interpreted as follows. For $x \in \mathcal{D}(C_L)$, $\tilde{A}x + BJC_Lx$ can be evaluated in X_{-1}. Then $\mathcal{D}(A + BJC_L) = \{x \in \mathcal{D}(C_L) \mid \tilde{A}x + BJC_Lx \in X\}$. For a related result, see Salamon [18, Thms. 4.2 and 4.3].

To apply Theorem 2.5, stated below, to our system we need to show that it is *stabilizable* and *detectable*, as defined in Rebarber [16]. These definitions, which we now give, are nonstandard, but are more general than the more standard definitions (see, for example, Curtain [7]); this is because the stabilizing operator F and the detecting operator K are allowed to be unbounded.

DEFINITION 2.3. *(A, B) is stabilizable if there exists $F \in \mathcal{B}(X_1, U)$ such that (F, A, B) is regular, the identity I is an admissible feedback operator for (F, A, B), and $A + BF_L$ is the generator of an exponentially stable semigroup on X.*

DEFINITION 2.4. *(C, A) is detectable if there exists $K \in \mathcal{B}(U, X_{-1})$ such that (C, A, K) is regular, I is an admissible feedback for (C, A, K), and $A + KC_L$ is the generator of an exponentially stable semigroup on X.*

The following theorem from [16] is used to verify that $R(\lambda, A_k)$ is bounded on the imaginary axis.

THEOREM 2.5. *Suppose that (C, A, B) is a regular system with transfer function* $\mathbf{H}(\lambda)$, *(A, B) is stabilizable, and (C, A) is detectable. If* $\mathbf{H}(\lambda)$ *is bounded on the imaginary axis, then so is $R(\lambda, A)$.*

A controlled, observed system is said to be input-output (or externally) stable if its transfer function is bounded on $\mathbf{C}^{+}_{0}$. Hence, Theorem 2.5 gives conditions under which input-output stability implies exponential stability. To apply this theorem we break the work into several lemmas. In Lemma 2.6 it is shown that $A_k = A - BB^*_L$, and that the feedthrough for the system (1.2), (1.3), (2.10), (2.14) is 0, so the system can be represented by (B^*_L, A_k, B), where B^*_L is interpreted as the restriction of B^*_L to the domain of A_k. It is also shown in Lemma 2.6 that (B^*_L, A_k, B) is a regular system. In Lemma 2.8 we show that the transfer function $\mathbf{H}_k(\lambda)$ for (B^*, A_k, B) is bounded on the imaginary axis. In Lemma 2.10 we show that (A_k, B) is stabilizable and (B^*, A_k) is detectable when condition (1.5) holds. Putting this together with Lemma 2.6 and Theorem 2.5 shows that (1.5) is a sufficient condition for A_k to generate an exponentially stable semigroup. In Lemma 2.12 it is shown that (1.5) is a necessary condition for A_k to generate an exponentially stable semigroup.

LEMMA 2.6. (1) *The feedthrough for the system* (1.2), (1.3), (2.10), (2.14) *is* 0.

(2) *(B^*_L, A_k, B) is regular.*

(3) *$A_k = A - kBB^*_L$.*

Proof. We first show that the system (1.2), (1.3), (2.4), (2.10) is regular. We have already shown that B and B^* are admissible for $S(t)$. We can obtain the formula for $B^*_L R(\lambda, A)B$ by using the matrix representations of A and B in the basis $\{\Phi_k\}_{k \in I}$. Because B can be represented by

$$B = \sum_{k \in I} b_k \Phi_k,$$

(c.f. (2.12)), $R(\lambda, A)B$ can be represented by

$$R(\lambda, A)B = \sum_{k \in I} \frac{b_k}{\lambda - \lambda_k} \Phi_k. \tag{2.16}$$

To check whether $R(\lambda, A)B$ is in the domain of B_L^*, note that the right side of (2.13) is, with $x = R(\lambda, A)B$,

$$B_L^* R(\lambda, A)B = \lim_{\tau \to 0} \frac{1}{\tau} \int_0^\tau \sum_{k \in I} \frac{e^{\lambda_k t} |b_k|^2}{\lambda - \lambda_k} \, dt = \sum_{k \in I} \frac{|b_k|^2}{\lambda - \lambda_k}, \tag{2.17}$$

where the second inequality requires simple applications of Fubini's theorem and the dominated convergence theorem. It is clear that this converges for any $\lambda \neq \lambda_k$.

The proof of the following result is straightfoward but very technical, and is found in the Appendix. □

THEOREM 2.7. *Suppose* $\lambda_{\pm k} = \pm i\omega_k^2$ *for* $k \in \mathbf{Z}^+$, *where* $\omega_k \geq 0$ *and satisfies, for some real numbers* m, a, *and* b,

$$mk + b \leq \omega_k \leq mk + a. \tag{2.18}$$

Let

$$\mathbf{H}(\lambda) = \sum_{k \in I} \frac{d_k}{\lambda - \lambda_k},$$

where $\{d_k\} \in l^\infty$. *Then* $\mathbf{H} \in H_\alpha^\infty$ *for every* $\alpha > 0$, *and*

$$\lim_{\alpha \to \infty} \|\mathbf{H}\|_{H_\alpha^\infty} = 0. \tag{2.19}$$

This shows that the system (1.2), (1.3), (2.4), (2.10) is regular in the sense of [23] (with the feedthrough undetermined so far) and regular in the sense of Definition 2.1 if we can show the feedthrough is 0. We can easily obtain a transcendental description of the transfer function $\mathbf{H}(\lambda)$ by taking Laplace transforms of the partial differential equation system. Let $\lambda = i\omega^2$. There are of course two values of ω that satisfy this for any $\lambda \in \overline{\mathbf{C}_0^+}$, so we restrict ω to $\{\mathrm{r}e^{i\theta} \mid r \geq 0,\ \theta \in [-\pi/2, 0]\}$. The transfer function can then be written in the form

$$\mathbf{H}(\lambda) = \frac{i}{2\omega} \left\{ \frac{\sinh(\omega s_1) \cosh \omega(s_1 - 1)}{\cosh \omega} - \frac{\sin(\omega s_1) \cos \omega(s_1 - 1)}{\cos \omega} \right\}. \tag{2.20}$$

It is easy to use this form to see that $\mathbf{H}(\lambda) \to 0$ as $\lambda \to \infty$ along the positive real axis. Therefore, we can refer to the system as (B^*, A, B), which is regular.

Theorem 3.13 in Weiss [24] states that if (2.19) is true, then any feedback $u(t) = Jy(t)$ with $J \in \mathcal{B}(U)$ leads to a well-posed closed-loop system with generator $A + BJB_L^*$. Furthermore, from [24, Thm. 4.6], the closed-loop system is regular with feedthrough 0. We refer to this system as $(B_L^*, A + BJB_L^*, B)$, where the first operator B_L^* is interpreted as B_L^* restricted to the domain of $A + BJB_L^*$.

The system (1.2), (1.3), (2.4), (2.10) is a "boundary control system" discussed in Salamon [18]. By part (iii) of Lemma 4.4 and part (i) of Corollary 4.5 in that paper, $A - kBB_L^* = A_k$. Therefore, letting $J = -kI$, we see that (B_L^*, A_k, B) is regular.

LEMMA 2.8. *The transfer function* $\mathbf{H}_k(\lambda)$ *for* (B_L^*, A_k, B) *is bounded on the imaginary axis.*

Proof. $\mathbf{H}_k$ is that function such that $\hat{y}(\lambda) = \mathbf{H}_k(\lambda)\hat{u}(\lambda)$ when y and u are related by (2.11), (2.15), which is equivalent to

$$\dot{x}(t) = Ax(t) + Bv(t),$$

$$y(t) = B^*x(t),$$

$$v(t) = -ky(t) + u(t).$$

Therefore, $\hat{y}(\lambda) = \mathbf{H}(\lambda)\hat{v}(\lambda) = \mathbf{H}(\lambda)(-k\hat{y}(\lambda) + \hat{u}(\lambda))$. Solving for $\hat{y}$ in terms of $\hat{u}$, we see that

$$\mathbf{H}_k(\lambda) = ((\mathbf{H}(\lambda))^{-1} + k)^{-1}. \tag{2.21}$$

Because $\lambda = i\omega^2$, λ is on the imaginary axis when ω is on the real or imaginary axes. Therefore, it follows from (2.20) that $\mathbf{H}(\lambda)$ is imaginary on the imaginary axis. Hence, when λ is imaginary, (2.21) implies that $|\mathbf{H}_k(\lambda)| \leq 1/k$, finishing the lemma. □

From Lemma 2.8 and the dissipativity of A_k, it follows that $\mathbf{H}_k(\lambda)$ is bounded on $\{\text{Re}(z) \geq 0\}$, so (B^*, A_k, B) is input-output stable. This does not in general imply exponential stability, so we need to verify the conditions in Theorem 2.5 to show that we do indeed also have exponential stability. We first need to show that condition (1.5) is equivalent to a condition on the input coefficients b_k given in (2.12).

LEMMA 2.9. *Condition* (1.5) *is true if and only if there exists* $m > 0$ *such that*

$$|\sin((\pi/2 + \pi k)s_1)| > m \quad \textit{for all } k \in Z^+. \tag{2.22}$$

Proof. First suppose (1.5) is true, so $s_1 = p/q$, where p is odd. Then

$$\sin(\pi/2 + \pi k)s_1 = \sin \pi(p(2k+1))/2q.$$

Note that if j is an integer

$$|p(2k+1)/2q - j| = (1/2q)|p(2k+1) - 2qj| \geq (1/2q),$$

the last inequality following from the fact that $p(2k+1)$ is odd and $2qj$ is even. Therefore, $\sin(\pi/2 + \pi k)s_1$ is always bounded away from $|\sin j\pi| = 0$, proving (2.22).

Now suppose (1.5) is not true. Then either s_1 is irrational, or s_1 is of the form p/q, where p and q are coprime integers with p even, so q is odd and $p = 2\tilde{p}$ for some integer $\tilde{p}$. In this last case,

$$\sin(\pi/2 + \pi k)s_1 = \sin(\pi\tilde{p}(2k+1)/q).$$

When $k = (q-1)/2$, this becomes $\sin \pi\tilde{p} = 0$, so (2.22) cannot be true.

If s_1 is irrational, note that for any integer m,

$$|\sin(\pi/2 + \pi k)s_1| = |\sin \pi(s_1/2 + ks_1 + m)|.$$

The set $\{ks_1 + m \mid k, m \in \mathbf{Z}\}$ is dense in $\mathbf{R}$ if s_1 is irrational (see [10, Thm. 438]), which shows that (2.22) cannot be true, finishing the proof of Lemma 2.9. □

LEMMA 2.10. *If* (1.5) *is true, then* (A_k, B) *is stabilizable as in Definition* 2.3, *and* (B_L^*, A_k) *is detectable as in Definition* 2.4.

Proof. We first prove that (A, B) is stabilizable. In this proof we consider A to be the diagonal operator in the basis $\{\Phi_k\}_{k\in I}$ with eigenvalues $\{\lambda_k\}_{k\in I}$, and B to be the column vector $[b_k]_{k\in I}$, where b_k is given by (2.12).

We apply the results in Rebarber [15] to construct our stabilizing feedback. Let

$$p(\lambda) = \prod_{k\in I} \frac{\lambda - \lambda_k}{\lambda - \lambda_k + 2}, \qquad q(\lambda) = \prod_{k\in I} \frac{\lambda - \lambda_k + 1}{\lambda - \lambda_k + 2}.$$

Let

$$\chi_k = \sum_{j\in I} \frac{p(\lambda_k - 1)b_j}{b_k p'(\lambda_k)(\lambda_k - \lambda_j - 1)}\Phi_j,$$

$$h_j = \sum_{k\in I} \frac{q(\lambda_k)b_j p'(\lambda_j)}{q'(\lambda_j - 1)(\lambda_k - \lambda_j + 1)b_k p'(\lambda_k)}\Phi_k.$$

It is shown in [15] that $\{\chi_k\}_{k\in I}$ is a Riesz basis for X and $\langle h_k, \chi_j\rangle = \delta_{jk}$. Let

$$Fx = \sum_{k\in I} \frac{p(\lambda_k - 1)}{b_k p'(\lambda_k)}\langle h_k, x\rangle, \qquad \mathcal{D}(F) = \mathcal{D}(A).$$

We first need to show that (F, A, B) is regular. In [15] it is shown that F can also be written as

$$Fx = \sum_{k\in I} f_k\langle\Phi_k, x\rangle,$$

where

$$f_k = \frac{q(\lambda_k)}{b_k p'(\lambda_k)}\left(\sum_{j\in I} \frac{p(\lambda_j - 1)}{q'(\lambda_j - 1)(\lambda_k - \lambda_j + 1)}\right).$$

In [15] it is shown that $\{p(\lambda_j - 1)q(\lambda_k)/q'(\lambda_j - 1)p'(\lambda_k)\}_{k,j\in I}$ is a bounded set. Because $\lambda_{\pm k} = \pm i(\pi/2 + \pi k)^2$, it follows that $\{f_k\}_{k\in I} \in \ell_\infty$. Therefore, using the Carleson measure criterion as in [9] or [20], we see that F is an admissible observation operator for $S(t)$. As in the derivation of (2.17), we see that $R(\lambda, A)B \in \mathcal{D}(F_L)$ and

$$\mathbf{H}_F(\lambda) := F_L R(\lambda, A)B = \sum_{k\in L} \frac{f_k b_k}{(\lambda - \lambda_k)}.$$

Because $\{f_k b_k\} \in \ell^\infty$, we see from Theorem 2.7 that $\mathbf{H}_F(\lambda)$ is bounded in $\mathbf{C}^+_\alpha$ for any $\alpha > 0$, and that $\mathbf{H}_F(\lambda) \to 0$ as $\lambda \to \infty$ along the real axis. Therefore, (F, A, B) is regular.

Theorem 2.7 shows that (2.19) is true for this transfer function, so Theorem 3.13 in [24] shows that I is an admissible feedback for (F_L, A, B). Therefore Theorems 4.6 and 6.6 in [24] show that $(F_L, A + BF_L, B)$ is a regular system, where as usual we interpret the first F_L as the restriction of F_L to the domain of $A + BF_L$.

If (1.5) is true, then (2.12) and Lemma 2.9 show that the sequence of b_k's is bounded above and below, and Theorem 2 in [15] shows that $A + BF_L$ has eigenvalues $\{\lambda_k - 1\}_{k\in I}$ and eigenvectors $\{\chi_k\}_{k\in I}$. Because these eigenvectors are a Riesz basis for X, $A + BF_L$ generates an exponentially stable semigroup. This shows that (A, B) is stabilizable in the sense of Definition 2.3.

An obvious candidate for the operator that stabilizes $(A_k, B) = (A - kBB_L^*, B)$ is $\tilde{F} = F_L + kB_L^*$, because $(A - kBB_L^*) + B\tilde{F} = A + BF_L$ is exponentially stable. Because the transfer functions for (B^*, A, B) and (F, A, B) satisfy (2.19), it follows from Theorem 6.6 in [24] that $(B_L^*, A - kBB_L^*, B)$ and $(F_L, A - kBB_L^*, B)$ are also regular. Therefore $(\tilde{F}, A - kBB_L^*, B)$ is regular. To show that I is an admissible feedback for $(\tilde{F}, A - kBB_L^*, B)$, first note that we know from our previous work that I is an admissible feedback for (F, A, B) and (kB^*, A, B), and then apply the following lemma, which is proved in the Appendix. □

LEMMA 2.11. *Suppose K is an admissible feedback for the regular systems (C, A, B) and (C', A, B). Then K is admissible for $(C'_L - C_L, A + BKC_L, B)$.*

This finishes the proof that (A_k, B) is stabilizable in the sense given in Definition 2.3. We can use a similar dual argument to show that (B^*, A_k) is detectable in the sense of Definition 2.4, finishing the proof of Lemma 2.10.

Putting together Lemmas 2.6, 2.8, and 2.10 allows us to apply Theorem 2.5, so it follows that condition (1.5) is sufficient for exponential stability of the system (1.2), (1.3), (1.4). Therefore, to finish the proof of Theorem 1.1 we need to prove the necessity of condition (1.5). It is clear that in our case the eigenvalues satisfy the following condition.

Condition 1. For any $h > 0$, the number of eigenvalues N_h in $\{z \,|\, 0 \leq \mathrm{Re}(z) \leq h,$ $a - h \leq \mathrm{Im}(z) \leq a + h\}$ satisfies $N_h \leq Mh$ for some $M < \infty$ that is independent of a.

LEMMA 2.12. *Suppose A is a diagonal operator with eigenvalues that satisfy Condition* 1, *and suppose B is an admissible input operator for $S(t)$ that is represented by the column vector $[b_k]_{k \in I}$. Suppose there exists a functional F on X such that $A_F = A + BF$ generates an exponentially stable semigroup $S_F(t)$, and F is an admissible observation operator* (*or an extension of an admissible observation operator*) *for $S_F(t)$. Then there exists $M > 0$ such that*

$$\sum_{k \in I} |x_k / b_k|^2 \leq M \sum_{k \in I} |x_k|^2 \tag{2.23}$$

for all $\{x_k\}_{k \in I} \in \ell^2$.

Proof. Let $M > 0$, $\alpha < 0$ be such that $\|S_F(t)\| \leq Me^{\alpha t}$, and let $\beta \in (\alpha, 0)$. The variation of parameters formula implies that for $x \in X$,

$$S_F(t)x = S(t)x + \int_0^t S(t - \tau)Bu(\tau; x)d\tau, \tag{2.24}$$

where $u(\tau; x) = FS_F(\tau)x$. Then $\hat{u}(\cdot\,; x)$ is in the Hardy space $H^2(\mathbf{C}_\beta^+)$, which is the space of Laplace transforms of all u in $L^2_{\mathrm{loc}}[0, \infty)$ such that $e^{-\beta \cdot} u \in L^2[0, \infty)$. Because B is admissible, we can take the Laplace transform of (2.24), obtaining for $\mathrm{Re}(\lambda) > 0$,

$$R(\lambda, A_F)x = R(\lambda, A)x + R(\lambda, A)B\hat{u}(\lambda; x).$$

Because $R(\lambda, A_F)x$ and $\hat{u}(\lambda; x)$ are holomorphic in $\mathbf{C}_\beta^+$ and $R(\lambda, A)x$ is meromorphic in $\mathbf{C}_\beta^+$, we can consider the left side of this equation to be a holomorphic extension of the right side. Representing x in a basis $\{\Phi_k\}_{k \in I}$ of X by $\sum_{k \in I} x_k \Phi_k$, this equation is equivalent to

$$R(\lambda, A_F)x = \sum_{k \in I} \frac{x_k + b_k \hat{u}(\lambda; x)}{\lambda - \lambda_k} \Phi_k. \tag{2.25}$$

Because $R(\lambda, A_F)x$ is holomorphic on the imaginary axis, (2.25) is true only if

$$\hat{u}(\lambda_k; x) = -x_k / b_k. \tag{2.26}$$

Now define the following measure μ on the Borel subsets of $\mathbf{C}_\beta^+$:

$$\mu(\lambda_k) = 1, \qquad \mu(\{z \,|\, \mathrm{Re}(z) > \beta\} \backslash \{\lambda_k\}_{k\in I}) = 0. \tag{2.27}$$

Because $\{\lambda_k\}_{k\in I}$ satisfies Condition 1, we easily see that μ is a Carleson measure on $\mathbf{C}_\beta^+$ (see [9] and [20] for the definition of Carleson measure and for other examples along these lines). The Carleson measure theorem as stated in [9] implies that there exists $M_1 > 0$ such that for all $\phi \in H^2(\mathbf{C}_\beta^+)$,

$$\sum_{k\in I} |\phi(\lambda_k)|^2 \le M_1 \int_{-\infty}^{+\infty} \lim_{\rho\to\beta} |\phi(\rho + i\eta)|^2 \, d\eta. \tag{2.28}$$

Note that by the Plancherel theorem,

$$\int_{-\infty}^{+\infty} \lim_{\rho\to\beta} |\hat{u}(\rho + i\eta, x)|^2 \, d\eta = 2\pi \int_0^\infty |e^{-\beta t} u(t; x)|^2 \, dt. \tag{2.29}$$

Because $u(t; x) = FS_F(t)x$, F is an extension of an admissible observation operator for $S_F(t)$, and $S_F(t)$ has exponential decay α, there exist M_2, $M_3 > 0$ such that

(2.30)

$$\int_0^\infty |e^{-\beta t} u(t; x)|^2 \, dt \le M_2 \int_0^\infty |e^{-\beta t} S_F(t)x|^2 \le M_3 \left(\sum_{k\in I} |x_k|^2 \right) \int_0^\infty e^{2(\alpha-\beta)t} \, dt.$$

Let $\phi(\lambda) = \hat{u}(\lambda, x)$ in (2.28), keeping in mind (2.26), (2.29), and (2.30), and we get (2.23), finishing the lemma. □

Remark 2.13. The proof of Lemma 2.12 is motivated by the idea of *open-loop stabilizability* (see Zwart [26]). In the proof we show that (2.23) is a necessary condition for open-loop stabilizability, hence also for closed-loop stabilizability.

Now suppose that the system (1.2), (1.3), (1.4) is exponentially stable, so $A_k = A - kBB_L^*$ generates an exponentially stable semigroup. Then Lemma 2.12 is applicable with $F = -kB_L^*$, so (2.23) is true, and $\{1/b_k\}_{k\in I}$ must be a bounded sequence. This means that condition (1.5) is a necessary condition for stabilization, which completes the proof of Theorem 1.1.

3. Proof of Theorem 1.3. We need to approach the system (1.2), (1.6), (1.7) differently than the system (1.2), (1.3), (1.4). To see why this is the case, we briefly discuss the state-space formulation for the controlled, observed system associated with (1.2), (1.6), (1.7) and explain why our approach in §2 cannot be applied directly to this system.

Let

$$X = \{[x_1, x_2]^T \in H^2[0,1] \oplus L^2[0,1] \,|\, x_1(0) = x_1(1) = 0\}, \tag{3.1}$$

and let A be the matrix (2.2) with domain

$$\begin{aligned} \mathcal{D}(A) = \{[x_1, x_2]^T \in H^4[0,1] \oplus H^2[0,1] \,|\, x_1(0) &= x_1(1) \\ = D^2x_1(0) = D^2x_1(0) = x_2(0) &= x_2(1) = 0\}, \end{aligned} \tag{3.2}$$

so it is easy to see that A generates a C_0-semigroup $S(t)$ on X. If we let $x(t) = [w(\cdot, t), \dot{w}(\cdot, t)]^T$ and $B = [0, \delta'(\cdot - s_1)]^T$, then (1.2), (1.6) with control

$$D^2w(s_1^-, t) - D^2w(s_1^+, t) = u(t) \tag{3.3}$$

can be seen to be equivalent to (2.9). Similarly, (1.2), (1.6), (1.7) with $\hat{k} = 0$, and observation

$$y(t) = D\dot{w}(s_1, t) \tag{3.4}$$

is equivalent to $\dot{x}(t) = Ax(t)$ and (2.11). However, B is not an admissible input operator for $S(t)$, B^* is not an admissible observation operator for $S(t)$, and the transfer function for (1.2), (1.6), (3.3), (3.4) can easily be computed and is not bounded on any half plane $\mathbf{C}_\alpha^+$. Such a system is ill posed, and we cannot apply the approach of §2 directly to this system.

If we formally close the loop in (1.2), (1.6), (3.3), (3.4) with the feedback $u(t) = -\hat{k}y(t)$, $\hat{k} > 0$, then the resulting system is (1.2), (1.6), (1.7) is shown in [5] to have solutions with nonincreasing energy, and the underlying semigroup generator is easily seen to be dissipative using the Lumer–Phillips theorem. This implies that the controlled, observed system (1.2), (1.6), (3.3), (3.4), while possibly natural from a physical point of view, is not the appropriate systems-theoretic framework for studying the feedback system (1.2), (1.6), (1.7). This is because the input-output relation defined by (1.2), (1.6), (3.3), (3.4) is not well posed in the sense of [8], [18], [22], and [23]. In Rebarber and Townley [17] an ill-posed system is analyzed by reversing the roles of $u(t)$ and $y(t)$, which is what we do here. Therefore, we consider the "inverse" system (1.2), (1.6),

$$D\dot{w}(s_1, t) = u(t), \tag{3.5}$$

$$y(t) = D^2w(s_1^-, t) - D^2w(s_1^+, t). \tag{3.6}$$

When we put this system into state-space form, we see that it is regular. When the loop in this system is closed by $u(t) = -(1/\hat{k})y(t)$, we obtain the system (1.2), (1.6), (1.7), and we apply the approach from §2 to the inverse system. The difficulties in this approach arise in characterizing which joint placements lead to stabilizability and detectability. This is because the underlying semigroup generator is not standard, and the eigenvalues $\{\lambda_k\}$ and input coefficients $\{b_k\}$ are not easily computed. The analysis of the eigenvalues and input coefficients is done in Lemmas 3.2, 3.5, 3.6, 3.7, and 3.8.

We start by discussing the state-space formulation for (1.2), (1.6), (3.5), (3.6). We note here that we will be using most of the notation as in §2, to mirror the development in that section. However, much of the notation needs to be redefined for this new system, so the notation will not represent the same things unless explicitly stated. Once the analysis of the eigenvalues and input coefficients is completed the rest of the proof of Theorem 1.3 is almost identical to the proof of Theorem 1.1. Because many of the proofs in this section proceed like the proofs in §2, we will not repeat the details and merely refer to §2.

The state space is X, given by (3.1), with the inner product (2.1). Let A be the operator given by the matrix (2.2) with domain

$$\begin{aligned}\mathcal{D}(A) = \{[x_1, x_2]^T \in X \mid\ & x_1|_{[0,s_1)} \in H^4[0, s_1),\ x_1|_{(s_1,1]} \in H^4(s_1, 1],\\ & x_2 \in H^2[0,1],\ D^2x_1(0) = D^2x_1(1) = x_2(0) = x_2(1) = 0,\\ & Dx_2(s_1) = 0,\ D^3x_1(s_1^-) = D^3x_1(s_1^+)\}.\end{aligned} \tag{3.7}$$

Remark 3.1. The operator A here is the generator of the *zero dynamics* of (1.2), (1.6), (1.7) (see Byrnes and Gilliam [2]). The zero dynamics of a feedback system with gain parameter $\hat{k} \in \mathbf{R}$ is that system obtained formally by letting $\hat{k} \to \infty$. It is shown in Rebarber and Townley [17] that if the inverse system is regular with underlying semigroup $S(t)$, and if $T_{\hat{k}}(t)$ is the semigroup associated with $y(t) = -\hat{k}u(t)$ in the original system, then $\lim_{\hat{k}\to\infty} T_{\hat{k}}(t) = S(t)$ in the operator norm for every $t > 0$.

This generator A is not as familiar as the generator A in §2, so we briefly discuss here how to show that A is skew adjoint on X. Let $x = [x_1, x_2]^T \in \mathcal{D}(A)$. Let M be the operator given by (2.2) and $v = [v_1, v_2] \in X$ be such that $v_1|_{[0,s_1)} \in H^4[0, s_1)$, $v_1|_{s_1,1]} \in H^4(s_1, 1]$, $v_2|_{[0,s_1)} \in H^2[0, s_1)$, and $v_2|_{(s_1,1]} \in H^2(s_1, 1]$. Using integration by parts we see that $\langle Ax, v\rangle = \langle x, -Mv\rangle$ plus terms at 0, 1, and s_1. Thus A^* is given by $-M$ on the domain of all v so that the terms at 0, 1, and s_1 are zero. The terms at 0 and 1 are fairly standard and imply that $v \in \mathcal{D}(A^*)$ must satisfy $v_2(0) = D^2v_1(0) = v_2(1) = D^2v_1(1) = 0$. The terms at s_1 are as follows:

(1) $x_2(s_1^-)\overline{D^3v_1(s_1^-)} - x_2(s_1^+)\overline{D^3v_1(s_1^+)}$. Because $x_2 \in H^2[0, 1]$, $x_2(s_1^-) = x_2(s_1^+)$, which implies that $D^3v_1(s_1^-) = D^3v_1(s_1^+)$.

(2) $Dx_2(s_1^-)\overline{D^2v_1(s_1^-)} - Dx_2(s_1^+)\overline{D^2v_1(s_1^+)}$. Because $Dx_2(s_1) = 0$, this implies no condition on $D^2v_1(s_1)$.

(3) $D^2x_1(s_1^-)\overline{Dv_2(s_1^-)} - D^2x_1(s_1^+)\overline{Dv_2(s_1^+)}$. Because there is no condition on D^2x_1 at s_1, this implies that $Dv_2(s_1^+) = Dv_2(s_1^-) = 0$.

(4) $D^3x_1(s_1^-)\overline{v_2(s_1^-)} - D^3x_1(s_1^+)\overline{v_2(s_1^+)}$. Because $D^3x_1(s_1^-) = D^3x_1(s_1^+)$, this implies that $v_2(s_1^-) = v_2(s_1^+)$.

Putting this all together, we see that $\mathcal{D}(A^*) = \mathcal{D}(A)$, so $A^* = -A$ and A is skew adjoint. Because it is clear from the Rellich–Kondrachov theorem (see [1, Thm. 6.2]) that the $\mathcal{D}(A)$ is compactly embedded in X, A has compact resolvent. Therefore, the spectrum of A consists solely of eigenvalues $\{\lambda_k\}$ on the imaginary axis with associated eigenvectors $\{\Phi_k\}$ that form an orthogonal basis of X. Hence A generates a $\mathbf{C}_0$-semigroup $S(t)$ on X such that $\|S(t)\| = 1$.

To identify the appropriate input operator for the control system (1.2), (1.6), (3.5), we let $\hat{A}$ be the extension of A with domain

$$\begin{aligned}\mathcal{D}(\hat{A}) = \{[x_1, x_2]^T \in X \mid\, & x_1|_{[0,s_1)} \in H^4[0, s_1),\ x_1|_{(s_1,1]} \in H^4(s_1, 1],\\ & x_2 \in H^2[0, 1],\ D^2x_1(0) = D^2x_1(1) = x_2(0) = x_2(1) = 0,\\ & x_2(s_1^-) = x_2(s_1^+),\ D^3x_1(s_1^-) = D^3x_1(s_1^+)\}.\end{aligned}$$

Let $x = [x_1, x_2]^T \in \mathcal{D}(\hat{A})$ and $v = [v_1, v_2]^T \in \mathcal{D}(A^*)$. Using integration by parts we find that

$$\langle \hat{A}x, v\rangle = \langle x, -Av\rangle + Dx_2(s_1)\overline{D^2v_1(s_1^-) - D^2v_1(s_1^+)}. \tag{3.8}$$

Let $B \in \mathcal{D}(A^*)'$ be given by

$$B = \begin{bmatrix} \delta''(\cdot - s_1^+) - \delta''(\cdot - s_1^-) \\ 0 \end{bmatrix}. \tag{3.9}$$

We interpret B as a functional on $\mathcal{D}(A^*) = \mathcal{D}(A)$ as

$$B[v_1, v_2]^T = D^2v_1(s_1^-) - D^2v_1(s_1^+).$$

This interpretation of B is not consistent with the inner product (2.1) but is consistent with the usual interpretation of δ''. Let $\tilde{A}$ be defined by (2.7). Then from (3.8) we see that $\hat{A}x = \tilde{A}x + Dx_2(s_1)B$ in X_{-1}. As in §2, we see that if $x(t) = [w(\cdot, t), \dot{w}(\cdot, t)]^T$ and $w(s, t)$ satisfies (1.2), (1.6), and (3.5), then $\hat{A}x(t) = \tilde{A}x(t) + Bu(t)$ in X_{-1}. Therefore, the state-space equation for (1.2), (1.6), (3.5) is (2.9).

The observation (3.6) is given by $y(t) = B^*x(t)$. We prove in Lemma 3.7 that for any $s_1 \in [0, 1]$, B is an admissible input operator, and hence B^* is an admissible observation operator.

Letting $k = (1/\hat{k})$, the system (1.2), (1.6), (1.7) is equivalent to $\dot{x}(t) = A_k x(t)$, where A_k is given by the matrix (2.2) with domain

$$\mathcal{D}(A_k) = \{[x_1, x_2]^T \in \mathcal{D}(\hat{A}) \mid Dx_2(s_1) = -k(D^2x_1(s_1^-) - D^2x_1(s_1^+))\}.$$

It is again an easy consequence of [5, (1.4)] and the Lumer–Phillips theorem that A_k is a dissipative operator that generates a C_0-semigroup of contractions $S_k(t)$ on X.

As in the proof of Theorem 1.1, we show that $S_k(t)$ is exponentially stable by showing that $R(\lambda, A_k)$ is bounded on the imaginary axis. To do this, we apply Theorem 2.5 to the controlled, observed system for A_k given by (1.2), (1.6), (3.6) and

$$D\dot{w}(s,t) + k[D^2w(s_1^+,t) - D^2w(s_1^-,t)] = u(t). \tag{3.10}$$

As in the proof of Theorem 1.1, when it is shown that the feedthrough for this system is zero, it will follow that the system is equivalent to (2.15) with observation (2.11). Let $\mathbf{H}_k$ be the transfer function for this system.

To apply the approach in §2 to the generator A for this reversed system, we need to analyze the eigenvalues and eigenvectors of A in detail. Although this is a simple task if we are dealing with the generator for the "forward" system with domain (3.2) (which is unfortunately associated with a control system that is not well posed), the operator A we are working with here, with domain (3.7), is not standard. The asymptotic analysis of the eigenvalues is simpler here than in Chen et al. [5] because in that paper the eigenvalues for A_k were analyzed, whereas we are only interested in the eigenvalues for $k = 0$. However, in this case we also need a formula for the normalized eigenvectors because these are used for the computation of the input coefficients b_k. The following lemma characterizes the eigenvalues and eigenvectors of A.

LEMMA 3.2. *All of the eigenvalues and eigenvectors of A are of one of the three forms described below in parts* (a), (b), *and* (c).

(a) $\lambda = 0$ *is an eigenvalue of multiplicity one with associated eigenvector* $\Psi_0 = [\psi_0, 0]^T$, *where*

$$\psi_0(s) = s^3 + (3s_1^2 - 6s_1 + 2)s \quad \textit{for } s \in [0, s_1), \tag{3.11}$$

$$\psi_0(s) = (s-1)^3 + (3s_1^2 - 1)(s-1) \quad \textit{for } s \in (s_1, 1]. \tag{3.12}$$

(b) *If* $\cos(\pi n s_1) = 0$ *for some* $n \in \mathbf{Z}^+$, *then* $\lambda = \pm i(\pi n)^2$ *are eigenvalues of multiplicity one with associated eigenvectors*

$$\Psi_{\pm n} = \begin{bmatrix} \psi_n / \pm i\pi n^2 \\ \psi_n \end{bmatrix}, \tag{3.13}$$

where

$$\psi_n(s) = \sin(\pi n s) \quad \textit{for } s \in [0, 1]. \tag{3.14}$$

(c) *Let ω be a real positive solution of*

$$g(\omega) := \tanh(\omega s_1) - \tan(\omega s_1) - \tanh \omega(s_1 - 1) + \tan \omega(s_1 - 1) = 0. \tag{3.15}$$

Then $\lambda = \pm i\omega^2$ are eigenvalues with associated eigenvectors

$$\Psi_{\pm\omega} = \begin{bmatrix} \psi_\omega / \pm i\omega^2 \\ \psi_\omega \end{bmatrix}, \tag{3.16}$$

where

$$\psi_\omega(s) = \frac{\sinh(\omega s)}{\cosh(\omega s_1)} - \frac{\sin(\omega s)}{\cos(\omega s_1)} \quad \textit{for } s \in [0, s_1), \tag{3.17}$$

$$\psi_\omega(s) = \frac{\sinh \omega(s-1)}{\cosh \omega(s_1 - 1)} - \frac{\sin \omega(s-1)}{\cos \omega(s_1 - 1)} \quad \textit{for } s \in (s_1, 1]. \tag{3.18}$$

Proof. A is skew adjoint, so the eigenvalues are on the imaginary axis, and it is easy to show that the eigenvalues come in complex conjugate pairs. Therefore, the eigenvalues are of the form $\lambda = \pm i\omega^2$, where ω is on the nonnegative real axis.

Assume that $[\psi, \chi]^T$ is an eigenvector associated with the eigenvalue λ. Then $\chi = \lambda\psi$ and ψ satisfies

$$\lambda^2\psi(s) + D^4\psi(s) = 0 \tag{3.19}$$

with boundary conditions

$$\begin{aligned} \psi(0) &= D^2\psi(0) = 0, \\ \psi(1) &= D^2\psi(1) = 0, \\ \psi(s_1^-) &= \psi(s_1^+), \\ D\psi(s_1^-) &= D\psi(s_1^+), \\ D^3\psi(s_1^-) &= D^3\psi(s_1^+), \end{aligned} \tag{3.20}$$

$$\lambda D\psi(s_1) = 0. \tag{3.21}$$

It is easy to verify that if $\lambda = 0$, then all solutions of (3.19), (3.20), (3.21) are given by $c\psi(s)$, where c is a scalar and ψ is given by (3.11), (3.12).

Now assume $\lambda \neq 0$, so (3.21) becomes

$$D\psi(s_1) = 0. \tag{3.22}$$

Equation (3.19) and the first line of (3.20) imply that

$$\psi(s) = C_1 \sinh(\omega s) + C_2 \sin(\omega s) \quad \text{for } s \in [0, s_1). \tag{3.23}$$

Equation (3.19) and the second line of (3.20) imply that

$$\psi(s) = C_3 \sinh \omega(s-1) + C_4 \sin \omega(s-1) \quad \text{for } s \in (s_1, 1]. \tag{3.24}$$

The third line of (3.20) is true if and only if

$$C_1 \sinh(\omega s_1) + C_2 \sin(\omega s_1) = C_3 \sinh \omega(s_1 - 1) + C_4 \sin \omega(s_1 - 1). \tag{3.25}$$

The fourth and fifth lines of (3.20) are true if and only if

$$C_1 \cosh(\omega s_1) = C_3 \cosh \omega(s_1 - 1), \tag{3.26}$$

$$C_2 \cosh(\omega s_1) = C_4 \cos \omega(s_1 - 1). \tag{3.27}$$

Equation (3.22) is true if and only if

$$C_1 \cosh(\omega s_1) = -C_2 \cos(\omega s_1), \tag{3.28}$$

$$C_3 \cosh \omega(s_1 - 1) = -C_4 \cos \omega(s_1 - 1). \tag{3.29}$$

To find real positive ω so that (3.25)–(3.29) have nonzero solutions for C_1, C_2, C_3, and C_4, we consider four cases.

Case 1. Suppose $\cos \omega(s_1 - 1) = 0$ and $\cos(\omega s_1) \neq 0$. Then (3.27) implies that $C_2 = 0$, (3.28) implies that $C_1 = 0$, and (3.26) implies that $C_3 = 0$. Finally, (3.25) implies that $C_4 = 0$, so $\pm i\omega^2$ cannot be an eigenvalue.

Case 2. Suppose $\cos \omega(s_1 - 1) \neq 0$ and $\cos(\omega s_1) = 0$. Then (3.27) implies that $C_4 = 0$, (3.29) implies that $C_3 = 0$, and (3.26) implies that $C_1 = 0$. Finally, (3.25) implies that $C_2 = 0$, so $\pm i\omega^2$ cannot be an eigenvalue.

Case 3. Suppose $\cos \omega(s_1 - 1) = 0$ and $\cos(\omega s_1) = 0$. Then $\omega(s_1 - 1) = \pi(1/2 - j)$ and $\omega s_1 = \pi(-1/2 + k)$ for some j, $k \in Z^+$. Therefore $\omega = \pi n$ for some $n \in Z^+$. Then (3.28) is true if and only if $C_1 = 0$, and (3.29) is true if and only if $C_3 = 0$. Equation (3.25) is true if and only if

$$C_2 \sin(n\pi s_1) = C_4 \sin n\pi(s_1 - 1).$$

Therefore

$$C_4 = C_2(-1)^n.$$

Hence, using (3.23) and (3.24), we see that $\psi(s)$ is a solution of (3.25)–(3.29) if and only if

$$\psi(s) = C_2 \sin(\pi n s) \quad \text{for } s \in [0, s_1),$$

$$\psi(s) = C_2(-1)^n \sin \pi n(s - 1) = C_2 \sin(\pi n s) \quad \text{for } s \in (s_1, 1].$$

Therefore (3.13), (3.14) give eigenvectors associated with $\omega = \pi n$, where n is such that $\cos(\pi n s_1) = 0$, and $\pm i(\pi n)^2$ are eigenvalues of multiplicity one.

Case 4. Suppose $\cos \omega(s_1 - 1) \neq 0$ and $\cos(\omega s_1) \neq 0$. Then (3.26) is true if and only if

$$C_3 = C_1(\cosh(\omega s_1)/\cosh \omega(s_1 - 1)). \tag{3.30}$$

Equation (3.28) is true if and only if

$$C_2 = C_1(-\cosh(\omega s_1)/\cos(\omega s_1)). \tag{3.31}$$

Using this, we see that (3.27) is true if and only if

$$C_4 = C_1(-\cosh(\omega s_1)/\cos \omega(s_1 - 1)). \tag{3.32}$$

Equation (3.29) is also satisfied if C_3 and C_4 are given by (3.30) and (3.32). Plugging (3.30), (3.31), and (3.32) into (3.25) and dividing by $\cosh(\omega s_1)$, we see that $\pm i\omega^2$ is an eigenvalue in Case 4 if and only if (3.15) is true. Plugging (3.30), (3.31), (3.32) into (3.23) and (3.24) and letting $C_1 = 1/\cosh(\omega s_1)$, we obtain the eigenvectors given by (3.16), (3.17), (3.18). □

It is proved in Proposition 3.3, below, that the eigenvalues of A described in part (c) of Lemma 3.2 are also of multiplicity 1.

For our purposes we need the normalized eigenvectors of A. For the eigenvector in part (a) of Lemma 3.2, let $C_0 = 1/\|\Psi_0\|$ and let Φ_0 be the normalized eigenvector $C_0\Psi_0$. For the eigenvectors in part (b) of Lemma 3.2, we easily find that the normalizing constant is 1, so let $\Phi_{\pm n} = \Psi_{\pm n}$. For the eigenvectors in part (c) of Lemma 3.2, let ω be a positive solution of (3.15). Let

$$\Phi_{\pm\omega} = C_\omega \Psi_{\pm\omega}, \tag{3.33}$$

where

$$1/C_\omega^2 = \|[\psi_\omega/\omega^2, \psi_\omega]^T\|^2 = \int_0^1 \{|D^2\psi_w(s)/\omega^2|^2 + |\psi_\omega(s)|^2\}ds. \tag{3.34}$$

Using (3.17) on $[0, s_1)$, we compute that

$$\begin{aligned}\int_0^{s_1} \{|D^2\psi_\omega(s)/\omega^2|^2 + |\psi_\omega(s)|^2\}ds = {}& \frac{1}{\cosh^2(\omega s_1)}\left[\frac{\sinh(2\omega s_1)}{2\omega} - s_2\right] \\ & - \frac{1}{\cos^2(\omega s_1)}\left[\frac{\sin(2\omega s_1)}{2\omega} - s_1\right].\end{aligned} \tag{3.35}$$

Using (3.18) on $(s_1, 1]$, we compute that

(3.36)

$$\begin{aligned}\int_{s_1}^1 \{|D^2\psi_\omega(s)/\omega^2|^2 + |\psi_\omega(s)|^2\}ds = {}& \frac{-1}{\cosh^2\,\omega(s_1-1)}\left[\frac{\sinh\,2\omega(s_1-1)}{2\omega} - (s_1-1)\right] \\ & + \frac{1}{\cos^2\,\omega(s_1-1)}\left[\frac{\sin\,2\omega(s_1-1)}{2\omega} - (s_1-1)\right].\end{aligned}$$

Putting (3.35) and (3.36) together, we see that

(3.37)

$$\begin{aligned}1/C_\omega^2 = {}& (1/\omega)(\tanh(\omega s_1) - \tan(\omega s_1) - \tanh\,\omega(s_1-1) + \tan\,\omega(s_1-1)) \\ & + s_1(\sec^2(\omega s_1) - \operatorname{sech}^2(\omega s_1)) + (s_1-1)(\operatorname{sech}^2\omega(s_1-1) - \sec^2\,\omega(s_1-1)).\end{aligned}$$

Because ω satisfies (3.15), we can simplify this expression to

(3.38)

$$C_\omega = [s_1(\sec^2(\omega s_1) - \operatorname{sech}^2(\omega s_1)) + (s_1-1)(\operatorname{sech}^2\omega(s_1-1) - \sec^2\,\omega(s_1-1))]^{-1/2}.$$

PROPOSITION 3.3. *The eigenvalues of A are all of multiplicity one.*

Proof. We only need to show that the zeros of g given by (3.15) are of multiplicity one. If ω is a zero of g, then comparing (3.15) to (3.38) we see that $g'(\omega) = -1/C_\omega^2$, which is always negative. □

We can now compute the input coefficients $B^*\Phi$ associated with B given by (3.9). The input coefficient associated with Φ_0 is $b_0 = -6C_0$, and all we need to note about this is that it is not 0. The input coefficients associated with $\Phi_{\pm n}$ are

$$b_{\pm n} = 0. \tag{3.39}$$

The input coefficients associated with $\Phi_{\pm\omega}$ are $b_{\pm\omega} = B\Phi_{\pm\omega}$, where B is given by (3.9). Using (3.17) and (3.18), these are easily computed to be $b_{\pm\omega} = \pm b_\omega$, where

$$b_\omega = iC_\omega[\tanh\,\omega(s_1-1) + \tan\,\omega(s_1-1) - \tanh(\omega s_1) - \tan(\omega s_1)],$$

and C_ω is given by (3.38). Using (3.15) we see that

$$b_\omega = -2iC_\omega[\tanh(\omega s_1) - \tanh \omega(s_1 - 1)]. \tag{3.40}$$

The next lemma is proved in almost exactly the same way as Lemma 2.9, so the proof is omitted.

LEMMA 3.4. *Condition* (1.8) *is true if and only if there exists $m > 0$ such that*

$$|\cos(\pi k s_1)| > m \quad \text{for all } k \in Z^+. \tag{3.41}$$

Now we can relate the behavior of the eigenvalues to condition (1.8).

LEMMA 3.5. (a) *Suppose that* (1.8) *is true for some $m > 0$. Then there exists $m_1 > 0$ and $m_2 > 0$ such that for any positive zero ω* of g,

$$|\cos(\omega s_1)| > m_1, \tag{3.42}$$

$$|\cos \omega(s_1 - 1)| > m_2. \tag{3.43}$$

Furthermore, there exists $m_3 > 0$ such that for any two zeros ω_1 and ω_2 of g,

$$|w_1 - w_2| > m_3. \tag{3.44}$$

(b) *Suppose there exists a sequence of positive integers $\{n_l\}$ such that*

$$\lim_{l \to \infty} \cos(n_l \pi s_1) = 0. \tag{3.45}$$

There there exists a sequence $\{\omega_k\}$ of positive zeros of g such that

$$\lim_{k \to \infty} |\cos(\omega_k s_1)| = 0. \tag{3.46}$$

Proof. We first prove part (a), so we assume that (1.8), hence (3.41), is true. Suppose $\{\omega_k\}$ is a sequence of positive zeros of g such that (3.46) is true. Examining g in (3.15), we see that this can only happen if $\lim_{k\to\infty} |\cos \omega_k(s_1 - 1)| = 0$. Therefore there exists a sequence of integers $\{m_k\}$ and $\{n_k\}$ such that

$$\lim_{k \to \infty} |w_k s_1 - \pi(1/2 + n_k)| = 0,$$

$$\lim_{k \to \infty} |w_k(s_1 - 1) - \pi(1/2 + m_k)| = 0.$$

This means that

$$\lim_{k \to \infty} |w_k - \pi j_k| = 0,$$

where $j_k = n_k - m_k$, so

$$\lim_{k \to \infty} |\cos \pi j_k s_1| = 0,$$

contradicting (3.41). From this we see that if (3.41) is true, then (3.46) cannot be true. Note that if ω is a zero of g, (3.15) shows that $\cos(\omega s_1) \neq 0$ and $\cos \omega(s_1 - 1) \neq 0$. Therefore, if (3.41) is true, then (3.42) is true for some m_1 and all real positive zeros of g. It is clear from the form of g given in (3.15) that if (3.42) is true, then (3.43) must also be true.

The vertical asymptotes of the graph of $g(\omega)$ for ω on the real positive axis are the lines $\omega = \eta_k$ and $\omega = \tilde{\eta}_j$, where

$$\eta_k = \pi(k + 1/2)/(1 - s_1), \qquad \tilde{\eta}_j = \pi(j + 1/2)/s_1, \tag{3.47}$$

where k and j are nonnegative integers. Note that $g'(\omega) = (s_1 - 1)[\sec^2 \omega(s_1 - 1) - \operatorname{sech}^2 \omega(s_1 - 1)] + s_1[\operatorname{sech}^2(\omega s_1) - \sec^2(\omega s_1)]$, which is always negative where it is defined. Therefore, between any adjacent numbers in

$$\Lambda := \{\eta_k\}_{k=0}^{\infty} \cup \{\tilde{\eta}_j\}_{j=0}^{\infty} \tag{3.48}$$

there is exactly one positive zero of g.

We show that when (3.41) is true the points in Λ are separated in the sense that there exists $m > 0$ such that if $\eta, \mu \in \Lambda$, then $|\eta - \mu| > m$. Because $|\eta_{k_1} - \eta_{k_2}| \geq \pi|k_1 - k_2|/(1 - s_1)$ and $|\tilde{\eta}_{j_1} - \tilde{\eta}_{j_2}| \geq \pi|j_1 - j_2|/s_1$, we need to find $m > 0$ so that $|\eta_k - \tilde{\eta}_j| > m$ for all nonnegative integers k and j. To that end, note that

$$|\eta_k - \tilde{\eta}_j| = \frac{\pi|s_1(1 + k + j) - (1/2 + j)|}{s_1(1 - s_1)}. \tag{3.49}$$

By (3.41), there exists an $m_1 > 0$ such that

$$|ns_1 - (1/2 + j)| > m_1$$

for all nonnegative integers n and j. Comparing this with (3.49) (with $n = 1 + k + j$), we see that

$$|\eta_k - \tilde{\eta}_j| > \pi m_1/s_1(1 - s_1)$$

for all nonnegative integers k and j, showing that the points in Λ are separated.

Let ω be a positive zero of g, so (3.42) and (3.43) are true. Because $\cos \eta_k(s_1 - 1) = 0$ and $\cos(\tilde{\eta}_j s_1) = 0$, we see that there exists $m_2, m_3 > 0$ such that

$$|\omega s_1 - \tilde{\eta}_j s_1| > m_2,$$

$$|\omega(s_1 - 1) - \eta_k(s_1 - 1)| > m_3.$$

Let m_4 be the minimum of $\{m_2/s_1, m_3/(1 - s_1)\}$. Therefore, if $\eta \in \Lambda$, $|\eta - \omega| > m_4$. Because the points in Λ are separated and there is only one positive real zero·ω of g between any adjacent points of Λ, this shows that the positive real zeros of f are separated in the sense given in (3.44).

We now prove part (b). Suppose (3.45) is true for a sequence of positive integers $\{n_l\}$. Therefore there exists a sequence of positive integers $\{m_l\}$ such that

$$\lim_{l \to \infty} |n_l s_1 - (1/2 + m_l)| = 0. \tag{3.50}$$

In (3.49) let $j = m_l$ and $k = n_l - m_l - 1$, so we obtain

$$|\eta_{n_l - m_l - 1} - \tilde{\eta}_{m_l}| = \frac{\pi|s_1 n_l - (1/2 + m_l)|}{s_1(1 - s_1)}.$$

Comparing this with (3.50) we see that

$$\lim_{l \to \infty} |\eta_{n_l - m_l - 1} - \tilde{\eta}_{m_l}| = 0.$$

Because there is at least one real positive zero ω_l of g between $\eta_{n_l-m_l-1}$ and $\tilde{\eta}_{m_l}$, we see that

$$\lim_{l\to\infty} |\cos(\omega_l s_1) - \cos(\tilde{\eta}_{m_l} s_1)| = 0. \tag{3.51}$$

Applying, in order, (3.51), (3.47), (3.50), and (3.45), we see that

$$\lim_{l\to\infty} |\cos(\omega_l s_1)| = \lim_{l\to\infty} |\cos(\tilde{\eta}_{m_l} s_1)| = \lim_{l\to\infty} |\cos \pi(1/2 + m_l)| = \lim_{l\to\infty} |\cos(\pi n_l s_1)| = 0.$$

This finishes the proof of part (b). □

In the proof of Lemma 3.5 it is shown that the positive zeros of $g(\omega)$ are countable. We append to this set the values $\{\pi n\}$, where n satisfies $\cos(\pi n s_1) = 0$, and label them $\{\omega_k\}_{k\in\mathbf{Z}^+}$, where we assume that the sequence is increasing. Recall from Lemma 3.2 that these are the nonzero values of ω for which $\pm i\omega^2$ are eigenvalues.

The following lemma is useful in the proofs of Lemmas 3.7 and 3.9.

LEMMA 3.6. $\{\omega_k\}_{k\in\mathbf{Z}^+}$ *satisfies condition* (2.18).

Proof. For some values of s_1 all of the η_k's and $\tilde{\eta}_j$'s, as defined in (3.47), are distinct, whereas for some values of s_1 there are nonnegative integers k and j such that $\eta_k = \tilde{\eta}_j$. It is easy to verify that the latter values of s_1 are the set of rational numbers in [0,1] of the form n/m, where n is an odd integer and m is an even integer. If s_1 is of this form, let l be any odd integer, and we see that if $k = (l(m-n)-1)/2$ and $j = (nl-1)/2$, then $\eta_k = \tilde{\eta}_j = \pi lm/2$. It is also easy to see that these are the only nonnegative integer values of j and k such that $\eta_k = \tilde{\eta}_j$. The integers $lm/2$ are precisely the integers ω such that $\cos(\pi\omega s_1) = 0$. The statement "between any two adjacent elements of Λ (see (3.48)) there is exactly one zero of g" (see (3.15)) was proved in the proof of Lemma 3.5. This statement can be extended to "between any two adjacent elements of Λ there is exactly one element of $\{\omega_k\}$," interpreted to mean that if two elements of Λ are equal, then the number "between" them is the common value.

In a closed interval of length M, there are at most $1 + M(1-s_1)/\pi$ values of η_k, and at most $1 + Ms_1/\pi$ values of $\tilde{\eta}_j$. Therefore there are at most $2 + M/\pi$ elements of Λ (not necessarily distinct) in an interval of length M. In light of the above paragraph, we see that there are at most $3 + M/\pi$ values of ω_k in an interval of length M. Therefore, ω_{k+4} must be outside of the interval $[0, k\pi]$, that is,

$$\omega_{4+k} > k\pi,$$

or

$$\omega_k > k\pi - 4\pi. \tag{3.52}$$

In a closed interval of length M, there are at least $-1 + M(1-s_1)/\pi$ values of η_k and at least $-1 + Ms_1/\pi$ values of $\tilde{\eta}_j$. Therefore there are at least $-2 + M/\pi$ elements of Λ (not necessarily distinct) in an interval of length M. We then see that there are at least $-3 + M/\pi$ values of ω_k in an interval of length M. Hence ω_{k-3} must be inside of the interval $[0, k\pi]$, that is,

$$\omega_{k-3} \leq k\pi$$

or

$$\omega_k \leq k\pi + 3\pi.$$

Combining this with (3.52) completes the proof of the lemma. □

We can now prove the admissibility of the input and observation operators.

LEMMA 3.7. *For any s_1, the input operator B and the observation operator B^* are admissible for $S(t)$.*

Proof. It follows from (3.38) that for positive zeros ω of g,

$$\lim_{\omega\to\infty} |C_\omega - [s_1 \sec^2(\omega s_1) + (1-s_1)\sec^2 \omega(s_1 - 1)]^{-1/2}| = 0. \tag{3.53}$$

Therefore $\{C_\omega\}$ is a bounded sequence. It is then clear from (3.40) for b_ω that $\{b_\omega\}$ is a bounded sequence. Taking (3.39) into account, we see that the input coefficients are all bounded.

To conclude that B is an admissible input operator, we use the Carleson measure criterion found in Ho and Russell [9] or Weiss [20]. This criterion applies here if the eigenvalues are on the imaginary axis and Condition 1, defined before Lemma 2.12, is satisfied. In the proof of Lemma 3.6 we see that the number of ω_k's in an interval of length M is at most $3 + M/\pi$, so the number of eigenvalues of A in

$$S_{a,M} = \{z \,|\, \mathrm{Re}(z) = 0, a^2 \le \mathrm{Im}(z) \le (a+M)^2\}$$

is less than or equal to $3 + M/\pi$. This shows that Condition 1 preceding Lemma 2.12 is satisfied, so B is an admissible input operator. By duality, B^* is an admissible observation operator. □

We now see how condition (1.8) effects the behavior of the input coefficients $\{b_\omega\}$.

LEMMA 3.8. (a) *Suppose* (1.8) *is true. Then there exists $m > 0$ such that for all positive zeros ω of g,*

$$|b_\omega| > m. \tag{3.54}$$

(b) *Suppose there exists a sequence of positive integers $\{n_l\}$ such that* (3.45) *is true. Then there exists a sequence of zeros $\{\omega_k\}$ of g such that*

$$\lim_{k\to\infty} |b_{\omega_k}| = 0. \tag{3.55}$$

Proof. We first prove part (a). Because (1.8) is true, (3.42) and (3.43) are true. Combining this with (3.53), we see that $\{C_\omega^{-1}\}$ is a bounded sequence, so $\{C_\omega\}$ is bounded below away from zero. Because $(\tanh(\omega s_1) + \tanh \omega(1-s_1)) \to 2$ as $\omega \to \infty$ along the real axis, we see from (3.40) that (3.54) is true.

To prove part (b), note that if (3.45) is true, then part (b) of Lemma 3.5 gives a sequence $\{\omega_k\}$ such that (3.46) is true. Therefore, we see from (3.53) that $\{C_{\omega_k}^{-1}\}$ is an unbounded sequence. Because $\{\tanh(\omega_k s_1) - \tanh \omega_k(s_1 - 1)\}$ is a bounded sequence, we see from (3.40) that (3.55) is true. □

We are now in a position to follow the approach given in the proof of Theorem 1.1. Because many of the details are the same, we frequently refer to the corresponding proofs in §2.

LEMMA 3.9. (1) *The feedthrough for the system* (1.2), (1.6), (3.6), (3.10) *is* 0.

(2) *(B^*, A_k, B) is regular.*

(3) $A_k = A - kBB_L^*$.

Proof. We first show that the system (1.2), (1.6), (3.5), (3.6) is regular. We have already shown that B and B^* are admissible for $S(t)$. As in the proof of Lemma 2.6, we obtain the formula for $B_L^*R(\lambda, A)B$ by using the matrix representations of A and B in the basis $\{\Phi_k\}_{k\in I}$, so

$$B_L^*R(\lambda, A)B = \frac{|b_0|^2}{\lambda} + \sum_{k=1}^{\infty}\left\{\frac{|b_{\omega_k}|^2}{\lambda - i\omega_k^2} + \frac{|b_{\omega_k}|^2}{\lambda + i\omega_k^2}\right\}.$$

Because Lemma 3.6 shows that $\{\omega_k\}$ satisfies condition (2.18), we can apply Theorem 2.7 to conclude that the system (1.2), (1.6), (3.5), (3.6) is regular in the sense of [23] (with the feedthrough undetermined so far), and in fact satisfies (2.19).

We can compute the transfer function $\mathbf{H}(\lambda)$ of this system by taking the Laplace transform of (1.2), (1.6), (3.5), (3.6). Recall that in this paper $\lambda = i\omega^2$, where we can restrict ω to $\{\mathrm{r}e^{i\theta} \mid r \geq 0,\ \theta \in [-\pi/2, 0]\}$.

Let

$$p(s_{\mathrm{I}}, \omega) = \frac{\sinh(\omega s_{\mathrm{I}}) - \cosh(\omega s_{\mathrm{I}}) \tanh \omega(s_{\mathrm{I}} - 1)}{\cos(\omega s_{\mathrm{I}}) \tan \omega(s_{\mathrm{I}} - 1) - \sin(\omega s_{\mathrm{I}})}. \tag{3.56}$$

We find that

$$\mathbf{H}(\lambda) = \frac{-2i}{\omega}\left\{\frac{\sinh(\omega s_{\mathrm{I}}) - \cosh(\omega s_{\mathrm{I}}) \tanh \omega(s_{\mathrm{I}} - 1)}{\cosh(\omega s_{\mathrm{I}}) + p(s_{\mathrm{I}}, \omega)\cos(\omega s_{\mathrm{I}})}\right\}. \tag{3.57}$$

From this it is easy to see that $\mathbf{H}(\lambda) \to 0$ as $\lambda \to \infty$ along the positive real axis. Therefore, the system (1.2), (1.6), (3.5), (3.6) is regular in the sense of Definition 2.1 and can be represented by (B^*, A, B).

As in the proof of Lemma 2.6, if $J \in \mathcal{B}(U)$, then $A + BJB_L^*$ is the generator of a C_0-semigroup, and the closed-loop system obtained by letting $u(t) = Jy(t)$ is regular in the sense of Definition 2.1 and can be represented by $(B_L^*, A - BJB_L^*, B)$. Also as in the proof of Lemma 2.6, $A - kBB_L^* = A_k$, so we can take $J = -kI$ to finish the proof of Lemma 3.9. □

The proof of Lemma 2.8 applies to this system, so $\mathbf{H}_k(\lambda)$ is bounded on the imaginary axis.

LEMMA 3.10. *If* (1.8) *is true, then* (A_k, B) *is stabilizable as in Definition* 2.3, *and* (B^*, A_k) *is detectable as in Definition* 2.4.

Proof. This lemma is proved in exactly the same way as Lemma 2.10. Therefore, rather than give the proof, we explain why the proof of Lemma 2.10 also goes through in this case. That proof relied on three properties of A that are needed to be able to apply results in [15]. These follow from Lemmas 3.2, 3.6, and 3.8 and Proposition 3.3.

(1) The eigenvectors of A form a Riesz basis for X.

(2) The eigenvalues of A are of multiplicity 1 and grow quadratically on the imaginary axis.

(3) The input coefficients for (A, B) are bounded above and bounded below away from zero. □

Therefore, if (1.8) is true, then an application of Theorem 2.5 shows that A_k is exponentially stable. To finish the proof of Theorem 1.3, suppose (1.8) is not true, so (3.41) is not true by Lemma 3.5. There are then two possibilities. If $\cos(\pi m s_{\mathrm{I}}) = 0$ for some positive integer m, then part (b) of Lemma 3.2 and (3.39) show that $\lambda = \pm i(\pi m)^2$ are eigenvalues with associated input coefficients $b_{\pm m} = 0$. The other possibility is that there exists a sequence of positive integers $\{n_l\}$ such that (3.45) is true. Then part (b) of Lemma 3.8 gives a sequence of eigenvalues $\{\pm i\omega_k^2\}$ with associated input coefficients satisfying (3.55). In either case we see that (3.54) is not true. Therefore, (2.23) cannot be satisfied. Because the eigenvalues of A are shown in the proof of Lemma 3.7 to satisfy Condition 1 stated before Lemma 2.12, Lemma 2.12 shows that $A_k = A - kBB^*$ cannot generate an exponentially stable semigroup. This finishes the proof of Theorem 1.3.

4. Appendix.

In this Appendix we prove Lemma 2.11 and Theorem 2.7.

We start with Lemma 2.11. The result seems almost obvious formally, because $(A + BKC_L) + BK(C_L' - C_L) = A + BKC_L'$ and K is an admissible feedback for (C_L', A, B).

However, to prove the result rigorously we need to show that $(C'_L - C_L, A + BC_L, B)$ satisfies Definition 2.2 with feedback K.

Proof of Lemma 2.11. Let $\mathbf{H}_1$ be the transfer function for (C, A, B) and $\mathbf{H}_2$ be the transfer function for (C', A, B). Let $\mathbf{H}_1^K$ be the closed-loop transfer function for (C, A, B) under the feedback $u = Ky + v$, that is, the transfer function for the system

$$\dot{x}(t) = (A + BKC_L)x(t) + Bv(t) \tag{4.1}$$

$$y(t) = C_L x(t). \tag{4.2}$$

Then it is shown in [24] that

$$\mathbf{H}_1^K = \mathbf{H}_1(I - K\mathbf{H}_1)^{-1}, \tag{4.3}$$

where all inverses are in H_∞^∞.

Let $\mathbf{H}$ be the transfer function for $(C'_L - C_L, A + BKC_L, B)$, so $\mathbf{H}$ can be written as $\mathbf{H}_3 - \mathbf{H}_1^K$, where $\mathbf{H}_3$ is the transfer function for $(C'_L, A + BKC_L, B)$. We need to show that $(I - K\mathbf{H})$ is invertible in H_∞^∞. To do this we need to find a formula for $\mathbf{H}_3$ in terms of $\mathbf{H}_2$ and $\mathbf{H}_1$.

Consider the system

$$\dot{x}(t) = Ax(t) + Bu(t),$$

$$y(t) = [C_L, C'_L]^T x(t).$$

The transfer function for this system is $\tilde{\mathbf{H}} = [\mathbf{H}_1, \mathbf{H}_2]^T$. It is easy to see that $\tilde{K} = [K, 0]$ is an admissible feedback for $\tilde{\mathbf{H}}$, because $(I - \tilde{K}\tilde{\mathbf{H}}) = I - K\mathbf{H}_1$, which is invertible by hypothesis. The closed-loop system with the feedback $u = \tilde{K}y + v$ is then

$$\dot{x}(t) = (A + BKC_L)x(t) + Bv(t),$$

$$y(t) = [C_L, C'_L]x(t),$$

which has the transfer function $\tilde{\mathbf{H}}^{\tilde{K}}$. Because $\mathbf{H}_1^K$ is the transfer function for (4.1), (4.2), $\tilde{\mathbf{H}}^{\tilde{K}} = [\mathbf{H}_1^K, \mathbf{H}_3]$. Using Proposition 3.6 in [24], we see that the admissibility of $\tilde{K}$ for $\tilde{\mathbf{H}}$ implies that

$$\tilde{\mathbf{H}}^{\tilde{K}} - \tilde{\mathbf{H}} = \tilde{\mathbf{H}}^{\tilde{K}}\tilde{K}\tilde{\mathbf{H}}.$$

Writing down the second component in this identity gives

$$\mathbf{H}_3 - \mathbf{H}_2 = \mathbf{H}_3 K\mathbf{H}_1.$$

Solving this for $\mathbf{H}_3$, we obtain

$$\mathbf{H}_3 = \mathbf{H}_2(I - K\mathbf{H}_1)^{-1}.$$

Combining this with (4.3) we see that

$$\mathbf{H} = \mathbf{H}_3 - \mathbf{H}_1^K = (\mathbf{H}_2 - \mathbf{H}_1)(I - K\mathbf{H}_1)^{-1}. \tag{4.4}$$

Let

$$\mathbf{H}^K = (\mathbf{H}_2 - \mathbf{H}_1)(I - K\mathbf{H}_2)^{-1}, \tag{4.5}$$

where we are using the fact that K is an admissible feedback for (C_L', A, B), so $(I - K\mathbf{H}_2)$ is invertible. It is now easy but tedious to use (4.4) and (4.5) to verify algebraically that

$$(I - K\mathbf{H})^{-1} = (I + K\mathbf{H}^K).$$

This shows that $(I - K\mathbf{H})$ is invertible in H_∞^∞, so K is an admissible feedback for $\mathbf{H}$ according to Definition 2.2, finishing the lemma. □

Proof of Theorem 2.7. Because $\{d_k\}_{k\in I} \in l^\infty$, there exists $M > 0$ such that

$$|\mathbf{H}(s)| \leq M \sum_{k=1}^{\infty} \left| \frac{1}{s - i\omega_k^2} \right| + M \sum_{k=1}^{\infty} \left| \frac{1}{s + i\omega_k^2} \right| .$$

We analyze the first sum and note that the second sum can be analyzed in the same way. Let a, b, and m be as in the statement of Theorem 2.9, and let K be the smallest nonnegative integer greater than or equal to $-b/m$. So if $k > K$, then $(mk + a)$ and $(mk + b)$ are both positive. We can write the first sum above as

$$\sum_{k=1}^{K} \left| \frac{1}{s - i\omega_k^2} \right| + \sum_{k=K+1}^{\infty} \left| \frac{1}{s - i\omega_k^2} \right|, \tag{4.6}$$

where the first finite sum (interpreted as 0 if $K = 0$) clearly satisfies (2.19). To analyze the second sum in (4.6), let $s = x + iy$, where $x > 0$. The second sum is then less than or equal to

$$\frac{1}{\sqrt{2}} \sum_{k=K+1}^{\infty} \frac{1}{x + |y - \omega_k^2|}. \tag{4.7}$$

We show that this sum has a bound that is independent of y and dependent of x. If $y \geq (m(K+1) + a)^2$, let $K_1(y)$ be the largest integer less than or equal to $(\sqrt{y} - a)/m$. If $y < (m(K+1) + a)^2$, let $K_1(y) = K$. Then

$$y - \omega_k^2 \geq y - (mk + a)^2 \geq 0 \quad \text{for } K < k \leq K_1(y). \tag{4.8}$$

If $y \geq (mK + b)^2$, let $K_2(y)$ be the smallest integer greater than $(\sqrt{y} - b)/m$. If $y < (mK + b)^2$, let $K_2(y) = K + 1$. Then

$$y - \omega_k^2 \leq y - (mk + b)^2 < 0 \quad \text{for } k \geq K_2(y). \tag{4.9}$$

Now we see that (4.7) is equal to

$$\frac{1}{\sqrt{2}} \left(\sum_{k=K_1(y)+1}^{K_2(y)} \frac{1}{x + |y - \omega_k^2|} + \sum_{k=K+1}^{K_1(y)} \frac{1}{x + y - \omega_k^2} + \sum_{k=K_2(y)+1}^{\infty} \frac{1}{x + \omega_k^2 - y} \right)$$
$$=: \frac{1}{\sqrt{2}}(S_1 + S_2 + S_3),$$

where S_2 is interpreted as 0 if $K_1(y) = K$. To analyze S_1, note that there are $K_2(y) - K_1(y)$ terms in S_1, each of modulus less than or equal to $1/x$. Because $K_2(y) > (\sqrt{y} - b)/m + 1$ and $K_1(y) \leq (\sqrt{y} - a)/m - 1$, we see that

$$S_1 \leq (2 + (a - b)/m)/x. \tag{4.10}$$

We now analyze S_2. Note that if $y < (m(K+1)+a)^2$ then $S_2 = 0$, so we assume that $y \geq (m(K+1)+a)^2$. Using (4.8), we see that

$$S_2 \leq \sum_{k=K+1}^{K_1(y)} \frac{1}{x+y-(mk+a)^2}.$$

Therefore,

$$\begin{aligned} S_2 &\leq \int_{K+1}^{K_1(y)} \frac{d\eta}{x+y-(m\eta+a)^2} \\ &= \frac{1}{m}\int_{m(K+1)+a}^{mK_1(y)+a} \frac{d\zeta}{x+y-\zeta^2}. \end{aligned}$$

Because $mK_1(y)+a \leq \sqrt{y}$ and $m(K+1)+a > 0$, this is

$$\leq \frac{1}{m}\int_0^{\sqrt{y}} \frac{d\zeta}{x+y-\zeta^2}.$$

This integral is evaluated as

$$\frac{1}{m}\frac{1}{\sqrt{x+y}}\ln\left|\sqrt{1+\frac{y}{x}}+\sqrt{\frac{y}{x}}\right| = \frac{1}{m\sqrt{x}}\frac{1}{\sqrt{1+\psi}}\ln\left|\sqrt{1+\psi}+\sqrt{\psi}\right|,$$

where $\psi = y/x$. This is

$$\leq \frac{1}{m\sqrt{x}}\left(1+\sqrt{\psi/(1+\psi)}\right) \leq \frac{2}{m\sqrt{x}}$$

when $\psi > 0$, which is the case here, because $y \geq (m(K+1)+a)^2$ and $x > 0$. Therefore there exists a constant $M > 0$ independent of y such that

$$S_2 \leq \frac{M}{\sqrt{x}}. \tag{4.11}$$

We now analyze S_3. Using (4.9) we see that

$$\begin{aligned} S_3 &\leq \sum_{k=K_2(y)+1}^{\infty} \frac{1}{x+(mk+b)^2-y} \\ &\leq \int_{K_2(y)}^{\infty} \frac{d\eta}{x+(m\eta+b)^2-y} = \frac{1}{m}\int_{mK_2(y)+b}^{\infty} \frac{d\zeta}{x-y+\zeta^2}. \end{aligned} \tag{4.12}$$

To analyze this integral we consider several cases. If $x > y \geq 0$, we use the fact that $mK_2(y)+b > \sqrt{y}$, so this integral is

$$\begin{aligned} &\leq \frac{1}{m}\int_{\sqrt{y}}^{\infty} \frac{d\zeta}{x-y+\zeta^2} = \frac{1}{m\sqrt{x-y}}\left(\frac{\pi}{2} - \tan^{-1}\sqrt{\frac{y}{x-y}}\right) \\ &= \frac{1}{m\sqrt{x}}\frac{1}{\sqrt{1-1/\psi}}\left(\frac{\pi}{2} - \tan^{-1}\sqrt{\frac{1}{\psi-1}}\right), \end{aligned}$$

where $\psi = x/y > 1$. (If $y = 0$ the modifications are obvious.) If we let $g(\psi) = (\frac{\pi}{2} - \tan^{-1}((\psi - 1)^{-1/2}))/\sqrt{1 - 1/\psi}$, it is easy to use L'Hôpital's rule to see that $g(\psi)$ is bounded on $(1, \infty)$. Therefore, for $x > y \geq 0$, there exists $M > 0$ such that

$$S_3 \leq \frac{M}{\sqrt{x}}. \tag{4.13}$$

If $x > y$ and $y < 0$, then the last integral in (4.12) is less than or equal to

$$\frac{1}{m}\int_0^\infty \frac{d\zeta}{x - y + \zeta^2} = \frac{\pi}{2m\sqrt{x - y}} \leq \frac{\pi}{2m\sqrt{x}},$$

so in this case (4.13) is also true.

If $x = y$, then the last integral in (4.12) is

$$= \frac{1}{m(mK_2(y) + b)} \leq \frac{1}{m\sqrt{y}} = \frac{1}{m\sqrt{x}},$$

so (4.13) is true in this case.

If $x < y$, then the last integral in (4.12) is less than or equal to

$$\frac{1}{m}\int_{\sqrt{y}}^\infty \frac{d\zeta}{x - y + \zeta^2} = \frac{1}{m\sqrt{y - x}} \ln\left|\sqrt{\frac{y}{x}} + \sqrt{\frac{y}{x} - 1}\right|.$$

Letting $\psi = y/x > 1$, we see that this is

$$= \frac{1}{m\sqrt{x}}\frac{1}{\sqrt{\psi - 1}} \ln\left|\sqrt{\psi} + \sqrt{\psi - 1}\right|.$$

It is easy to use L'Hôpital's rule to show that $g(\psi) = \ln|\sqrt{\psi} + \sqrt{\psi - 1}|/\sqrt{\psi - 1}$ is bounded on $(1, \infty)$, so we see that in this case (4.13) is again true. Therefore, there exists an M such that for all $x > 0$, (4.13) is true.

Combining (4.10), (4.11), and (4.13), we see that for any $\alpha > 0$, the second sum in (4.6) is bounded by $M/\sqrt{x}$ for some M and all $x \geq \alpha$. This shows that the conclusions of Theorem 2.7 are true, finishing the proof. □

Acknowledgments. The author thanks Dr. Stuart Townley and Dr. George Weiss for helpful discussions and correspondence.

REFERENCES

[1] R. A. Adams, *Sobolev Spaces*, Academic Press, New York, 1975.

[2] C. Byrnes and D. Gilliam, *Stability of certain distributed parameter systems by low dimensional controllers: A root locus approach*, Proc. 29th Conference on Decision and Control, Honolulu, HI, 1990, pp. 1871–1872.

[3] G. Chen, M. C. Delfour, A. M. Krall, and G. Payre, *Modeling, stabilization and control of serially connected beams*, SIAM J. Control Optim., 25 (1987), pp. 526–546.

[4] G. Chen, S. G. Krantz, D. W. Ma, C. E. Wayne, and H. H. West, *The Euler–Bernoulli beam equation with boundary energy dissipation*, in Operator Methods for Optimal Control Problems, Sung J. Lee, ed., Marcel Dekker, New York, 1988, pp. 67–96.

[5] G. Chen, S. G. Krantz, D. L. Russell, C. E. Wayne, H. H. West, and M. P. Coleman, *Analysis, design and behavior of dissipative joints for coupled beams*, SIAM J. Control Optim., 49 (1989), pp. 1665–1693.

[6] F. Conrad, *Stabilization of beams by pointwise feedback control*, SIAM J. Control Optim., 28 (1990), pp. 423–437.

[7] R. F. Curtain, *Equivalence of input-output stability and exponential stability*, Systems Control Lett., 12 (1989), pp. 235–239.

[8] R. F. Curtain and G. Weiss, *Well-posedness of triples of operators (in the sense of linear systems theory)*, in Distributed Parameter Systems, F. Kappel, K. Kunisch, and W. Schappacher, eds., Proc. of the Conf. in Vorau, Austria, July 1988, Birkhäuser, Basel, 1989.

[9] L. F. Ho and D. L. Russell, *Admissible input elements for systems in Hilbert space and a Carleson measure criterion*, SIAM J. Control Optim., 21 (1983), pp. 614–640.

[10] G. H. Hardy and E. M. Wright, *An Introduction to the Theory of Numbers*, 5th ed., Clarendon Press, Oxford, UK, 1979.

[11] F. Huang, *Characteristic conditions for exponential stability of linear dynamical systems in Hilbert space*, Ann. Differential Equations, 1 (1985), pp. 43–56.

[12] K. S. Liu, *Energy decay problems in the design of a point stabilizer for coupled string vibrating systems*, SIAM J. Control Optim., 26 (1988), pp. 1348–1356.

[13] K. S. Liu, F. L. Huang, and G. Chen, *Exponential stability analysis of a long chain of coupled vibrating strings with dissipative linkage*, SIAM J. Appl. Math., 49 (1989), pp. 1694–1707.

[14] J. Prüss, *On the spectrum of C_0-semigroups*, Trans. Amer. Math. Soc., 284 (1984), pp. 847–857.

[15] R. Rebarber, *Spectral assignability for distributed parameter systems with unbounded scalar control*, SIAM J. Control and Optim., 27 (1989), pp. 148–169.

[16] ———, *Conditions for the equivalence of internal and external stability for distributed parameter systems*, IEEE Trans. Automat. Control, 38 (1993), pp. 994–998.

[17] R. Rebarber and S. Townley, *High gain robustness of distributed parameter systems*, Proc. 1993 Conference on the Mathematical Theory of Networks and Systems, Regensburg, Germany, August, 1993.

[18] D. Salamon, *Infinite-dimensional linear systems with unbounded control and observation: A functional analytic approach*, Trans. Amer. Math. Soc., 300 (1987), pp. 383–431.

[19] G. Weiss, *Admissible observation operators for linear semigroups*, Israel J. Math., 65 (1989), pp. 17–43.

[20] ———, *Admissibility of input elements for diagonal semigroups on l^2*, Systems Control Lett., 10 (1988), pp. 79–82.

[21] ———, *Admissibility of unbounded control operators*, SIAM J. Control Optim., 27 (1989), pp. 527–545.

[22] ———, *The representation of regular linear systems on Hilbert spaces*, Proc. Conference on Distributed Parameter Systems, Vorau, Austria, July, 1988, Birkhäuser, Basel, 1989.

[23] ———, *Transfer functions of regular linear systems, part* I: *Characterizations of regularity*, Trans. Amer. Math. Soc., to appear.

[24] ———, *Regular linear systems with feedback*, Math. Control Signals Sys., to appear.

[25] J. Zabczyk, A note on C_0-semigroups, Bull. Acad. Pol. Sci. Serie Math., 23 (1985), pp. 895–898.

[26] H. Zwart, *Some remarks on open- and closed-loop stabilizability for infinite-dimensional systems*, in Control and Estimation of Distributed Parameter Systems, International Series on Numerical Mathematics, Vol. 91, F. Kappel, K. Kunisch, and W. Schappacher, eds., Birkhäuser-Verlag, Basel, 1989, pp. 425–434.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 33, No. 1, pp. 29–66, January 1995

© 1995 Society for Industrial and Applied Mathematics
002

DYNAMIC PROGRAMMING AND PRICING OF CONTINGENT CLAIMS IN AN INCOMPLETE MARKET*

NICOLE EL KAROUI† AND MARIE-CLAIRE QUENEZ†

Abstract. The problem of pricing contingent claims or options from the price dynamics of certain securities is well understood in the context of a complete financial market. This paper studies the same problem in an incomplete market. When the market is incomplete, prices cannot be derived from the absence of arbitrage, since it is not possible to replicate the payoff of a given contingent claim by a controlled portfolio of the basic securities. In this situation, there is a price range for the actual market price of the contingent claim. The maximum and minimum prices are studied using stochastic control methods.

The main result of this work is the determination that the maximum price is the smallest price that allows the seller to hedge completely by a controlled portfolio of the basic securities. A similar result is obtained for the minimum price (which corresponds to the purchase price).

Key words. option pricing, incomplete market, equivalent martingale measures, portfolio processes, stochastic control

AMS subject classifications. 90A09, 90C39, 93E25

Introduction. We study the problem of determining the price of a contingent claim from the price dynamics of certain securities (such as stocks and bonds). However, the price system alone cannot give a complete description of the exogenous uncertain environment; other information that is inside or outside the market is available and might influence market fluctuations. Therefore, the information structure used is as general as possible; in particular, it is not supposed to be generated by Brownian motions.

The primitive securities consist of a bond and n stocks, the latter being driven by a d-dimensional Brownian motion. Absence of arbitrage is assumed. The fluctuation in those prices is linked to the rest of the market (market fluctuations, change in rates, prices of other securities, and so on) and to other factors that are outside the market. The contingent claim is not linked only to the basic securities.

In §1, we formulate the basic problem of hedging by constructing a portfolio of the basic securities that attains (at least) the payoff of the contingent claim as its terminal wealth. Unlike in the complete market case, it is not possible to replicate the payoff of every contingent claim by a portfolio, and there are several probability measures that are equivalent to the initial probability, such that the discounted price processes are martingales. Several price systems are associated by duality to those martingale measures. Thus in this situation there is a price range for the actual market price of the contingent claim.

In §2, we study the maximum price using stochastic control methods. We show that the maximum price can be written as the difference of the (discounted) value of a portfolio and an optional increasing process that is equal to zero at time zero. We then state that the maximum price is the selling price defined as the smallest price that allows the seller to hedge completely by a controlled portfolio of the basic securities. A similar result is obtained for the minimum price (which corresponds to the purchase price). We also state that the class of contingent claims for which the supremum price is obtained for an optimal martingale measure is exactly the set of attainable claims.

In §3, we give a few methods for computing the maximum price (approximation methods, use of Bellman equations in the Markovian case). We also give two examples that illustrate the obtained results.

* Received by the editors June 1, 1992; accepted for publication (in revised form) July 16, 1993.

† Laboratoire de Probabilités, Université Paris VI, 4 place Jussieu, Tour 56, $3^{ème}$ étage, 75252 Paris Cedex 05, France.

1. Formulation of the problem.

1.1. The model. The basic securities consist of $n + 1$ assets; they are the only assets that are available to agents for trading. One of them is a nonrisky asset (the bond), with price-per-unit $P_0(t)$ governed by the equation

$$dP_0(t) = P_0(t)r(t)dt. \tag{1}$$

The interest rate $r(t)$ is positive and bounded.

In addition to the bond, there are n risky securities (stocks). The price $P_i(t)$ for one share of the ith stock is modeled by the linear stochastic differential equation

$$dP_i(t) = P_i(t)\left[b_i(t)dt + \sum_{j=1}^{d} \sigma_{i,j}(t)dW_t^j\right]. \tag{2}$$

The information structure is modeled by a filtration $(F_t,\ 0 \leq t \leq T)$ that satisfies the standard hypotheses and is left quasi-continuous. The coefficients of the model r, b_i, $\sigma_{i,j}$ are taken to be predictable with respect to $\{F_t\}$. $W = (W_1, \ldots, W_d)^*$ is a d-dimensional (F_t)-Brownian motion under P, with the "objective" probability taken as a primitive. We suppose that $n \leq d$.

To be a reasonable model of securities markets, the prices of the basic securities should not allow one to create something out of nothing or to create free lunches. Thus, we will suppose the existence of d coefficients $\theta_1, \theta_2, \ldots, \theta_d$ that are (F_t)-predictable processes such that

$$b_i(t) - r(t) = \sum_{j=1}^{d} \sigma_{ij}(t)\theta_j(t), \quad P\text{ a.s.}, \quad 1 \leq i \leq n.$$

$\theta_j(t)$ represents the risk premium associated with the source of uncertainty W_j; we suppose that θ_j is bounded.[1]

We adopt the following notation and make the following assumptions:

- We denote by b the column vector of stock appreciation rates $b = (b_1, \ldots, b_n)^*$.
- For $1 \leq i \leq n$, let σ_i be the volatility row vector of the ith stock $\sigma_i = (\sigma_{i,1}, \ldots, \sigma_{i,n})$
- Let $\sigma(t)$ be the volatility $(n \times d)$ matrix whose rows are $\sigma_1(t), \ldots, \sigma_n(t)$. $\sigma(t)$ is supposed to be bounded.
- Let θ_t be the relative risk column vector $\theta = (\theta_1, \theta_2, \ldots, \theta_d)^*$.

If we denote by $\mathbf{1}$ the vector whose every component is 1, the above equations are written as

$$b(t) - r_t\mathbf{1} = \sigma_t\theta_t, \quad P\text{ a.s.}$$

Furthermore, we will suppose that there exists a predictable vector α_t for which $\theta_t = \sigma_t^*\alpha_t$. (This hypothesis, which is not restrictive, will be explained in §1.8.)

Remark. Notice that if σ_t is taken to be injective, the above hypotheses are satisfied.

In the next section, we recall the characteristics of portfolios.

[1] For information concerning the equivalence between the assumption of no arbitrage and the existence of θ, one is referred to [Fö-Sc], [H-J-M], [An-St]. We thank the reviewers for the following references concerning this point: [Ch-Mu], [He-Pe], [Sch].

1.2. The portfolios. Let us consider an investor who can invest in the $n+1$ basic securities. At time 0, he invests the amount $x \geq 0$ in the $n+1$ securities. We shall denote by $X(t)$ the value of the amount invested at time t in the $n+1$ securities; $X(0) = x$. For $i = 1, \dots, d$, we denote by $\pi_i(t)$ the amount that he invests in the ith stock at time t.

DEFINITION 1.2.1. *A portfolio strategy* $\pi(t) = (\pi_1(t), \dots, \pi_n(t))$; $0 \leq t \leq T$, *is an* R^n*-valued process that is predictable with respect to* (F_t) *and satisfies*

$$\int_0^T \|\sigma_t^* \pi_t\|^2 \, dt < +\infty, \quad P\text{a.s.}$$

The amount invested in the bond at time t is then given by

$$X(t) - \sum_{i=1}^n \pi_i(t).$$

Let $N(t)$ be the cumulative total sum of additional amounts that have been invested by the agent in the $n+1$ securities between 0 and t; N is a right-continuous left-limited (RCLL), (F_t)-optional process that satisfies $N_0 = 0$.

The value of the portfolio $X(t)$ is given by

$$dX_t = \sum_{i=1}^n \pi_i(t)[b_i(t) + \sigma_i(t) dW_t] + \left(X_t - \sum_{i=1}^n \pi_i(t)\right) r(t)dt + dN_t,$$

or equivalently,

$$dX_t = r_t X_t \, dt + \pi_t^*(b_t - r_t \mathbf{1})dt + \pi_t^* \sigma_t \, dW_t + dN_t.$$

Remark. Loosely speaking, dN_t is the amount saved (or consumed if negative) during the time period $[t, \, t + dt]$.

We denote by $X^{\pi,N,x}(t)$ the value of the portfolio corresponding to the strategy (π, N) and the initial investment x.

- If $N = 0$, the portfolio is called *self-financing*, then $X^{\pi,0,x}(t)$ is the value at time t of the self-financing portfolio corresponding to the initial investment x and the portfolio π.
- If $N_t = -C_t$, where $(C_t, 0 \leq t \leq T)$ is an RCLL, (F_t)-optional increasing process that satisfies $C_0 = 0$, the portfolio is called a *portfolio strategy with consumption*; $X^{\pi,-C,x}(t)$ is the value at time t of the portfolio corresponding to the initial investment x, the portfolio π, and the process C_t that represents the cumulative amount the agent withdraws up to time t for consumption.
- If $N_t = D_t$ where $(D_t, 0 \leq t \leq T)$ is an RCLL, (F_t)-optional increasing process that satisfies $D_0 = 0$, the portfolio is called a *portfolio strategy with savings*; $X^{\pi,D,x}(t)$ is the value at time t of the portfolio corresponding to the initial investment x, the portfolio π, and the process D_t that represents the cumulative amount the agent adds in the portfolio up to time t.

1.3. Contingent claim B, selling price. Let T, a positive constant, be the terminal time for the problem.

DEFINITION 1.3.1. *A* contingent claim *B is a nonnegative, F_T-measurable random variable. It can be thought of as a contract or agreement that pays B at maturity T.*

The problem is to price this contingent claim. Let us consider a seller who wants to sell some contingent claim with payoff B and maturity T, between time 0 and time T. Suppose the seller chooses the price Y_t for B at any time t; more precisely, the seller must choose his price process $(Y_t,\ t \geq 0)$ (i.e., an RCLL optional nonnegative process that satisfies $Y_T = B$). Also, the seller does not want to run any risk of losing money. Therefore, he will only choose price processes that allow him to hedge completely by a controlled portfolio of the basic securities in the sense that, if at any time t, he sells the option B, and if at a later date, he buys it back, he wants to make a profit. More precisely, suppose that at time t, the seller sells the contingent claim at the price Y_t. He then invests this amount in the self-financing portfolio determined by π_t. At time $t + dt$, he buys back the contingent claim at the price Y_{t+dt} and sells the portfolio; he then makes a profit equal to

$$r_t Y_t\, dt + \pi_t^* \sigma_t (dWt + \theta_t\, dt) - dY_t.$$

Hence, we have the following definition.

DEFINITION 1.3.2. *A process Y_t is called a* price admissible for sellers *if Y_t is an* RCLL *nonnegative optional process that satisfies $Y_T = B$ and such that there exist a portfolio process π_t and an* RCLL, *(F_t)-optional increasing process $(C_t,\ 0 \leq t \leq T)$ such that $C_0 = 0$ and $dY_t = r_t Y_t\, dt + \pi_t^*\, \sigma_t (dWt + \theta_t\, dt) - dC_t$.*

Remark. It is equivalent to say that a process Y_t is a price admissible for sellers if there exists a hedging portfolio of B that is a portfolio with consumption (i.e., with withdrawals) whose value is equal to the price, that is, if there exists a portfolio process π_t and an RCLL, (F_t)-optional increasing process $(C_t,\ 0 \leq t \leq T)$ that satisfies $C_0 = 0$ and an initial investment x such that

$$X^{\pi,-C,x}(T) = B \quad \text{and} \quad Y_t = X^{\pi,-C,x}(t) \geq 0, \qquad 0 \leq t < T.$$

Note that every price admissible for sellers corresponds to one (and only one) hedging portfolio with consumption.

For the seller who sells the contingent claim B to an investor, such a strategy can be interpreted as follows:

- x is the price paid by the investor to the seller at time 0.
- π characterizes the hedging portfolio held by the seller.
- $C(t)$ represents the cumulative amount the seller withdraws from the hedging portfolio up to time t (which should be given to the investor, and since it is not, it is a profit for the seller).

Hence, it follows that x is the price paid not only for getting B at maturity T but also for getting additional amounts represented by the process $(C(t),\ 0 \leq t \leq T)$. Also, the seller's price will be the lowest price admissible for sellers. Thus, we define the selling price by Definition 1.3.3.

DEFINITION 1.3.3. *If it exists, the lowest price process admissible for sellers is called the* selling price.

Remark 1. We will see in §2 that such a process always exists, which is not obvious.

Remark 2. We could have defined the selling price as the essential infimum of the price processes admissible for sellers (but this would not define a stochastic process).

(Note that by symmetry, we can define the purchase price for B; it will be studied in §4.)

1.4. Discounting. Let β_t be the discount process given by

$$\beta_t = \exp\left\{-\int_0^t r_s\, ds\right\}, \qquad 0 \leq t \leq T.$$

We denote by $P_i^d(t)$, $\pi^d(t)$, $C^d(t)$ the discounted price process, the discounted portfolio process associated with the portfolio $\pi(t)$, and the discounted consumption process associated with the consumption $C(t)$, respectively. For $0 \le t \le T$, we have

$$\begin{aligned} P_i^d(t) &= \beta_t P_i(t) \qquad (1 \le i \le n),\\ \pi^d(t) &= \beta_t \pi(t),\\ C_t^d &= \int_0^t \beta_s \, dC_s. \end{aligned}$$

We have the following equation for the discounted price process:

$$dP_i^d(t) = P_i^d(t)[(b_i(t) - r(t))dt + \sigma_i(t)dW_t]. \tag{3}$$

The discounted value of the portfolio with withdrawals X^d associated with the portfolio-consumption strategy (π_t, C_t) is governed by the equation

$$dX_t^d = (\pi_t^d)^*[(b(t) - r(t)\,\mathbf{1})dt + \sigma(t)dW_t] - dC_t^d,$$

that is,

$$dX_t^d = (\pi_t^d)^* \sigma_t(\theta_t \, dt + dW_t) - dC_t^d. \tag{4}$$

Also, for a contingent claim B, we denote by B^d the discounted contingent claim

$$B^d = \beta_T B.$$

We now come back to the problem of pricing the contingent claim B from the price dynamics of the $n+1$ securities. We begin by recalling the theory of contingent claim valuation in the context of a complete market (see [Ha-Kr], [Ha-Pl], [Duf], and [Kar]). It is well known that the reference probability Q defined below has a fundamental role. Hereafter, we will often use the Girsanov theorem. (See §A.1, where we recall a general form of this theorem.)

1.5. The reference probability Q.

Notation. Let N_t be a local martingale (RCLL) under P with respect to $\{F_t\}$, such that $N_0 = 0$. We denote by $\mathcal{E}(N)_t$ the *exponential of N*, that is, the solution of the stochastic differential equation (SDE)

$$dZ_t = Z_{t^-} \, dN_t, \quad 0 \le t \le T, \quad Z_0 = 1.$$

This process is a local martingale under P.

Let $Z_0(t)$ be the exponential local martingale of

$$\left(- \int_0^t \theta_s^* \, dW_s, \, 0 \le t \le T \right)$$

that is,

$$Z_0(t) = \exp\left\{ - \int_0^t \theta_s^* \, dW_s - \frac{1}{2} \int_0^t \|\theta_s\|^2 \, ds \right\}.$$

Because θ is bounded, $Z_0(t)$, $0 \le t \le T$ is martingale under P.

We then define Q as the probability measure equivalent to P on F_T that admits the Radon–Nikodym derivative $Z_0(T)$. Let

$$\widetilde{W}_t = W_t + \int_0^t \theta_s \, ds, \qquad 0 \le t \le T.$$

By the Girsanov theorem, $(\widetilde{W}_t,\ 0 \leq t \leq T)$ is an (F_t)-Brownian motion under Q.

The SDE (3) relative to the discounted prices may be written as

$$dP_i^d(t) = P_i^d(t)\sigma_i(t)d\widetilde{W}_t. \tag{5}$$

Also, the discounted value X^d of the portfolio with withdrawals associated with the portfolio π_t, consumption C_t, and initial investment x is given by

$$X_t^d = x + \int_0^t (\pi_s^d)^* \sigma_s \, d\widetilde{W}_s - C_t^d, \quad Q\text{ a.s.}, \quad 0 \leq t \leq T. \tag{6}$$

Notice that the prices of the basic securities are Q-martingales and the prices (of the contingent claim) admissible for sellers are Q-supermartingales.

1.6. Pricing in a complete market. We review briefly in this section important results for use later in the treatment of the incomplete market case.

DEFINITION 1.6.1. *The security market is said to be* complete *if the filtration (F_t) is that generated by the Brownian motion W_t, $n = d$, and σ has full rank. It means that all the sources of uncertainty can be explained by the price dynamics of the basic securities. If it is not complete, the market will be called* incomplete.

Recall that if the market is complete, it is possible to construct a portfolio that attains as its final wealth any contingent chain B that is integrable under Q, that is, there exist some $x \geq 0$ and some portfolio (π_t) satisfying

$$B^d = x + \int_0^T (\pi_u^d)^* \sigma_u \, d\widetilde{W}_u, \quad P\text{ a.s.}$$

and such that the process defined by

$$x + \int_0^t (\pi_u^d)^* \sigma_u \, d\widetilde{W}_u, \qquad 0 \leq t \leq T$$

is a martingale under Q.

It is clear that $x = E_Q(B^d)$. This property allows us to derive the price for any contingent claim from an absence of arbitrage.

PROPOSITION 1.6.1. *In a complete market, every contingent claim (Q-integrable) is priced by arbitrage. This price is given by the expectation of the discounted contingent claim under Q, which is the unique probability measure under which the discounted prices of the basic stocks are martingales.*

Proof. Let P_B be the price for the contingent claim B. Suppose that P_B is strictly greater than $E_Q(B^d)$; then there exists some opportunity for arbitrage. For example, you can sell the contingent claim at $t = 0$ at the price P_B and invest the amount P_B in the hedging portfolio determined by π. At time T, you pay the amount B to your buyer and sell the portfolio, whose value is given by

$$\beta_T^{-1}\left(P_B + \int_0^T (\pi_u^d)^* \sigma_u \, d\widetilde{W}_u\right).$$

From an initial wealth equal to 0, at time T, you make a strictly positive profit equal to $(\beta_T)^{-1}(P_B - x)$. Also, if $P_B < E_Q(B^d)$, there exists an arbitrage opportunity. Hence, $P_B = E_Q(B^d)$.

The fact that Q is the unique probability that is equivalent to P such that the discounted price processes of the basic stocks are martingales can be easily proved using the Girsanov theorem. (The proof is similar to that of Proposition 1.8.1.) □

Remark 1. The price at time t for the contingent claim B can be determined by an absence of arbitrage and it is given by $E_Q[B^d/F_t]$.

Remark 2. It should be emphasized that the price system is associated by duality to the probability measure Q, which is equivalent to P and under which the discounted price processes are martingales.

Remark 3. In the case of a complete market, the arbitrage-free price coincides with the selling price. The hedging portfolio is a self-financing portfolio associated with the portfolio process π and the initial investment x. Clearly, if the market is complete the consumption processes are unnecessary since it is always possible to replicate the payoff of a given contingent claim by a self-financing portfolio. In the complete market contingent claim valuation, notice the fundamental roles of the construction of a hedging portfolio and the reference probability Q, which is the unique probability measure equivalent to P and under which the discounted price processes are martingales. Those remarks should be kept in mind when considering the more difficult case of an incomplete market.

We now turn to the consideration of an incomplete market. Recall that, contrary to the complete market case, the price system cannot suffice in itself to give a complete description of the environment. As a result, agents will not be able to replicate the payoff of every contingent claim by a self-financing portfolio of the basic securities and to price every contingent claim by arbitrage. Also, there exist several probability measures that are equivalent to P and under which the discounted price processes are martingales, associated by duality to different price systems.

1.7. The P-martingale measures and the attainable contingent claims.

DEFINITION 1.7.1. *Any probability measure that is equivalent to P on F_T and is such that the discounted price processes* (*of the basic claims*) *are martingales is called a P-martingale measure.*

We denote by $\mathcal{P}$ the set of all P-martingale measures. Notice that Q belongs to $\mathcal{P}$ and that, if the market is complete, $\mathcal{P} = \{Q\}$. If the market is incomplete, there are several P-martingale measures. Each martingale measure can naturally be associated by duality with a price system (for more details see [Ha-Kr]).

Assumption. Hereafter, the contingent claim B is supposed to be such that there exists a price admissible for sellers, or equivalently B is supposed to be smaller than the value of a self-financing portfolio, that is, B satisfies

$$B \leq X^{H,0,y}(T), \quad P\,\text{a.s.}$$

for some portfolio strategy H and initial investment $y \geq 0$. Also, we make the following technical assumption: $(X^{H,0,y})^d$ is supposed to be a square-integrable martingale under each P-martingale measure. This assumption will allow us to work with square integrable martingales but the main results of this paper remain true under the weaker hypothesis (see §A.3).

$$\sup_{R \in \mathcal{P}} E_R(B^d) < +\infty.$$

DEFINITION 1.7.2. *Every real that can be written $E_R(B^d)$, where R is a P-martingale measure, is called a* possible price at time $t = 0$ for B.

More generally, any random variable $E_R(B^d/F_t)$, where R is a P-martingale measure, is called a possible price at time t for B. However, there exists a class of contingent claims such that the price is unique, that is, the set of possible prices contains a unique element. We will show that this class is exactly the set of contingent claims that can be synthetized by a controlled portfolio of the basic securities (called "attainable").

DEFINITION 1.7.3. *A contingent claim B is said to be* attainable *if there exist some $x \geq 0$ and some portfolio process π such that*

$$B^d = x + \int_0^T (\pi_u^d)^* \sigma_u \, d\widetilde{W}_u, \quad P\, a.s.$$

and

$$E_Q\left[\int_0^T \|\sigma_s^* \pi_s^d\|^2 \, ds\right] < +\infty.$$

As in the complete market case, an attainable contingent claim B can be priced by arbitrage; therefore, $x = E_Q(B^d)$ is the arbitrage-free price at $t = 0$. Also, we state the following property.

PROPOSITION 1.7.1. *B is attainable if and only if $E_R(B^d)$ is constant over all $R \in \mathcal{P}$, that is, there is only one possible price.*

Remark. We show only one implication. The opposite implication follows from Theorem 2.3.2.

Proof. Let $x \geq 0$ and a portfolio process π such that

$$B^d = x + \int_0^T (\pi_u^d)^* \sigma_u \, d\widetilde{W}_u, \quad P \text{ a.s.}$$

and

$$E_Q\left[\int_0^T \|\sigma_s^* \pi_s^d\|^2 \, ds\right] < +\infty.$$

Let R be a P-martingale measure. The discounted prices $P_i^d(t)$ are martingales under R. Hence, the process defined by

$$x + \int_0^t \sum_{i=1}^n \pi_i^d(u)\sigma_i(u) d\widetilde{W}_u = x + \int_0^t \sum_{i=1}^n \frac{\pi_i^d(u)}{P_i^d(u)} \, dP_i^d(u), \qquad 0 \leq t \leq T$$

is clearly a local martingale under R. It is lower than $(X^{H,0,y})^d$, the discounted value of the portfolio (associated with H and y), because it is equal to $E_Q[B^d/F_t]$. Therefore, it is a martingale under R (by Proposition 1.a in the Appendix) because $(X^{H,0,y})^d$ is a martingale under R. Hence, $E_R(B^d) = x$. □

Remark. If B is attainable, then the price for B at $t \geq 0$ can be derived by an absence of arbitrage and $E_R[B^d/F_t]$ does not depend on $R \in \mathcal{P}$, that is, the set of possible prices contains only one element given by

$$x + \int_0^t \pi_u^{d^*} \sigma_u \, d\widetilde{W}_u.$$

If B is not attainable, there are several prices for B and B cannot be priced by arbitrage. Thus, it seems interesting to determine the bounds of the set of possible prices for B. At $t = 0$, the price for B is worth not less than $\inf_{R \in \mathcal{P}} E_R(B^d)$ and not more than $\sup_{R \in \mathcal{P}} E_R(B^d)$.

Using optimal control techniques, we shall study dynamically those maximum and minimum prices. In particular, we shall show that the supremum of the possible prices is equal to the selling price.

Before proceeding with the analysis of the maximum price, let us characterize the set of P-martingale measures. Recall that Pagès [Pag] has already characterized this set in the context of a Brownian model (see also [KLSX] for utility maximization problems in that context). Also, Ansel and Stricker [An-St] have shown that the market model contains no arbitrage opportunities if and only if $\mathcal{P}$ is nonempty. Furthermore, they have characterized the set of P-martingale measures in a different context: $n = 1$ and the price process (one-dimensional) is supposed to be any continuous semimartingale (actually, the arbitrage-free hypothesis implies that it can be written $M + \int \alpha_s d\langle M\rangle_s$ for some continuous local martingale M and some predictable process α).

1.8. Characterization of the P-martingale measures.

PROPOSITION 1.8.1. *The following properties are equivalent*:

(i) *R is a P-martingale measure.*

(ii) *R is a probability equivalent to P that admits the density*

$$\frac{dR}{dP}\bigg|_{F_T} = \varepsilon\left(-\int_0^{\cdot} \theta_s^*\, dW_s + N\right)_T$$

where N_t is a local martingale that is orthogonal (*in the quadratic variation sense*) *to the prices of the basic securities*

$$\left\langle N, \int_0^{\cdot} \sigma_i(s)d\widetilde{W}_s\right\rangle_T = 0 \quad \forall i \in \{1, 2, \dots, n\}, \quad Q\,\text{a.s.}$$

Proof. Let us show that (i) and (ii) are equivalent. Let R be some probability measure that is equivalent to P on F_T. Put

$$L_t = \frac{dR}{dP}\bigg|_{F_t}, \qquad 0 \le t \le T.$$

L_t is a strictly positive martingale under P. We introduce the local martingale M given by

$$M_t = \int_{0+}^t \frac{1}{L_{s-}}\, dL_s, \qquad 0 \le t \le T,$$

so that

$$L_t = \varepsilon(M)_t, \qquad 0 \le t \le T.$$

Note that, for any $i \le n$, $P_i^d(t)$ is continuous, and hence locally bounded; therefore, by Proposition 1.b in the Appendix, it follows that $\langle M, P_i^d\rangle$ exists. By the Girsanov theorem (Appendix Corollary 1.A (i)), for any $i = 1, \dots, n$, $P_i^d(t)$ is a local martingale under R if and only if

$$(b_i(t) - r_t)dt + d\left\langle M, \int_0^{\cdot} \sigma_i(s)dW_s\right\rangle_t = 0,$$

which can also be written, because $b_i - r = \sigma_i\theta$, as

$$\left\langle M + \int_0^{\cdot} \theta_s^*\, dW_s, \int_0^{\cdot} \sigma_i(s)dW_s\right\rangle_T = 0, \quad P\,\text{a.s.}$$

The result therefore follows. □

In particular, the reference probability Q is an equivalent martingale measure. Also, if we suppose that θ belongs to the range of σ^*, then Q is a minimal P-martingale measure, in the following sense (see [Fö-Sc] or [An-St]).

DEFINITION 1.8.1. *A P-martingale measure R will be called minimal if any local P-martingale that is orthogonal to $\int \sigma_i \, dW$, for $1 \leq i \leq n$, under P, remains a local martingale under R.*

We state the following property.

PROPOSITION 1.8.2. *The following properties are equivalent:*

(i) *Q is a minimal probability.*

(ii) *There is a predictable vector process α_t such that $\theta_t = \sigma_t^* \alpha_t$.*

Remark. Suppose that (i) or (ii) is satisfied. Let R be a P-martingale measure. By Proposition 1.8.1, there exists some local martingale N with $N_0 = 0$, orthogonal to $\int \sigma_i \, dW$, $1 \leq i \leq n$, such that

$$\left.\frac{dR}{dP}\right|_{F_T} = \varepsilon\left(-\int_0^{\cdot} \theta_s^* \, dW_s + N\right)_T .$$

Recall that

$$\left.\frac{dQ}{dP}\right|_{F_T} = \varepsilon\left(-\int_0^{\cdot} \theta_s^* \, dW_s\right)_T .$$

If N is locally square integrable, then R is minimal if and only if $N = 0$, that is, $R = Q$. Indeed, if R is minimal, then N is a local martingale under R (because it is a local P-martingale orthogonal to $\int \sigma_i \, dW$, $1 \leq i \leq n$). If N is supposed to be locally square integrable, then $\langle -\int \theta^* \, dW + N, N \rangle$ exists and we can apply the Girsanov theorem (Appendix Corollary 1.A (i))

$$\left\langle -\int \theta^* \, dW + N, \, N \right\rangle = 0,$$

that is, $\langle N, N \rangle = 0$ (because $\theta_t = \sigma_t^* \alpha_t$). Hence, $N = 0$.

Proof. The definition gives that the fact that Q is minimal is equivalent to the following property:

(**) Any local P-martingale orthogonal to $\int \sigma_i \, dW$, for $1 \leq i \leq n$, is a local martingale under Q.

By the Girsanov theorem, Corollary 1.A (i), property (**) is equivalent to the following one. Any local P-martingale orthogonal to $\int \sigma_i \, dW$, $1 \leq i \leq n$, is orthogonal to $\int \theta^* \, dW$. By some results on stable subspaces of martingales and orthogonality (see [De-Me, pp. 371, 372, VIII-46–VIII-49]), (**) is equivalent to the fact that $\int \theta^* \, dW$ belongs to the space generated by $\int \sigma_i \, dW$, $1 \leq i \leq n$, which is equivalent to the existence of n predictable processes $\alpha_i(t)$, $1 \leq i \leq n$, such that

$$\int \theta^* \, dW = \sum_{i=1}^{n} \int \alpha_i \sigma_i \, dW,$$

that is, $\theta_t = \sigma_t^* \alpha_t$ where $\alpha_t = (\alpha_1(t), \ldots, \alpha_n(t))^*$. □

Hereafter, the above hypothesis ((i) or (ii)) is supposed to be satisfied. [2]

[2] Actually, this is not restrictive because if this hypothesis is not satisfied, just replace θ_t by its orthogonal projection θ_t^{l} on the range of σ_t^* in all the equations (indeed, $\sigma_t \theta_t = \sigma_t \theta_t^{\mathrm{l}}$).

We introduce the following notation.

Notation. We denote by D the set of local (F_t)-martingales $(N_t,\, 0 \leq t \leq T)$, with $N_0 = 0$, satisfying the following three properties.

(i) The jumps of N are strictly greater than -1 so that $\varepsilon(N)_t$, $0 \leq t \leq T$, is a strictly positive local martingale.

(ii) $\varepsilon(N)_t$, $0 \leq t \leq T$, is a martingale under Q.

(iii)

$$\left\langle N, \int_0^{\cdot} \sigma_i(s) d\widetilde{W}_s \right\rangle_T = 0 \quad \forall i \in \{1,2,\dots,n\}, \quad Q\,\text{a.s.}$$

For any local (F_t)-martingale N belonging to D, define Q^N as the probability measure equivalent to Q that admits $\mathcal{E}(N)_T$ as a Radon–Nikodym derivative with respect to Q on F_T. Q^N is then a P-martingale measure.

PROPOSITION 1.8.3. *The mapping $N \longrightarrow Q^N$ is a one-to-one mapping that carries D onto $\mathcal{P}$.*

Proof. We have clearly

$$\varepsilon\left(-\int \theta_u^* \, dW_u + N\right)_T = \varepsilon\left(-\int \theta_u^* \, dW_u\right)_T \varepsilon(N)_T.$$

Indeed, $\langle N, \int \theta_u^* \, dW_u \rangle = 0$ because it is supposed that $\theta_t = \sigma_t^* \alpha_t$. The result now follows easily. □

Hereafter, to simplify notation, we will suppose that $r = 0$. Then the SDEs relative to the prices are written as

$$dP_i(t) = P_i(t)\sigma_i(t)d\widetilde{W}_t.$$

Also, the value $X^{\pi,-c,x}(t)$ of the portfolio with consumption associated with portfolio π, consumption C, and initial investment x is given by

$$X_t^{\pi,-C,x} = x + \int_0^t \pi_t^* \sigma_t \, d\widetilde{W}_t - C_t \quad \forall t \in [0,T], \quad Q\,\text{a.s.}$$

This seems highly restrictive but it is not. All the results we obtain can be generalized to the case $r \neq 0$. In all the properties, just replace B by B^d and the prices, portfolio, and consumption processes by the discounted processes defined above.

We now turn to the study of the maximum price.

2. Dynamical study of the maximum price.

2.1. Predictable decomposition of the maximum price. The supremum of the possible prices for B at time 0 is given by

$$\sup_{R \in \mathcal{P}} E_R(B) = \sup_{N \in D} E_{Q^N}(B).$$

Also, the essential supremum of the possible prices for B at time t is given by

$$\operatorname{ess\,sup}_{R \in \mathcal{P}} E_R(B/F_t) = \operatorname{ess\,sup}_{N \in D} E_{Q^N}(B/F_t).$$

Using dynamical programming methods (see [ElK]), we have the following theorem. (The proof is given in the appendix.)

THEOREM 2.1.1. *There exists an* RCLL *process* $(J_t,\ 0 \leq t \leq T)$ *so that, for each* t

$$J_t = \text{ess}\sup_{N \in D} E_{Q^N}[B/F_t].$$

J_t *is characterized as the smallest right continuous supermartingale under* Q_N, *for every* N *belonging to* D, *which is equal to* B *at time* T. *Also,* N^* *is optimal* (i.e., $J_t = E_{Q^{N^*}}(B/F_t)$, Q *a.s.*, $0 \leq t \leq T$) *if and only if* J_t *is a martingale under* Q_{N^*}.

Before continuing the dynamical study of J_t, recall the hypothesis satisfied by B (introduced in §1.7):

$$B \leq y + \int_0^T H_s^* \sigma_s \, d\widetilde{W}_s, \quad Q\,\text{a.s.},$$

where y is a positive constant and H_t is a portfolio process that satisfies

$$E_{Q^N}\left[\int_0^T \|\sigma_s^* H_s\|^2 \, ds\right] < +\infty$$

for each $N \in D$.

This hypothesis implies the following result.

PROPOSITION 2.1.1. J_t *is of class* D *and satisfies*

$$J_t \leq y + \int_0^t H_s^* \sigma_s \, d\widetilde{W}_s, \quad Q\,\text{a.s.}, \quad 0 \leq t \leq T.$$

Proof. The process given by

$$y + \int_0^t H_s^* \sigma_s \, d\widetilde{W}_s$$

is a square integrable martingale under each P-martingale measure. Thus, for each $N \in D$,

$$E_{Q^N}[B/F_t] \leq y + \int_0^t H_s^* \sigma_s \, d\widetilde{W}_s, \quad Q\,\text{a.s.}, \quad 0 \leq t \leq T,$$

and the desired result clearly follows. □

The fact that J_t is a supermartingale under Q shows that J_t can be written under Q as the difference between a local Q-martingale and a predictable increasing process (and this decomposition is unique). The above properties relative to J_t will allow us to write the Q-martingale as the sum of a portfolio and a martingale j that may be characterized.

THEOREM 2.1.2. *There exist a portfolio process* φ_t, *a right continuous increasing predictable process* A_t *with* $A_0 = 0$, *and a purely discontinuous* Q*-martingale* j_t *such that*

$$\forall t \in [0,T],\ J_t = J_0 + \int_0^t \varphi_s^* \sigma_s \, d\widetilde{W}_s + j_t - A_t, \quad Q\,\text{a.s.} \tag{7}$$

Remark. Notice that the assumption made on B is not necessary. Theorem 2.1.2 remains true under the hypothesis

$$\sup_{N \in D} E_{Q^N}(B) < \infty,$$

but in this case J_t is not generally of class D and the process j is only a local martingale. The arguments of the proof still hold, but it is a bit more complicated technically because $\langle j \rangle$ is not always defined (see §A.3).

Scheme of the proof. J_t is a Q-supermartingale; hence, it admits a unique decomposition as a local martingale M_t minus an increasing predictable process $A_t : J_t = M_t - A_t$. The martingale M_t admits the Kunita–Watanabe decomposition

$$M_t = J_0 + \int_0^t \varphi_s^* \sigma_s \, d\widetilde{W}_s + j_t \quad \forall t \in [0,T], \quad Q \,\text{a.s.}$$

for some predictable process φ and some Q-local martingale j, such that

$$\left\langle j, \int_0^\cdot \sigma_i(s) dW_s \right\rangle_T = 0, \quad Q \,\text{a.s.} \quad \forall i \in \{1, \ldots, n\}.$$

Using the fact that J_t is a supermartingale under each P-martingale measure, we show that the continuous part of j is equal to zero.

Proof. $J_t - E_Q[B/F_t]$ is a positive RCLL optional supermartingale that is equal to zero at time T and lower than the Q-martingale $X^{H,0,y}$. By the Doob–Meyer decomposition theorem (cf. [De-Me, Thm. VII.8, p. 211]), there exists a Q-integrable right continuous increasing predictable process A_t with $A_0 = 0$ such that

$$J_t - E_Q[B/F_t] = E_Q[A_T/F_t] - A_t, \qquad 0 \leq t \leq T.$$

A_T is square integrable under Q since

$$E_Q[(A_T)^2] \leq 4 \, E_Q[(\sup_{0 \leq t \leq T} J_t)^2] \leq 4 \, E_Q[(X_T^{H,0,y})^2]$$

(see [De-Me, inequality VII.15.1, p. 221]).

Put $M_t = E_Q[A_T + B/F_t]$, $0 \leq t \leq T$. $M_t = J_t + A_t$ is square integrablė martingale under Q; hence, $\langle M \rangle$ exists. Thus, M_t admits the Kunita–Watanabe decomposition with respect to the Q-square integrable martingales $\int \sigma_i \, dW_t$, that is, there exist a predictable process φ_t and a square integrable Q-martingale (RCLL) j_t with $j_0 = 0$ such that

$$E_Q \left[\int_0^T \|\sigma_s^* \varphi_s\|^2 \, ds \right] < +\infty,$$

$$\left\langle j, \int_0^\cdot \sigma_i(s) dW_s \right\rangle_T = 0, \quad Q \,\text{a.s.} \quad \forall i \in \{1, \ldots, n\},$$

and

$$M_t = J_0 + \int_0^t \varphi_s^* \sigma_s \, d\widetilde{W}_s + j_t \quad \forall t \in [0,T], \quad Q \,\text{a.s.}$$

(for the Kunita–Watanabe decompositions, see [De-Me, p. 374, VIII-52]).

Now, j can be written as the sum of its continuous martingale part j^c and its purely discontinuous martingale part j^d

$$j = j^c + j^d.$$

It remains to be shown that j^c is equal to zero.

This result can be obtained using the following lemma (which comes from the fact that J_t is a supermartingale under each P-martingale measure).

LEMMA 2.1.1. *For all $N \in D$, $A_t - \langle N, j \rangle_t$ is an increasing process.*

Proof. By Theorem 2.1.1, we have the following property: for each $N \in D$, J_t is a supermartingale under Q^N. By the Girsanov theorem (Corollary 1A (ii) in the Appendix), this property is equivalent to the following:

$$\forall N \in D, \; -A_t + \langle N, M \rangle_t \text{ is a decreasing process,}$$

which may be written, because $N \in D$, as

$$\forall N \in D, \; A_t - \langle N, j \rangle_t \text{ is an increasing process.} \qquad \square$$

Lemma 2.1.1, applied to some N that can be written as a stochastic integral with respect to j^c, will allow us to show that j^c is equal to zero. We take

$$N_t = \int_0^t n_s \, dj_s^c,$$

where n is a bounded predictable process. If $\mathcal{E}(N)_t$ is a martingale, then N belongs to D and Lemma 2.1.1 applied to N yields the fact that $A - \int n_s \, d\langle j^c \rangle_s$ is increasing. It now remains to decompose the measure dA_t with respect to $d\langle j^c \rangle_t$ and to choose n.

$\langle j^c \rangle$ is integrable because j is square integrable. By the Lebesgue decomposition theorem, there exist a positive predictable process h which belongs to $L^1([0,T] \times \Omega, d\langle j^c \rangle_t \, dQ)$ and an integrable predictable increasing process B such that

$$dA_t = h_t d\langle j^c \rangle_t + dB_t$$

and such that, Q almost surely, the measure dB_t is singular with respect to $d\langle j^c \rangle_t$. For each integer p, we can write the following decomposition:

$$dA_t = h_t \, \mathbf{1}_{\{h(t) \le p\}} d\langle j^c \rangle_t + dB_t^p,$$

where, Q almost surely, the measure dB^p is singular with respect to the measure $\mathbf{1}_{\{h(t) \le p\}} d\langle j^c \rangle_t$.

Let N be given by

$$N_t = \int_0^t \mathbf{1}_{\{h_s \le p\}} (1 + h_s) dj_s^c.$$

Clearly, we may choose a sequence of stopping times T_n, $n \ge 0$, such that $T_n \uparrow T$ almost surely as n tends to infinity and for each $n \in \mathbb{N}$, $\mathcal{E}(N)^{T_n}$ $(= \mathcal{E}(N^{T_n}))$ is a martingale. Thus, for each $n \in \mathbb{N}$, N^{T_n} belongs to D. Lemma 2.1.1 applied to N^{T_n} shows that for each $n \in \mathbb{N}$, $A_t - \langle N^{T_n}, j \rangle_t$ is an increasing process. Hence, for each $n \in \mathbb{N}$, $A^{T_n} - \langle N, j \rangle^{T_n}$ is an increasing process and thus $A - \langle N, j \rangle$ is an increasing process.

Because, Q almost surely, the measure dB^p is singular with respect to the measure $\mathbf{1}_{\{h(t) \le p\}} d\langle j^c \rangle_t$, it follows that the process given by

$$\int_0^t \mathbf{1}_{\{h_s \le p\}} h_s d\langle j^c \rangle_s - \int_0^t \mathbf{1}_{\{h_s \le p\}} (1 + h_s) d\langle j^c \rangle_s$$

is an increasing process. Hence, for all $p \in \mathbb{N}$, Q almost surely, $-\langle j^c \rangle_t$ is increasing on $\{t / h(t) \le p\}$ and this yields the equality

$$\langle j^c \rangle_T = 0, \quad Q \text{ a.s.}$$

Hence, $j^c = 0$. $\square$

2.2. The Brownian case. We now turn to the case when the filtration is generated by the Brownian motion. In this case, Theorem 2.1.2 takes a more simple form: j is equal to zero because every local martingale is continuous. Thus, the maximum price process can be written as the difference of a self-financing portfolio and a predictable increasing process that is equal to zero at time zero. This remarkable decomposition should be stressed; it is similar to that of the price for an American option in a complete market. (One is referred to the theory of American option pricing [Kar].)

THEOREM 2.2.1. *There exist a portfolio process φ_t and an increasing (F_t)-predictable right continuous process A_t with $A_0 = 0$ such that*

$$\forall t \in [0,T],\ J_t = J_0 + \int_0^t \varphi_s^* \sigma_s \, d\widetilde{W}_s - A_t, \quad Q\, q.s. \tag{8}$$

Remark. Because the process A_t is predictable, we can consider only portfolios with consumption for which the consumption process is predictable. We have the following corollaries.

COROLLARY 2.2.1. *J_t is the lowest price process admissible for sellers*; *it is the selling price for B.*

Proof. J_t is a price process admissible for sellers because

- φ_t is a portfolio process,
- A_t is an optional (because it is predictable) RCLL increasing process with $A_0 = 0$, and
- J_t is positive and $J_T = B$.

Let us show that it is the lowest, that is, that for every price process admissible for sellers X_t, we have $J_t \leq X_t$. Let X_t be a price process admissible for sellers. Suppose that X is a supermartingale under any P-martingale measure. Then, by the characterization of J_t (Theorem 2.1.1), it follows that $J_t \leq X_t$. It remains to show that X is a supermartingale under every P-martingale measure. Now, X is a positive process such that $X_T = B$ and such that there exist a portfolio process π_t and an RCLL optional increasing process C_t with $C_0 = 0$ satisfying $X_t = X^{\pi, -C, x}(t)$ (the value of the portfolio with consumption associated with (π_t, C_t)), that is,

$$X_t = x + \int_0^t \pi_u^* \sigma_u \, d\widetilde{W}_u - C_t \geq 0, \quad Q \text{ a.s.}$$

Let N be a local martingale of D. Under Q^N, $x + \int_0^t \pi_u^* \sigma_u \, d\widetilde{W}_u$ is a positive local martingale, hence, a supermartingale. Also, C_T is integrable under Q^N. Hence, X is a supermartingale under Q^N. □

2.3. Optional decomposition of the maximum price. When the filtration is not that generated by the Brownian motion, j is not equal to zero, as is shown by example 2 in §3.4. As a result of the constraint $\Delta N > -1$, Lemma 2.1.1 does not imply that j is equal to zero. Thus, the predictable decomposition of J_t under Q is not the good one. The good decomposition will be the optional decomposition (Theorem 2.3.1). Using Lemma 2.1.1 (in other words, the fact that J_t is a supermartingale under each P-martingale measure), we will show that j_t is a process with negative jumps only such that the process f defined by $f_t = A_t - j_t$ is a nondecreasing process.

Let us decompose j with respect to the sign of its jumps:

$$j = j^+ + j^-,$$

where j^+ (respectively, j^-) is the compensated integral of $\mathbf{1}_{\{\Delta j(t) > 0\}}$ (respectively, $\mathbf{1}_{\{\Delta j(t) < 0\}}$) with respect to j, that is,

$$j^+ = \mathbf{1}_{\{\Delta j > 0\}} \cdot j; \qquad j^- = \mathbf{1}_{\{\Delta j < 0\}} \cdot j.$$

Notice that j^+ and j^- are square integrable martingales.

Concerning j^+, the part of j corresponding to the positive jumps of j, Lemma 2.1.1 applied to a well-chosen local martingale N will allow us to show that $j^+ = 0$. (The proof is the same as that for $j^c = 0$ in the proof of Theorem 2.1.2.)

PROPOSITION 2.3.1. *The jumps of the purely discontinuous martingale j are negative.*

Proof. We take $N_t = \int_0^t n_s \, dj_s^+$, where n is a bounded positive predictable process. Notice that N is a square integrable martingale that admits only positive jumps.

It now remains to decompose the measure dA_t with respect to $d\langle j^+ \rangle_t$ and choose the process n_t. Using the same arguments as those used in the proof of Theorem 2.1.2 (replacing j^c by j^+), we obtain the desired result. □

We now come to the most important result, which will allow us to characterize the supremum of the possible prices for B.

THEOREM 2.3.1. *The process $A_t - j_t$ is an increasing process, hence, if we denote it by f_t,*

$$\forall t \in [0,T],\ J_t = J_0 + \int_0^t \varphi_s^* \sigma_s \, d\widetilde{W}_s - f_t, \quad Q\,a.s. \tag{9}$$

In particular,

$$B = J_0 + \int_0^T \varphi_s^* \sigma_s \, d\widetilde{W}_s - f_T, \quad Q\,a.s.$$

Remark. The results of Proposition 2.3.1 and Theorem 2.3.1 still hold under the weaker hypothesis

$$\sup_{N \in D} E_{Q^N}(B) < +\infty$$

(see §A.3).

Proof. Put $f_t = A_t - j_t$. By definition, f is an RCLL process. Hereafter, we will adopt the following notation: for any locally integrable finite variation RCLL adapted process V, we denote by V^P its predictable compensator.

For any $\mathcal{E} \in \,]0,1[$, define

$$u_t^{\mathcal{E}} = \sum_{s \le t} \Delta j_s \mathbf{1}_{\Delta j_s < -\mathcal{E}}$$

and $j^{\mathcal{E}} = u^{\mathcal{E}} - (u^{\mathcal{E}})^P$. Recall the following result (see [De-Me, p. 369, VII-44]): $j^{\mathcal{E}}$ locally converges in M^1 to j as $\mathcal{E} \to 0$, that is, there exists a sequence of stopping times T_k, $k \in \mathbb{N}$, such that $T_k \uparrow T$ almost surely as k tends to infinity and for every $k \in \mathbb{N}$,

$$E[\sup_{s \le T_k} (j_s^{\mathcal{E}})] < +\infty$$

and

$$\lim_{\mathcal{E} \to 0} E[\sup_{s \le T_k} |j_s^{\mathcal{E}} - j_s|] = 0.$$

Suppose we have shown that for all $\mathcal{E} \in \,]0,1[$, $A_t + (u^{\mathcal{E}})_t^P$ is an increasing process. Put $f^{\mathcal{E}} = A - j^{\mathcal{E}}$. We have $f_t^{\mathcal{E}} = (A_t + (u^{\mathcal{E}})_t^P) - u_t^{\mathcal{E}}$. The fact that the jumps of j are negative shows that $u^{\mathcal{E}}$ is a decreasing process. Hence, $f^{\mathcal{E}}$ is an increasing process.

We have $f(t) = f^{\mathcal{E}}(t) + j^{\mathcal{E}}(t) - j(t)$. Using the above property of convergence, it follows that

$$\lim_{\mathcal{E}\to 0} E[\sup_{s\leq T_k} |f_s^{\mathcal{E}} - f_s|] = 0.$$

Hence, f is an increasing process as the limit of increasing processes.

It remains to be shown that for all $\mathcal{E} \in]0, 1[$, $A_t + (u^{\mathcal{E}})_t^P$ is an increasing process.

For any $\mathcal{E} \in]0, 1[$, define

$$N_t^{\mathcal{E}} = -\sum_{s\leq t} \mathbf{1}_{\Delta j_s < -\mathcal{E}} + \left(\sum_{s\leq t} \mathbf{1}_{\Delta j_s < -\mathcal{E}}\right)_t^P .$$

Because the filtration is left quasi-continuous,

$$[N^{\mathcal{E}}, j] = -u^{\mathcal{E}}, \quad \text{hence } \langle N^{\mathcal{E}}, j\rangle = -(u^{\mathcal{E}})^P.$$

Then it is clear that Lemma 2.1.1 applied to $N^{\mathcal{E}}$ shows that the process $A + (u^{\mathcal{E}})^P$ is an increasing process. However, $N^{\mathcal{E}}$ does not belong to D because $N^{\mathcal{E}}$ admits only jumps equal to -1. To solve this, put

$$N = \alpha N^{\mathcal{E}} \quad \text{for } 0 < \alpha < 1,$$

so that $\Delta N > -1$.

Now, we may choose a sequence of stopping times T_n, $n \geq 0$, such that $T_n \uparrow T$ almost surely as n tends to infinity and for each $n \in \mathbb{N}$, $\mathcal{E}(N)^{T_n}$ $(= \mathcal{E}(N^{T_n}))$ is a martingale. Thus, for each $n \in \mathbb{N}$, N^{T_n} belongs to D. Lemma 2.1.1 applied to N^{T_n} shows that for each $n \in \mathbb{N}$, $A^{T_n} - \langle N, j\rangle^{T_n}$ is an increasing process; hence, $A - \langle N, j\rangle$ is an increasing process, that is, $A + \alpha(u^{\mathcal{E}})^P$ is an increasing process for each $\alpha < 1$. Hence, $A + (u^{\mathcal{E}})^P$ is an increasing process. □

To explain the consequences of the preceding theorem, we begin with the following corollary.

COROLLARY 2.3.1. *J_t is the lowest price process admissible for sellers; it is the selling price for B.*

Proof. The proof is similar to that of Corollary 2.2.1. □

The dynamic hedging strategy adopted by the seller happens continuously in time. At $t = 0$, the seller sells the claim at J_0. He invests this amount in the hedging portfolio (determined by φ). At time $t > 0$, the value of the self-financing portfolio determined by φ and J_0 is greater than the price for B at time t; hence, between 0 and t, the seller has withdrawn from the portfolio the nonnegative amount given by

$$J_0 + \int_0^t \varphi_s^* \sigma_s \, d\widetilde{W}_s - J_t.$$

At time t, the amount invested in the portfolio is only equal to J_t. At time $s > t$, the value of the self-financing portfolio determined by φ and J_t,

$$J_t + \int_t^s \varphi_u^* \sigma_u \, d\widetilde{W}_u,$$

is greater than the price for B at time s (J_s); between t and s, the seller has withdrawn the amount

$$J_t + \int_t^s \varphi_u^* \sigma_u \, d\widetilde{W}_u - J_s$$

from his portfolio. At time s, the amount invested in the portfolio is only equal to J_s, and so on.

The readjustments have to be done continuously in time so that the withdrawals correspond to the process f_t, which represents the cumulative amount withdrawn from the portfolio between 0 and t. As time passes, the amount invested in the portfolio is too high, so the seller withdraws some money from the portfolio that should be given to the buyer, which, because it is not, is a profit for the seller.

COROLLARY 2.3.2. (1) *For each* $N \in D$, $E_{Q^N}[B] = J_0 - E_{Q^N}[f_T]$.

(2) *Let* N_n, $n \geq 0$, *be an optimizing sequence of* D, *that is, such that*

$$\lim_{n \to +\infty} E_{Q^{N_n}}[B] = J_0.$$

Then

$$\lim_{n \to \infty} E_{Q^{N_n}}(f_T) = 0.$$

Remark 1. If B is only supposed to satisfy

$$\sup_{N \in D} E_{Q^N}(B) < +\infty,$$

then the second point of this corollary still holds but the first one does not (see §A.3).

Remark 2. If Q^{N_n} converges in a certain sense to a limit probability Q^0 and if f_T is smooth enough, then we have $E_{Q^0}[f_T] = 0$. In this case, J_t is a martingale under Q^0, but this limit probability is not necessarily equivalent to Q, as is shown by example 2 in §3.4. More generally, if there exists a subsequence still denoted by N_n such that $\mathcal{E}(N_n)_T$ converges almost surely and if the limit is denoted by L^*, then $f_T = 0$ on $\{L^* > 0\}$.

Proof of (1). We denote by V_t the process

$$J_0 + \int_0^t \varphi_s^* \sigma_s \, d\widetilde{W}_s.$$

Let N be an element of D. V_t is a positive continuous Q^N-local martingale, hence, a Q^N-supermartingale. f_T is Q^N-integrable because V_T and B are Q^N-integrable. We have, by the assumption made on B,

$$\sup_{t \in [0,T]} V_t \leq \sup_t (J_t + f_t) \leq \sup_t (X^{H,0,y}(t)) + f_T.$$

Now, $X^{H,0,y}(t)$ is a square integrable Q^N-martingale; hence, $\sup_t (X^{H,0,y}(t))$ is square integrable under Q^N. Hence, V_t is a uniformly integrable local martingale under Q^N, and thus it is a Q^N-martingale. It follows that for each $N \in D$,

$$E_{Q^N}[B] = J_0 - E_{Q^N}[f_T]. \qquad \square$$

Proof of (2). Equality (1) applied to the local martingales N_n gives the equalities

$$\forall n \in N, \; E_{Q^{N_n}}(B) = J_0 - E_{Q^{N_n}}(f_T).$$

Hence, if we let n tend to $+\infty$, we obtain the desired result of

$$\lim_{n \to +\infty} E_{Q^{N_n}}[f_T] = 0. \qquad \square$$

In the following theorem, we state that the class of contingent claims for which the supremum price is obtained for an optimal N is exactly the set of attainable claims. This result generalized those of Pagès [Pag] and Karatzas et al. ([KLSX, Thm. 8.5]) in the context of a Brownian model.

THEOREM 2.3.2. *The following properties are equivalent.*

(i) $\sup_{N \in D} E_{Q^N}[B]$ *is attained by* $\hat{N} \in D$.

(ii) B *is attainable, that is, there exist a constant* x *and a portfolio* π *such that*

$$B = x + \int_0^T \pi_u^* \sigma_u \, d\widetilde{W}_u, \qquad Q \text{ a.s.}$$

and

$$E_Q \left[\int_0^T \|\sigma_s^* \pi_s\|^2 \, ds \right] < +\infty.$$

(iii) *For each local martingale* $N \in D$, $E_{Q^N}(B) = E_Q(B)$.

Remark. If B is only supposed to satisfy

$$\sup_{N \in D} E_{Q^N}(B) < +\infty$$

then only a part of this theorem still holds (see §A.3).

Proof. Let us show that (ii) ⇒ (iii). Suppose that there exist a constant x and a portfolio π such that

$$B = x + \int_0^T \pi_u^* \sigma_u \, d\widetilde{W}_u \quad Q \text{ a.s.}$$

and

$$E_Q \left[\int_0^T \|\sigma_s^* \pi_s\|^2 \, ds \right] < +\infty.$$

Let S_t be the process

$$x + \int_0^t \pi_u^* \sigma_u \, d\widetilde{W}_u, \qquad 0 \leq t \leq T.$$

S_t is a square integrable Q-martingale; hence, $S_t = E_Q[B/F_t]$. Therefore, S_t is positive and lower than $X^{H,0,y}(t)$.

Let N be an element of D. Under Q^N, S_t is a local martingale of class (D), and hence, a martingale. Also, S_t is continuous; hence, the processes S and J are indistinguishable. Notice that we have also shown that each element N of D is optimal, which is equivalent to (iii).

It remains to show that (i) ⇒ (ii). Suppose that N^0 is an optimal local martingale of D. Item (1) of Corollary 2.3.2 applied to N^0 gives

$$E_{Q^0}(B) = J_0 - E_{Q^0}(f_T),$$

that is,

$$E_{Q^0}(f_T) = 0.$$

Hence,

$$f_T = 0, \quad Q^0 \text{ a.s.},$$

that is,

$$f_T = 0 \quad Q \text{ a.s.}$$

since Q and Q^0 are equivalent. □

After the study of the maximum price, we now turn to the study of the minimum price for B.

2.4. Study of the minimum price. Recall that B is lower than M_T where $M_t = X^{H,0,y}(t)$. Let us determine the buyer's price for B. Let us consider a buyer who wants to buy some contingent claims with payoff B and maturity T, between time 0 and time T. Suppose the buyer chooses the price X_t for B at any time t; more precisely, the buyer must choose his price process $(X_t, t \geq 0)$ (that is, an RCLL optional process lower than M_t with $X_T = B$) and a portfolio process π_t. Also, the buyer does not want to run any risk of losing money. Therefore, he will only choose strategies that allow him to hedge completely by the controlled portfolio of the basic securities in the sense that, for any t, any s such that $0 \leq t \leq s \leq T$.

Suppose that, at time t, the buyer buys the claim at price X_t and sells the self-financing portfolio (determined by π) at price X_t. At time s, he sells the claim at price X_s and buys the self-financing portfolio; he then makes the profit (nonnegative) given by

$$X_s - \left(X_t + \int_t^s \pi_u^* \sigma_u \, d\widetilde{W}_u \right).$$

More precisely, the process given by

$$X_t - \left(X_0 + \int_0^t \pi_u^* \sigma_u \, d\widetilde{W}_u \right), \qquad 0 \leq t \leq T$$

is an increasing process.

DEFINITION 2.4.1. *A process X_t is called a* price process admissible for buyers *if X_t is an* RCLL *optional process lower than M_t with $X_T = B$ such that there exist a portfolio π_t and an* RCLL *optional increasing process D_t with $D_0 = 0$ satisfying $dX_t = \pi_t^* \sigma_t (dW_t + \theta_t \, dt) + dD_t$.*

Remark. It is equivalent to say that a process X_t is a price process admissible for buyers if there exists a hedging portfolio of B that is a portfolio with savings whose value is equal to the price, that is, if there exist a portfolio process π_t and an RCLL optional increasing process D_t with $D_0 = 0$ satisfying

$$X^{\pi,D,x}(T) = B \quad \text{and} \quad X_t = X^{\pi,D,x}(t), \qquad 0 \leq t \leq T.$$

The purchase price is then given by Definition 2.4.2.

DEFINITION 2.4.2. *The greatest process admissible for buyers is called the* purchase price.

We have clearly the following property.

PROPOSITION 2.4.1. (i) *A process X_t is a price process for B admissible for buyers if and only if $M_t - X_t$ is a process price for $M_T - B$ admissible for sellers.*

(ii) *The purchase price for B is equal to the difference between M_t and the selling price for $M_T - B$.*

We now turn to the study of the essential infimum of the possible prices for B. Let K_t, $0 \leq t \leq T$, be the right continuous process satisfying

$$K_t = \operatorname*{ess\,inf}_{N \in D} E_{Q^N}[B/F_t], \qquad 0 \leq t \leq T.$$

Notice that the minimum price for B is given by the maximum price for $M_T - B$ by the equality

$$M_t - K_t^B = J_t^{M(T)-B}, \qquad 0 \leq t \leq T, \quad Q \text{ a.s.}$$

It follows that the properties of K can be derived from those of J.

THEOREM 2.4.1. *K_t, $0 \leq t \leq T$, is characterized as the greatest right continuous submartingale under Q_N, for any $N \in D$, with $N_T = B$. Also, N is optimal if and only if K_t is a martingale under Q_N.*

THEOREM 2.4.2. *There exist a portfolio process ψ_t and a right continuous increasing optional process g_t with $g_0 = 0$ such that*

$$K_t = K_0 + \int_0^t \psi_s^* \sigma_s \, d\widetilde{W}_s + g_t, \quad Q \text{ a.s.}, \quad 0 \leq t \leq T \tag{10}$$

Also, there exist a right continuous increasing predictable process B_t with $B_0 = 0$ and a purely discontinuous martingale i_t with negative jumps such that

$$g_t = -i_t + B_t, \qquad 0 \leq t \leq T.$$

COROLLARY 2.4.1. *K_t is the greatest of the price processes admissible for buyers, that is, K_t is the purchase price for B.*

Remark 1. The properties of K_t can be clearly derived directly without using the properties of J_t (by the use of stochastic control methods).

Remark 2. The assumption made on B is not necessary to obtain the above results. One should derive the properties of K directly; the price processes should be supposed to be lower than $E_Q(B/F_t)$ (instead of M_t).

Now, let us compare the hedging strategies associated with the maximum and minimum prices to the "optimal" strategy in Föllmer and Schweizer's sense. Recall that their optimal strategy is obtained by projecting the Q-martingale $E_Q[B/F_t]$ orthogonally on the stable subspace generated by $\int \sigma_i \, dW$, $1 \leq i \leq n$, that is,

$$B = E_Q[B] + \int_0^T \pi_u^* \sigma_u \, d\widetilde{W}_u + N_T,$$

where N is a local martingale under Q orthogonal to $\int \sigma_i \, d\widetilde{W}$, $1 \leq i \leq n$ and hence a local martingale under P, because Q is minimal. $E_Q[B]$ is called the "optimal" price for B.

Let us compare the three price-portfolio strategies:

- The price-portfolio strategy associated with the "optimal" price for B, whose value of the self-financing portfolio is given by

$$E_Q(B) + \int_0^t \pi_u^* \sigma_u \, d\widetilde{W}_u, \qquad 0 \leq t \leq T$$

and whose price process is given by $E_Q(B/F_t)$, $0 \leq t \leq T$, is characterized by the fact that the difference between the value of the self-financing portfolio and the price is a local martingale under P equal to zero at time zero that is orthogonal to $\int \sigma_i \, dW$, $1 \leq i \leq n$.

- The price-portfolio strategy associated with the maximum price for B, whose value of the self-financing portfolio is given by

$$J_0 + \int_0^t \varphi_u^* \sigma_u \, d\widetilde{W}_u, \qquad 0 \leq t \leq T$$

and whose price process is given by J_t, $0 \leq t \leq T$, is characterized by the fact that the difference between the value of the self-financing portfolio and the price is an increasing process equal to zero at time zero and by the fact that the price is minimal (in the sense defined above).

- Also, the price-portfolio strategy associated with the minimum price for B is characterized by the fact that the difference between the price and the value of the self-financing portfolio is an increasing process equal to zero at time zero and by the fact that the price is maximal (in the sense defined above).

In the next section, we give a few methods for computing the maximum price and a few examples that illustrate the obtained results.

3. Methods for computing.

3.1. Determination of J_t as the limit of a sequence of processes (in the Brownian case). The selling price for B is given by J_t, the essential supremum of $E_R(B/F_t)$, for $R \in \mathcal{P}$ (the set of martingale measures). However, J_t is generally difficult to compute. We may then restrict the control set to $\mathcal{P}_n$, $n \in N$, so that the essential supremum J^n taken over all the elements of $\mathcal{P}_n$ is attained. $\mathcal{P}_n$ is chosen so that J_t is the limit of $J^n(t)$ as n tends to infinity, and we then obtain a sequence approximating J that can be calculated explicitly. We develop this method in the context of a Brownian model. In this case, every local martingale is a stochastic integral with respect to a reference martingale (the Brownian); we will see later that this property allows us to obtain some precise results.

In this section, we will suppose that $\sup_{R \in \mathcal{P}} E_R(B) < \infty$ and $E_Q(B^2) < \infty$. Let us denote by $L^2[0,T]$ the class of predictable processes ϕ satisfying

$$\int_0^T \|\Phi_t\|^2 \, dt < \infty, \quad \text{a.s.}$$

Let $K(\sigma)$ be the subset of $L^2[0,T]$ defined by

$$\nu \in K(\sigma) \Leftrightarrow \nu \in L^2[0,T]/\sigma(t)\nu(t) = 0 \quad \forall t \in [0,T], \quad \text{a.s.}$$

and

$$\mathcal{E}\left(\int_0^{\cdot} \nu_s^* \, d\widetilde{W}_s\right)_t, \qquad 0 \leq t \leq T$$

is a Q martingale.

The following result is clear.

PROPOSITION 3.1.1. *The following properties are equivalent.*

(i) *N belongs to D.*

(ii) *There exists $\nu \in K(\sigma)$ such that*

$$N_t = \int_0^t \nu_s^* \, d\widetilde{W}_s \quad \forall t \in [0,T], \quad \text{a.s.}$$

It follows that $J_t = \text{ess sup}_{\nu \in K(\sigma)} E_{Q^\nu}[B/F_t]$, where Q^ν denotes the probability measure that admits the following Radon–Nikodym derivative with respect to Q:

$$\exp\left\{\int_0^T \nu_s^* \, d\widetilde{W}_s - \frac{1}{2}\int_0^T \|\nu_s\|^2 \, ds\right\}.$$

In this context, ν can be interpreted as a risk premium vector associated with the risks of the market that cannot be controlled using the prices of the basic securities (loosely speaking, those risks are in an another direction). Thus, this risk premium has the role of control in the determination of the maximum price of the contingent claim.

For any $n \in N$, define $K^n(\sigma)$, a subset of $K(\sigma)$ by

$$K^n(\sigma) = \{\nu \in K(\sigma)/\|\nu(t)\| \leq n \, \forall t \in [0,T], \text{ a.s.}\}.$$

For any $n \in N$, let $J^n(t)$ be the right continuous process satisfying

$$J^n(t) = \text{ess} \sup_{\nu \in K^n(\sigma)} E_{Q^\nu}[B/F_t].$$

It is characterized by the following property (similar to the characterization of J).

PROPOSITION 3.1.2. *$J^n(t)$ is characterized as the smallest right continuous supermartingale under Q^ν, for every $\nu \in K^n(\sigma)$, with $J^n(T) = B$. Also, ν is optimal (i.e., $\nu \in K^n(\sigma)$ with $J^n(t) = E_{Q^\nu}[B/F_t]$, $0 \leq t \leq T$, a.s.) if and only if $J^n(t)$ is a martingale under Q^ν.*

The properties relative to J and J^n, $n \in N$, allow us to state that J is the limit of J^n as n tends to infinity.

THEOREM 3.1.1.

$$J_t = \lim_{n \to +\infty} \uparrow J^n(t) \text{ a.s.} \quad \forall t \in [0,T].$$

Proof. Let J° be the process defined by $J_t^\circ = \lim_{n \to +\infty} \uparrow J^n(t)$. Let us show that $J_t^\circ = J_t$. We clearly have $J_t^\circ \leq J_t$. It remains to show that $J_t^\circ \geq J_t$.

J_t° is an RCLL supermartingale under every Q^ν, $\nu \in K(\sigma)$ and bounded, because it is the increasing limit of RCLL supermartingales. Using this property, one can show quite easily, by a proof similar to that of Theorem 2.1.2, that J_t° is a price process for B admissible for sellers in the sense that there exist a portfolio process π_t° and a predictable increasing process A_t° such that

$$J_t^\circ = J_0^\circ + \int_0^t (\pi_s^\circ)^* \sigma_s \, d\widetilde{W}_s - A_t^\circ, \qquad 0 \leq t \leq T,$$

hence, $J_t \leq J_t^\circ$ (because J_t is the lowest of the admissible prices). □

Whereas J_t is a priori difficult to calculate, we have a characterization of J^n that is linked to the fact that there exists an optimal control for J^n (contrary to J).

THEOREM 3.1.2. *There is an optimal control associated with J^n.*

Proof. The proof is an application of Theorem 3.30 in [ElK]. (Indeed, the model is strongly dominated and the space to which ν_t belongs is compact.) □

$J^n(t)$ can be determined explicitly as the unique solution of a backward stochastic differential equation of the type studied by E. Pardoux and S. G. Peng (see [Pa-Pe]). Thus, we have, in the general case, a construction of the value function similar to the construction of the solution of the Hamilton–Jacobi–Bellman equation in the Markovian case. We denote by $\Pi_{\text{Ker}\sigma(s)}$ the orthogonal projection that maps $\mathbb{R}^d$ onto the kernel of σ_s.

THEOREM 3.1.3. *Let $(X^n(t), Y^n(t))$ be the unique solution of the backward stochastic differential equation*

$$X_t^n - \int_t^T n \,\|\pi_{\operatorname{Ker}\sigma_s}(Y_s^n)\|\, ds + \int_t^T Y_s^{n^*}\, d\widetilde{W}_s = B, \qquad 0 \le t \le T. \tag{11}$$

Then

(1) $X^n(t) = J^n(t)$, $0 \le t \le T$, *almost surely.*

(2) *If ν^n is an optimal control associated with $J^n(t)$, then*

$$\nu_s^n = n \frac{\Pi_{\operatorname{Ker}\sigma_s}(Y_s^n)}{\|\Pi_{\operatorname{Ker}\sigma_s}(Y_s^n)\|}\, \mathbf{1}_{\{Y_s^n \neq 0\}}, \; ds\, dQ\, a.s.$$

Remark. Thus, we have $J_t = \lim_{n\to+\infty} E_{Q^{\nu^n}}(B/F_t)$ and $ds\, dQ$ almost surely. If a subsequence of $\{\|\nu^n(s)\|, n \in N\}$ converges, its limit is equal to 0 or $+\infty$. Loosely speaking, we see that the maximum price for the contingent claim is obtained by letting the norm of the risk premium tend to $+\infty$ or 0.

Proof. Let ν^n be an optimal control associated with J^n. Under Q, J^n is a supermartingale and admits the following decomposition

$$J_t^n = J_0^n + \int_0^t \varphi_s^{n^*}\, d\widetilde{W}_s - A_t^n, \qquad 0 \le t \le T,$$

where $\varphi^n \in L^2[0,T]$ and A^n is a predictable increasing process with $A^n(0) = 0$. Now, J^n is a martingale under Q^{ν_n}; hence, by the Girsanov theorem,

$$A_t^n = \int_0^t \varphi_s^{n^*} \nu_s^n\, ds, \qquad 0 \le t \le T.$$

Let $\nu \in K^n(\sigma)$. J^n is a supermartingale under Q^ν; hence, by the Girsanov theorem

$$\varphi^n(s)^* \nu^n(s) \ge \varphi^n(s)^* \nu(s), \quad ds\, dQ \text{ a.s.}$$

Because this inequality holds for each $\nu \in K^n(\sigma)$, we have

$$\varphi^n(s)^* \nu^n(s) = \operatorname*{ess\ sup}_{\nu \in K^n(\sigma)} [\varphi^n(s)^* \nu] = n \left\|\pi_{\operatorname{Ker}\sigma(s)}(\varphi^n(s))\right\|$$

and

$$\nu_s^n = n \frac{\Pi_{\operatorname{Ker}\sigma_s}(\varphi_s^n)}{\|\Pi_{\operatorname{Ker}\sigma_s}(\varphi_s^n)\|}\, \mathbf{1}_{\{\varphi_s^n \neq 0\}}, \; ds\, dQ \text{ a.s.}$$

Hence,

$$J_t^n - \int_t^T n \,\|\pi_{\operatorname{Ker}\sigma_s}(\varphi_s^n)\|\, ds + \int_t^T \varphi_s^{n^*}\, d\widetilde{W}_s = B, \qquad 0 \le t \le T. \qquad \square$$

Remark. For the results of existence and uniqueness of solutions of backward equations, see [Pa-Pe].

$M^2(0,T;\mathbb{R}^d)$ will denote the normed vectorial space of $\mathbb{R}^d$-valued processes that are predictable and belong to $L^2([0,T] \times \Omega, dt\, dQ)$.

For any $\phi \in M^2(0,T;\mathbb{R}^d)$, define its norm by

$$\|\phi\|^2_{M^2} = E_Q\left[\int_0^T \|\phi_s\|^2\,ds\right].$$

The backward equation (R) given by

$$X_t - \int_t^T n\,\|\pi_{\mathrm{Ker}\,\sigma_s}(Y_s)\|\,ds + \int_t^T Y_s^*\,d\widetilde{W}_s = B, \qquad 0 \le t \le T,$$

has Lipschitz coefficients (with respect to Y), and hence admits a unique solution $(X,Y) \in M^2(0,T;\mathbb{R}) \times M^2(0,T;\mathbb{R}^d)$. Recall that the solution can be constructed using a Picard type iteration. Y_0 is taken to be equal to 0. Let (X_p, Y_p), $p \in N^*$ be a sequence in $M^2(0,T;\mathbb{R}) \times M^2(0,T;\mathbb{R}^d)$ defined recursively by $X_p(0)$ and Y_p, which are constructed from Y_{p-1} by the representation theorem

$$E_Q\left[B - \int_0^T n\,\|\pi_{\mathrm{Ker}\,\sigma_s}(Y_{p-1}(s))\|\,ds/F_t\right] = X_p(0) + \int_0^t Y_p(s)^*\,d\widetilde{W}_s, \text{ a.s.}$$

$X_p(t)$ is then defined by

$$X_p(t) = \int_t^T n\,\|\pi_{\mathrm{Ker}\,\sigma_s}(Y_{p-1}(s))\|\,ds - \int_t^T Y_p(s)^*\,d\widetilde{W}_s + B, \qquad 0 \le t \le T.$$

Using Pardoux and Peng's result, X_p (respectively, Y_p) converges in $M^2(0,T;\mathbb{R})$ (respectively, $M^2(0,T;\mathbb{R}^d)$) to X, Y, the solution of (R) as p tends to $+\infty$.

Recall that the increasing process associated with the Q-supermartingale J is denoted by A. We denote by A^n the increasing process associated with J^n, that is,

$$A_t^n = \int_0^t n\,\|\pi_{\mathrm{Ker}\,\sigma_s}(\varphi_s^n)\|\,ds.$$

In general, the process A^n does not converge to A almost surely, but we have the following property (cf. [De-Me, Thm. VII.18, p. 223]). If J is of class D, then for each t, the sequence of random variables $A^n(t)$ converges to $A(t)$ weakly in L^1, that is, for each bounded F_T measurable variable U, $\lim_{n\to+\infty} E[A^n(t)U] = E[A(t)U]$.

In the next section, we study the Markovian case. We will see that, contrary to J, J_n is the solution of a usual Bellman equation (and this is linked to the existence of an optimal control associated with J_n, contrary to J).

3.2. The Bellman equation and the maximum price. In a particular case, we propose a numerical method to solve the problem. The general model is complete, that is, the filtration is that generated by the d-dimensional Brownian W. The market contains d securities whose volatility matrix has full rank, but only certain securities (the first n ones) can be traded. Therefore, the market is incomplete. We suppose that all the coefficients of the model are only functions of the time t and the price vector $P(t) = (P_1(t), \ldots, P_d(t))$, functions that are taken to be smooth enough so that the Bellman equations are satisfied. We denote by σ' (respectively, σ, δ) the volatility matrix of the d securities (respectively, the n first securities, the $(d-n)$ others). We have $\sigma' = \binom{\sigma}{\delta}$.

Under Q (the reference probability), the prices of the securities satisfy the following equations

$$(12) \qquad \begin{cases} dP_j(t) = P_j(t)[r(t,P_t)dt + \sigma_j(t,P_t)d\widetilde{W}_t], & 1 \le j \le n, \\ dP_k(t) = P_k(t)[\mu_k(t,P_t)dt + \delta_k(t,P_t)d\widetilde{W}_t], & n+1 \le k \le d. \end{cases}$$

The contingent claim B is taken to be equal to $g(P_1(T),\dots,P_d(T))$ for a $\mathbb{R}^+$-valued function g on $\mathbb{R}^d$ satisfying smoothness conditions (see [Kry, p. 205]). Recall that the selling price is given by

$$J_t = \operatorname{ess}\sup_{\nu\in K(\sigma)} E_{Q^\nu}[B/F_t],$$

where Q^ν denotes the probability measure that admits the following Radon–Nikodym derivative with respect to Q:

$$\exp\left\{\int_0^T \nu_s^* \, d\widetilde{W}_s - \frac{1}{2}\int_0^T \|\nu_s\|^2 \, ds\right\},$$

and $K(\sigma)$ denotes the set of Ker $\sigma(t,P(t))$-valued predictable processes ν_t that belong to $L^2[0,T]$.

Note that under Q^ν, $\widetilde{W}_t - \int_0^t \nu_s \, ds$ is a Q^ν-Brownian motion.

Using some of Krylov and El Karoui's results [Kry], [ElK], we see that the maximum price J_t (which is a function J of t and $P(t)$) is the value function for a more general problem. $(\Omega,\ F_t, P, W)$ is a fixed probability space on which W is an F-Brownian. The controlled system is described by the following equations:

$$(13)\quad \begin{cases} dP_j(t) = P_j(t)[r(t,P_t)dt + \sigma_j(t,P_t)dW_t], & 1\le j\le n,\\ dP_k(t) = P_k(t)[[\mu_k(t,P_t) + \delta_k(t,P_t)\nu_t]dt + \delta_k(t,P_t)dW_t], & n+1\le k\le d,\end{cases}$$

where the control ν belongs to $K(\sigma)$.

The value function is then given by

$$J(t,x) = \sup_{\nu\in K(\sigma)} E_{\nu,t}[g(P_T)/P_t = x]$$

for $x\in (R^+)^d$ and $t\in[0,T]$.

Because the coefficient $\mu_k + \delta_k\,\nu_t$ is not bounded, $J(t,x)$ is not the solution of the classical Bellman equation. However, if we impose some smoothness conditions on the coefficients, J satisfies the following inequality in terms of generalized derivatives:

$$(\alpha)\qquad L\varphi(t,x) + G\varphi(t,x)\nu \le 0 \quad \forall\,\nu\in \operatorname{Ker}\sigma(t,x),$$

where

$$\begin{aligned} L\varphi(t,x) &= \frac{\partial\varphi}{\partial t}(t,x) + \frac{1}{2}\sum_{1\le i,j\le d} x_i x_j(\sigma'\sigma'^*)_{ij}\frac{\partial^2\varphi}{\partial x_i\partial x_j}(t,x)\\ &\quad + \sum_{k=n+1}^{d}\frac{\partial\varphi}{\partial x_k}(t,x)\,x_k\mu_k(t,x) + \sum_{j=1}^{n}\frac{\partial\varphi}{\partial x_j}(t,x)\,x_j r(t,x).\\ G\varphi(t,x) &= \sum_{k=n+1}^{d}\frac{\partial\varphi}{\partial x_k}(t,x)x_k\delta_k(t,x). \end{aligned}$$

(J is characterized as the smallest solution of (α).) Also, equation (α) is clearly equivalent to the following system:

$$(\mathrm{A})\qquad \begin{cases} G\varphi(t,x)\nu = 0 \quad \forall\,\nu\in\operatorname{Ker}\sigma,\\ L\varphi(t,x)\le 0.\end{cases}$$

Thus, J is the smallest solution of system (A). It corresponds to the characterization of J as the smallest selling price (Thm. 2.2.1).

Notice that if J is $C^{1,2}([0,T] \times R^d)$, this system can be derived directly by applying Ito's formula to $J(T, P_1(T), \ldots, P_n(T))$ and using the fact that, by Theorem 2.2.1, J can be written as the difference of a portfolio (constructed from the first n securities) and an increasing process.

Notice that the above system can also be derived using some of Krylov's results on optimal control problems of Markov diffusion processes with unbounded coefficients. Using Krylov's results (cf. [Kry, p. 266]), one can show that, if the coefficients are taken to be smooth enough, $J(t,x)$ is solution of the normalized Bellman equation (in terms of generalized derivatives)

$$\sup_{\{\nu \in \operatorname{Ker} \sigma(t,x)\}} \left[\frac{1}{\lambda(\nu)} (L\varphi(t,x) + G\varphi(t,x)\nu) \right] = 0,$$

where $\lambda(\nu) = \sup(1, \|\nu\|)$. This equation is clearly equivalent to system (A).

To calculate J, we can use the following property (as in §3.1).

PROPOSITION 3.2.1.

$$J(t,x) = \lim_{n \to \infty} J_n(t,x),$$

where

$$J_n(t,x) = \sup_{\nu \in K^n(\sigma)} E_{t,\nu}[g(P_t)/P_t = x],$$

and $K^n(\sigma)$ *is the set of the processes* $\nu_t \in K(\sigma)$ *that are bounded by* n.

For each n, there exists an optimal control associated with J_n. Also, J_n is solution of the Bellman equation

$$\sup_{\{\nu \in \operatorname{Ker} \sigma(t,x), \|\nu\| \le n\}} [L\varphi(t,x) + G\varphi(t,x)\nu] = 0.$$

Notice that

$$\sup_{\{\nu \in \operatorname{Ker} \sigma(t,x), \|\nu\| \le 1\}} [G\varphi(t,x)\nu] = \|\pi_{t,x}(G\varphi(t,x))\|,$$

where $\Pi_{t,x}$ is the orthogonal projection from R^d onto Ker $\sigma(t,x)$ ($G\varphi(t,x)$ is a row vector). Hence, we have the following property.

PROPOSITION 3.2.2. $J_n(t,x)$ *is solution of the following equation*:

$$L\varphi(t,x) + n \|\pi_{t,x}(G\varphi(t,x))\| = 0.$$

Remark. Notice that if

$$H(\varphi)(t,x) = \sum_{j=1}^{n} \frac{\partial \varphi}{\partial x_j}(t,x)\, x_j \sigma_j(t,x) + G(\varphi)(t,x),$$

then $\Pi_{t,x}(H\varphi(t,x)) = \Pi_{t,x}(G\varphi(t,x))$. It follows that if φ is solution of the Bellman equation with $\varphi(T,x) = g(x)$, then $(\varphi(P_t), H\varphi(t,P_t))$ is solution of the backward equation given by

$$\varphi(t,P_t) - \int_t^T n \|\pi_{\operatorname{Ker}\sigma(s,P_s)}(H\varphi(s,P_s))\| \, ds + \int_t^T H\varphi(s,P_s) d\widetilde{W}_s = g(P_T).$$

We recognize the backward equation obtained in §3.1.

In the next section, we give an example that shows that there exist some discontinuous solutions of the normalized Bellman equation.

3.3. Example 1. Let W' be a (unidimensional) Brownian independent of the Brownian W_t (d-dimensional). The filtration is that generated by the Brownians W and W'. The prices of the different securities and the coefficients relative to those prices (appreciation rates and volatilities) depend only on W (that is, are adapted to the filtration of W). B is taken to be a function of the terminal value of W', that is, $B = f(W'_T)$ for a positive bounded real-valued function f on $\mathbb{R}$. In this case, we show that the maximum price for B is constant on $[0,T[$, equal to the supremum of the function f and jumps at time T to reach the value $f(W'_T)$.

Let us consider a more general case. The contingent claim B is taken to be positive bounded and to depend only on W'. Let us determine its maximum price J_t. We define the model more precisely.

Let Ω_1 be the space of all $\mathbb{R}^d$-valued continuous functions on $\mathbb{R}^+$,

$$\Omega_1 = \mathcal{C}(\mathbb{R}^+, \mathbb{R}^d).$$

We denote by F^1 the σ-field generated by the coordinate process $W_t : \omega_1 \to \omega_1(t)$ for $t \geq 0$. Let $(F_t^1, t \geq 0)$ be the filtration generated by the process W_t. Let P^1 be the Wiener measure on Ω_1 constructed so that the coordinate mapping process W_t is Brownian motion.

Let Ω_2 be the space of all real-valued continuous functions on $\mathbb{R}^+$,

$$\Omega_2 = \mathcal{C}(\mathbb{R}^+, \mathbb{R}).$$

We denote by F^2 the σ-field generated by the coordinate process $W'_t : \omega_2 \to \omega_2(t)$ for $t \geq 0$. Let (F_t^2) be the filtration generated by the process (W'_t). Let P^2 be the Wiener measure on Ω_2 constructed so that the coordinate mapping process W'_t is Brownian motion.

Let (Ω, F, P) be the cross-product probability space $(\Omega_1 \times \Omega_2, F_1 \otimes F_2, P_1 \otimes P_2)$. The filtration F_t is defined by $F_t = F_t^1 \otimes F_t^2$. We denote by ω the elements of Ω

$$\omega = (\omega_1, \omega_2), \quad \omega_1 \in \Omega_1, \quad \omega_2 \in \Omega_2.$$

We denote by W_t the first coordinate mapping process and by W'_t the second coordinate mapping process. W_t and W'_t are independent (F_t)-Brownian motions under P. The prices of the basic securities $P^i(t)$, the vector of stock appreciation $b(t)$, and the volatility matrix $\sigma(t)$ are taken to depend only on the first coordinate of the path (ω_1). B is taken to be F_T^2-measurable, positive bounded.

Clearly, if N is a stochastic integral with respect to W' and if the associated exponential martingale is a martingale, then N belongs to D. Hence,

$$J_t \geq \operatorname*{ess\,sup}_{\nu \in D'} E_{P^2}\left[B\mathcal{E}\left(\int \nu_s \, dW'_s\right)_T \Big/ F_t^2\right] \Big/ \mathcal{E}\left(\int \nu_s \, dW'_s\right)_t,$$

where D' is the set of all (F^2)-predictable processes ν defined on Ω^2 such that

$$E_{P^2}\left[\mathcal{E}\left(\int \nu_s \, dW'_s\right)_T\right] = 1.$$

Now, it follows by the representation theorem that

$$\sup_{\nu \in D'} E_{P^2}\left[B\mathcal{E}\left(\int \nu_s \, dW'_s\right)_T\right] = \sup_{\substack{X \in L^1 \\ \|X\|_1 \leq 1}} E_{P^2}[BX],$$

where $L^1 = L^1(\Omega^2, P^2, F_T^2)$ and $\|X\|_1 = E_{P^2}[|X|]$ for $X \in L^1$.

Now, because the L^∞ norm of any function on any measure space is equal to its norm as a linear functional on L^1, we have

$$\sup_{\substack{X\in L^1\\ \|X\|_1\leq 1}} E_{P^2}[BX] = \|B\|_{L^\infty(P^2)},$$

where $\|B\|_{L^\infty(P^2)}$ denotes the essential supremum of B under P^2. Therefore,

$$\sup_{\nu\in D'} E_{P^2}\left[B\mathcal{E}\left(\int \nu_s\, dW'_s\right)_T\right] = \|B\|_{L^\infty(P^2)}.$$

Hence, $J_0 = \|B\|_{L^\infty(P^2)}$.

Also, by the same argument we have

$$\operatorname{ess\,sup}_{\nu\in D'} E_{P^2}\left[B\mathcal{E}\left(\int \nu_s\, dW_s\right)_T \Big/ F_t^2\right] \Big/ \mathcal{E}\left(\int \nu_s\, dW'_s\right)_t = \|B\|_{L^\infty(P^2/F_t^2)},$$

where $\|B\|_{L^\infty(P^2/F_t^2)}$ denotes the essential supremum of B under the conditional probability measure of P^2 given F_t^2. It follows that $J_t = \|B\|_{L^\infty(P^2/F_t^2)}$ for each $t\in[0,T]$.

Remark. This example shows that there exist some discontinuous solutions of the normalized Bellman equation.

In the next section, we give an another example but not in a Brownian model. It illustrates the fact that the purely discontinuous Q-martingale j obtained in the decomposition of J_t may not be equal to zero. Consequently, the optional decomposition of J_t is the good one.

3.4. Example 2. Let N_t be a Poisson process with intensity 1 independent of W_t. The filtration is that generated by the Brownian W (d-dimensional) and the Poisson N (one-dimensional). The prices of the different securities and the coefficients relative to those prices (appreciation rates and volatilities) depend only on W (that is, are adapted to the filtration of W). The contingent claim B is taken to be positive bounded and to depend only on N. It is a contract that pays 1 if $N_T = 0$, and 0 if $N_T \neq 0$. Note that

$$B = \mathbf{1}_{N_T=0}.$$

The maximum price J_0 for B at time 0 is clearly equal to 1 (because it is impossible to hedge against the risk). To prove this result rigorously, choose P-martingale measure Q_α so that N_t is a Poisson process with intensity $1+\alpha$ under $Q_\alpha(\alpha > -1)$. Then it is easy to show that

$$J_0 = \sup_{\alpha\in]-1,+\infty[} E_Q\alpha(B) = \sup_{\alpha\in]-1,+\infty[} e^{-(1+\alpha)T} = 1.$$

Also, the maximum price J_t for B at time t will be equal to 0 if $N_t \geq 1$ (because then the event $N_T = 0$ is impossible) and 1 if $N_t = 0$.

It follows that the different processes of Theorem 2.3.1 are given by

- the price for B at time 0, $J_0 = 1$,
- the portfolio process $\varphi = 0$,
- the optional increasing process $f_t = N_{t\wedge T_1} = \mathbf{1}_{N_t\geq 1}$.

We see that $f_t = A_t - j_t$, where $j_t = -N_{t\wedge T_1} + t\wedge T_1$ is a purely discontinuous martingale with negative jumps, and $A_t = t\wedge T_1$ is a predictable increasing process. By defining the model on the canonical space, it is possible to show that the sequence Q_α converges weakly in distribution, as α tends to -1, to a probability measure that is not equivalent to P and under which the Poisson process is equal to 0 almost surely. We define the model more precisely below and give explicit calculations.

Let Ω_1 be the space of all $\mathbb{R}^d$-valued continuous functions on $\mathbb{R}^+$,

$$\Omega_1 = \mathcal{C}(\mathbb{R}^+, \mathbb{R}^d).$$

We denote by F^1 the σ-field generated by the coordinate process $W_t : \omega_1 \to \omega_1(t)$ for $t \geq 0$. Let $(F^{1_t}, t \geq 0)$ be the filtration generated by the process W_t. Let P^1 be the Wiener measure on Ω_1 constructed so that the coordinate mapping process W_t is Brownian motion.

Let Ω_2 be the space of all Radon positive measures on $\mathbb{R}^+$,

$$\Omega_2 = \mathcal{M}_+(\mathbb{R}^+).$$

We denote by F^2 the σ-field generated by all the random variables $N_t : \omega_2 \to \omega_2([0,t])$ for $t \geq 0$. Let (F_t^2) be the filtration generated by the process (N_t). Let P^2 be the probability measure on Ω_2 constructed so that the coordinate mapping process N_t is Poisson process with intensity 1.

Let (Ω, F, P) be the cross-product probability space $(\Omega_1 \times \Omega_2,\ F_1 \otimes F_2,\ P_1 \otimes P_2)$. The filtration F_t is defined by $F_t = F_t^1 \otimes F_t^2$. We denote by ω the elements of Ω

$$\omega = (\omega_1, \omega_2), \quad \omega_1 \in \Omega_1, \quad \omega_2 \in \Omega_2.$$

We denote by W_t the first coordinate mapping process and by N_t the second coordinate mapping process

$$W_t(\omega) = W_t(\omega_1) = \omega_1(t); \qquad N_t(\omega) = N_t(\omega_2) = \omega_2(0,t).$$

Under P, W_t is a (F_t)-Brownian motion, and N_t is a (F_t)-Poisson process with intensity 1.

The prices of the basic securities $P^i(t)$, the vector of stock appreciation rates $b(t)$, and the volatility matrix $\sigma(t)$ are taken to depend only on the first coordinate of the path (ω_1).

Let Q^1 be the probability measure that is equivalent to P^1 on F_T^1 such that

$$\widetilde{W_t} = W_t + \int_0^t \theta_s \, ds, \qquad 0 \leq t \leq T$$

is Brownian motion under Q^1. The reference probability measure Q on F_T is equal to $Q^1 \otimes P^2$. The contingent claim is equal to

$$B = \mathbf{1}_{N_T = 0}.$$

Let us calculate

$$J_0 = \sup_{M \in D} E_{Q^M}[\mathbf{1}_{N_T=0}].$$

Let J_0' be defined by

$$J_0' = \sup_{\alpha \in]-1, +\infty[} E_{P_\alpha^2}[\mathbf{1}_{N_T=0}].$$

Let P_α^2 be the probability measure defined on F_T^2 by

$$\frac{dP_\alpha^2}{dP^2} = \mathcal{E}(\alpha \widetilde{N})_T = e^{\log(1+\alpha)N_T - \alpha T},$$

where $\widetilde{N}_t = N_t - t$, $0 \le t \le T$. Let Q_α be defined by $Q_\alpha = Q^1 \otimes P^2_\alpha$. We have

$$\frac{dQ_\alpha}{dQ} = \mathcal{E}(\alpha\widetilde{N})_T \quad \text{and} \quad \alpha\widetilde{N} \in D,$$

hence, $J'_0 \le J_0$.

Let us show that $J'_0 = 1$. From the change of measure theorem for point processes ([Br-Ja, pp. 377–379]), we have that N is a Poisson process with intensity $1 + \alpha$ under P^2_α; hence, $P^2_\alpha(N_T = 0) = e^{-(1+\alpha)T}$ and

$$J'_0 = \sup_{\alpha \in]-1,+\infty[} P^2_\alpha(N_T = 0) = 1.$$

The supremum is obtained for $\alpha = -1$; hence

$$J_0 = 1.$$

Let us calculate

$$J_t = \text{ess}\sup_{M \in D} E_{Q^M}[\mathbf{1}_{N_T=0}/F_t].$$

Let J'_t be given by

$$J'_t = \text{ess} \sup_{\alpha \in]-1,+\infty[} E_{P^2_\alpha}[\mathbf{1}_{N_T=0}/F^2_t].$$

We have $J'_t \le J_t$. Let us determine J'_t:

$$E_{P^2_\alpha}[\mathbf{1}_{N_T=0}/F^2_t] = \mathbf{1}_{N_t=0}\, E_{P^2_\alpha}[\mathbf{1}_{N_T=0}/F^2_t] = \mathbf{1}_{N_t=0}\, e^{-(1+\alpha)(T-t)},$$

hence,

$$J'_t = \mathbf{1}_{N_t=0},$$

and the supremum is obtained for $\alpha = -1$. Now,

$$J_t \le \mathbf{1}_{N_t=0};$$

hence,

$$J_t = \mathbf{1}_{N_t=0} = 1 - N_{t\wedge T_1}.$$

The seller follows the following strategy. At $t = 0$, he receives 1 from the buyer. If the Poisson process remains equal to 0 until T, the seller does not make any profit; at time T, he gives 1 to the buyer. Otherwise, at the first instant the Poisson process is different from $0(T_1)$, the seller makes profit 1.

Let us determine the limit of the probability measures Q_α as $\alpha \to -1$:

$$\forall t \in R^+ \quad \text{and} \quad \lambda \in R^+, \lim_{\alpha \downarrow -1} E_{P^2_\alpha}[e^{-\lambda N_t}] = \lim_{\alpha \downarrow -1} e^{(1+\alpha)t(e^{-\lambda}-1)} = 1.$$

Also, for all $t_1, t_2, \ldots, t_k \in (R^+)^k$, the Laplace transform of $(N_{t_1}, \ldots, N_{t_k})$ tends to 1 under P^2_α as α tends to -1. Hence, $P^2_\alpha\,(N_{t_1}, \ldots, N_{t_k})^{-1}$ converges to the Dirac measure at zero on $(R^+)^k$ as α tends to -1, and $Q_\alpha = Q^1 \otimes P^2_\alpha$ converges in a weak sense to $Q_\alpha = Q^1 \otimes \delta_0$ as

α tends to -1. Under the limit probability measure $Q^1 \otimes \delta_0$, the increasing optional process f is equal to zero almost surely, that is,

$$Q^1 \otimes \delta_0(N_{t\wedge T_1} = 0) = 1,$$

$$J_t = 1, \quad Q^1 \otimes \delta_0 \text{ a.s.}$$

Hence, J is a martingale (constant) under the limit probability measure. But this probability measure is not equivalent to P. Notice that this probability measure makes the market complete (because it annuls the Poisson process).

The essential infimum of the possible prices at $t \geq 0$, $t < T$, is given by

$$\begin{aligned} K_t &= \text{ess} \inf_{\alpha > -1} Q^\alpha(N_T = 0/F_t), \\ &= \lim_{\alpha \to +\infty} Q^\alpha(N_T = 0/F_t), \\ &= 0. \end{aligned}$$

Hence, for each $t \in [0, T]$, we have

$$K_t = \mathbf{1}_{N_T=0}\, \mathbf{1}_{t=T}.$$

Let us determine the limit of the probability measures Q_α as $\alpha \to +\infty$. For all $t_1, t_2, \ldots, t_k \in (R^+)^k$, the Laplace transform of $(N_{t_1}, \ldots, N_{t_k})$ tends to zero under P^2_α as α tends to $+\infty$. Hence, Q_α converges in a weak sense to the null measure as $\alpha \to +\infty$. This result shows that an optimizing sequence of P-martingale measure does not necessarily converge to a probability measure.

A. Appendix.

A.1. A few useful (well-known) properties and theorems.

PROPOSITION 1.a. *A local martingale that is of class* (D) *is a martingale* (*see* [*De-Me*, *p.* 97, VI-30]).

Recall that if M and N are local martingales, their quadratic variation process $\langle M, N\rangle$ is defined only if MN is a special semimartingale. (For the definition, see [De-Me, p. 247, VII-39].) It always exists if one of the local martingales M, N is locally bounded by the following property (see [De-Me, p. 240, VII-32]).

PROPOSITION 1.b. *If X is a special semimartingale and Y is a locally bounded semimartingale, then XY is a special semimartingale.*

Proof. Using Doob's inequality, it is easy to show that $\sup_{s\leq t} |X_s Y_s|$ is locally integrable; it follows that XY is a special semimartingale (by using Thm. 25-d, p. 234 in [De-Me]). □

Notation. Let X_t be a local martingale (RCLL) under P with respect to $\{F_t\}$, such that $X_0 = 0$. We denote by $\mathcal{E}(X)_t$ the exponential of X, that is, the solution of the SDE

$$dU_t = U_{t^-}\, dX_t \quad \text{with}\, U_0 = 1.$$

The process $\mathcal{E}(X)$ is a local martingale under P.

We recall a general form of the Girsanov theorem that we shall use several times. (For more details see [De-Me, p. 259, VII-49] or [Mey, p. 377].)

Let P and Q be two probabilities equivalent on F_T, such that

$$\left.\frac{dQ}{dP}\right|_{F_T} = \mathcal{E}(N)_T,$$

where N is a local martingale which satisfies $N_0 = 0$. Let Z be a special semimartingale under P. It has the unique canonical decomposition under P

$$Z_t = Z_0 + M_t + A_t, \qquad 0 \leq t \leq T,$$

where M is a local martingale that satisfies $M_0 = 0$, and A is a predictable VF (finite variation) and RCLL process that satisfies $A_0 = 0$. Note that a supermartingale is special and that in this case, A_t is a decreasing process.

THEOREM 1.A. *Suppose that $\langle M, N\rangle$ exists. Then Z is a special semimartingale under Q and its canonical decomposition under Q is given by $Z_t = Z_0 + (M_t - \langle M, N\rangle_t) + (A_t + \langle M, N\rangle_t)$. The first term between brackets is a local martingale under Q, and the second one is a predictable finite variation process.*

COROLLARY 1.A. *Suppose that $\langle M, N\rangle$ exists. Then*

(i) *Z is a local martingale under Q if and only if the process $(A_t + \langle M, N\rangle_t)$ is equal to* 0.

(ii) *Z is a supermartingale under Q if and only if $(A_t + \langle M, N\rangle_t)$ is a decreasing process.*

A.2. Characterization of the essential supremum of all the possible prices (proofs). Let J_t be the essential supremum of the possible prices for B at time t and

$$J_t = \operatorname*{ess\ sup}_{R\in\mathcal{P}} E_R[B/F_t] = \operatorname*{ess\ sup}_{N\in D} E_{Q^N}[B/F_t].$$

We use dynamic programming methods [ElK] to solve the problem. Notice that (J_t) is not defined as a process yet because for each t, it is defined Q almost surely. Now,

$$\forall N \in D \quad \text{and} \quad t \in [0,T],\ E_{Q^N}[B/F_t] = E_{Q^{\tilde{N}}}[B/F_t],$$

where

$$\widetilde{N}_u = N_u - N_{t\wedge u},\ 0 \leq u \leq T.$$

Hence,

$$J_t = \operatorname*{ess\ sup}_{N\in D(t)} E_{Q^N}[B/F_t],$$

where $D(t) = \{N \in D/N_u = 0\ \forall u \in [0,t]\}$. For any $N \in D(t)$, put

$$\Gamma(t,N) = E_{Q^N}[B/F_t] = E[\mathcal{E}(N)_T B/F_t].$$

PROPOSITION 1. $\{\Gamma(t,N),\ N \in D(t)\}$ *is stable by supremum and infimum.*

By this property, it follows that for each t, there exists a sequence $N_p \in D(t)$ so that, almost surely, $\Gamma(t,N_p)$ is an increasing sequence of random variables that converges to J_t, that is,

$$J_t = \lim_{p\to+\infty} \uparrow \Gamma(t,N_p) = \lim_{p\to+\infty} \uparrow E_{Q^{N_p}}[B/F_t].$$

This property will allow us to invert supremum and expectation (using the monotone convergence theorem).

Proof. Let N_1, $N_2 \in D(t)$. There exists $N \in D(t)/\Gamma(t,N) = \Gamma(t,N_1) \vee \Gamma(t,N_2)$. Indeed, put $A = \{\Gamma(t,N_2) \geq \Gamma(t,N_1)\}$. We have $A \in F_t$. Put $N = N_1\mathbf{1}_{A^c} + N_2\mathbf{1}_A$. We have $N \in D(t)$;

$$\begin{aligned}\Gamma(t,N) &= E[\mathcal{E}(N_1)_T B/F_t]\mathbf{1}_{A^c} + E[\mathcal{E}(N_2)_T B/F_t]\mathbf{1}_A,\\ &= \Gamma(t,N_1)\mathbf{1}_{A^c} + \Gamma(t,N_2)\mathbf{1}_A.\end{aligned}$$

Hence, $\Gamma(t,N)=\Gamma(t,N_1)\vee\Gamma(t,N_2)$. □

For each t, take a sequence $N_p\in D(t)$ such that, Q almost surely,

$$J_t=\lim_{p\to+\infty}\uparrow\Gamma(t,N_p)=\lim_{p\to+\infty}\uparrow E_{Q^{N_p}}[B/F_t].$$

Now, J_t will denote a (F_t)-adapted process that is equal to the above limit almost everywhere.

PROPOSITION 2. *For any $N\in D$, (J_t) is a supermartingale under Q^N (that is, $\mathcal{E}(N)_t\,J_t$ is a supermartingale under Q).*

Proof. Let s, t be two positive reals such that $s<t\le T$. Take a sequence $N_p\in D(t)$ such that, Q almost surely,

$$J_t=\lim_{p\to+\infty}\uparrow\Gamma(t,N_p)=\lim_{p\to+\infty}\uparrow E_{Q^{N_p}}[B/F_t].$$

Let $N\in D$. Since we can invert limit and expectation (by the monotone convergence theorem), we have

$$\begin{aligned}E\left[\frac{\mathcal{E}(N)_t}{\mathcal{E}(N)_s}\,J_t/F_s\right]&=\lim_{p\to+\infty}\uparrow E\left[\frac{\mathcal{E}(N)_t}{\mathcal{E}(N)_s}\,E[\mathcal{E}(N_p)_TB/F_t]/F_s\right],\\&=\lim_{p\to+\infty}\uparrow E\left[\frac{\mathcal{E}(N)_t}{\mathcal{E}(N)_s}\,\mathcal{E}(N_p)_TB/F_s\right],\\&=\lim_{p\to+\infty}\uparrow E\left[\frac{\mathcal{E}(\widetilde{N}_p)_T}{\mathcal{E}(\widetilde{N}_p)_s}\,B/F_s\right],\end{aligned}$$

where $\widetilde{N}_p(u)=N(u\wedge t)+N_p(u)$, for $u\in[0,T]$. Now, $(\widetilde{N}_p(u),\ u\in[0,T])\in D$. Hence,

$$E\left[\frac{\mathcal{E}(N)_t}{\mathcal{E}(N)_s}\,J_t/F_s\right]\le J_s.\qquad\square$$

PROPOSITION 3. *(J_t) is the smallest supermartingale under Q^N, for any $N\in D$, which is equal to B at time T (unique up to a null set).*

Proof. Let (J'_t) be a supermartingale under Q^N, for any $N\in D$, which is equal to B at time T. Then,

$$\forall\,t\in[0,T]\quad\text{and}\quad N\in D,\ J'_t\ge E_{Q^N}[B/F_t],\ Q\text{ a.s.}$$

Hence,

$$\forall\,t\in[0,T],\ Q\text{ a.s., }J'_t\ge J_t.\qquad\square$$

We have also the following property.

PROPOSITION 4. *Let $\hat{N}$ be a local martingale that belongs to D. The following properties are equivalent:*

(i) *$\hat{N}$ is optimal, i.e., $\forall\,t\in[0,T]$, $J_t=E_{Q^{\hat{N}}}[B/F_t]$, Q a.s.*

(ii) *J_t is a martingale under $Q^{\hat{N}}$.*

PROPOSITION 5. *There exists an* RCLL *supermartingale still denoted by J_t so that for each $t\in[0,T]$,*

$$J_t=\operatorname*{ess\,sup}_{N\in D}E_{Q^N}[B/F_t].$$

Proof. Put $\mathbb{D} = [0,T] \cap \mathbb{Q}$. Because (J_t) is a supermartingale, we have that for P almost every ω, the mapping $t \longrightarrow J_t(\omega)$ defined on $\mathbb{D}$ has at each point t of $[0,T[$ a finite right limit

$$J_{t^+}(\omega) = \lim_{s\in\mathbb{D}, s\downarrow t} J_s(\omega)$$

and at each point of $]0,T]$ a finite left limit

$$J_{t^-}(\omega) = \lim_{s\in\mathbb{D}, s\uparrow t} J_s(\omega).$$

One can show (using a well-known property) that (J_{t^+}) is an (F_{t^+})-supermartingale under Q_N for all $N \in D$. Because the filtration is right continuous, J_{t^+} is an (F_t)-supermartingale under Q_N for all $N \in D$. Hence, by Proposition 3, for all $t \in [0,T]$, Q almost surely, $J_{t^+} \geq J_t$. Also, $J_t \geq E[J_{t^+}/F_t]$. Hence, Q almost surely, $J_t = J_{t^+}$ or else,

$$\forall t \in [0,T],\ J_{t^+} = \operatorname{ess}\sup_{N\in D} E_{Q^N}[B/F_t].$$

The result follows by taking J_t equal to the above process J_{t^+}. □

J_t is an RCLL process that satisfies

$$J_t = \operatorname{ess}\sup_{N\in D} E_{Q^N}[B/F_t].$$

J_t is characterized as the smallest right continuous supermartingale under Q_N, for every N belonging to D, which is equal to B at time T. Also, N is optimal if and only if J_t is a martingale under Q_N.

A.3. Generalization of the results of this paper. The results of Theorem 2.1.1, Theorem 2.1.2, Theorem 2.3.1, and Corollary 2.3.1 remain under the hypothesis

$$\sup_{N\in D} E_{Q^N}(B) < \infty.$$

(In fact, we shall see below that this hypothesis is equivalent to the fact that there exists a price admissible for sellers, or equivalently that B is smaller than the value of a self-financing portfolio, that is, B satisfies

$$B \leq X^{H,0,y}(T), \quad P\,\text{a.s.}$$

for some portfolio strategy H and initial investment $y \geq 0$.)

The whole proof of Theorem 2.1.1 still holds under the above hypothesis. In this case, J_t is not generally of class D. The results of Theorem 2.1.2, Proposition 2.3.1, and Theorem 2.3.1 still hold, but j is a Q-local martingale only (but not a martingale in general). The arguments of the proof still hold, but it is a bit more complicated technically because $\langle j \rangle$ is not always defined.

Proof of Theorem 2.1.2 *under the above hypothesis.* J_t is a Q-supermartingale; hence, it admits a unique decomposition as a local martingale M_t minus an increasing predictable process $A_t : J_t = M_t - A_t$. The local martingale M_t admits the following Kunita decomposition:

$$M_t = J_0 + \int_0^t \varphi_s^* \sigma_s \, d\widetilde{W}_s + j_t \quad \forall t \in [0,T],\ Q\,\text{a.s.}$$

for some predictable process φ and some Q-local martingale j such that

$$\left\langle j, \int_0^{\cdot} \sigma_i(s)dW_s \right\rangle_T = 0,\ Q\,\text{a.s.}, \quad \forall i \in \{1, \dots, n\}.$$

As in the proof of Theorem 2.1.2, we show that the continuous part j^c of j is equal to zero, using the following lemma (which follows from the fact that J_t is a supermartingale under each P-martingale measure).

LEMMA. *Let N be an element of D such that $\langle N, j\rangle$ exists. Then $A_t - \langle N, j\rangle_t$ is an increasing process.*

Now, $\langle j^c\rangle$ is locally integrable. By the Lebesgue Decomposition Theorem, there exist a positive predictable process h that belongs to $L^1_{\text{loc}}([0,T] \times \Omega, d\langle j^c\rangle_t\, dQ)$ and a locally integrable predictable increasing process B such that

$$dA_t = h_t d\langle j^c\rangle_t + dB_t$$

and such that, Q almost surely, the measure dB_t is singular with respect to $d\langle j^c\rangle_t$. Using the same arguments as in the proof of Theorem 2.1.2, we obtain the desired result. □

The proof of Theorem 2.3.1 under the weaker hypothesis is unchanged, and the result of Proposition 2.3.1 can be obtained by the same methods as before, but it is a bit longer because the lemma must be applied to some N such that $\langle N, j\rangle$ exists.

Remark. It follows from this, that the following properties are equivalent.

(i) $\sup_{N\in D} E_{Q^N}(B) < \infty$.

(ii) There exists a price admissible for sellers, or equivalently that B is smaller than the value of a self-financing portfolio, that is, B satisfies

$$B \le X^{H,0,y}(T),\ P\,\text{a.s.},$$

i.e.,

$$B \le y + \int_0^T H_s^* \sigma_s\, d\tilde{w}_s,\ Q\,\text{a.s.}$$

for some portfolio strategy H and initial investment $y \ge 0$.

Notice that the technical assumption on B given by

$$E_{Q^N}\left[\int_0^T \|\sigma_s^* H_s\|^2\, ds\right] < +\infty$$

for each $N \in D$ may be interpreted as the fact that the contingent claim B is not too risky. When we do not make the technical assumption on B, Corollary 2.3.1 still holds, but Corollary 2.3.2 must be replaced by the following.

COROLLARY 2.3.2$'$. (1) *For each $N \in D$, $E_{Q^N}(B) \le J_0 - E_{Q^N}(f_T)$.*

(2) *If N_n, $n \ge 0$, is an optimizing sequence belonging to D, that is, such that*

$$\lim_{n\to+\infty} E_{Q^{N_n}}[B] = J_0$$

then

$$\lim_{n\to\infty} E_{Q^{N_n}}(f_T) = 0.$$

Proof of (1). Let N be an element of D. The process given by

$$J_0 + \int_0^t \varphi_s^* \sigma_s \, d\widetilde{W}_s$$

is a positive continuous Q^N-local martingale, and hence a Q^N-supermartingale; this yields the inequality

$$J_0 + E_{Q^N}\left[\int_0^T \varphi_s^* \sigma_s \, d\widetilde{W}_s\right] \leq J_0.$$

Thus, f_T is Q^N-integrable and $E_{Q^N}(B) \leq J_0 - E_{Q^N}(f_T)$.

Proof of (2). Inequality (1) applied to the local martingales N_n gives the inequality

$$\forall n \in N,\ E_{Q^{N_n}}(B) \leq J_0 - E_{Q^{N_n}}(f_T).$$

Hence, if we let n tend to $+\infty$, we obtain the desired result

$$\lim_{n \to \infty} E_{Q^{N_n}}(f_T) = 0. \qquad \square$$

Also, when we do not make the technical assumption on B, Theorem 2.3.2 does not hold anymore, but we have the following result.

THEOREM 2.3.2′. *If* $\sup_{N \in D} E_{Q^N}[B]$ *is attained then* B *is attainable, that is, there exist a constant* x *and a portfolio* π *such that*

$$B = x + \int_0^T \pi_u^* \sigma_u \, d\widetilde{W}_u,\ Q \text{ a.s.}$$

Proof. Compare the proof of Theorem 2.3.2.

Thus, the contingent claims that satisfy the technical assumption (loosely speaking, those that are not too risky) are divided in two sets:

- The set of contingent claims that are attainable, which is equal to the set of contingent claims that admit a unique price.
- The set of contingent claims that are not attainable, which is equal to the set of contingent claims that admit several possible prices.

Things are not as clear for contingent claims that admit a finite selling price but do not satisfy the technical assumption. Even if they are attainable in the above sense, they may admit several possible prices. (Loosely speaking, this can be explained by the fact that, even if they are attainable, they can be too risky.) Nevertheless, we have the following properties that are equivalent.

(i) For each local martingale $N \in D$,

$$E_{Q^N}(B) = E_Q(B).$$

(ii) There exist a constant x and a portfolio π such that

$$B = x + \int_0^T \pi_u^* \sigma_u \, d\widetilde{W}_u,\ Q \text{ a.s.}$$

and such that the process given by

$$x + \int_0^{\cdot} \pi_u^* \sigma_u \, d\widetilde{W}_u$$

is a martingale under each P-martingale measure.

REFERENCES

[An-St] J. P. Ansel and C. Stricker, *Lois de martingale, densités et décomposition de Föllmer-Schweizer*, Ann. Inst. H. Poincaré, 28 (1992), pp. 375–392.

[Br-Ja] P. Bremaud and J. Jacod, *Processus ponctuels et martingales: Résultats récents sur la modélisation et le filtrage*. Adv. in Appl. Probab., 9 (1977), pp. 362–416.

[Ch-Mu] N. Christopeit and M. Musiela, *On the existence and characterization of arbitrage-free measures in contingent claim valuation*, SFB 303 discussion paper no. B-214.

[Duf] D. Duffie, *Security Markets: Stochastic Models*, Academic Press, New York, 1988.

[De-Me] C. Dellacherie and P. A. Meyer, *Probabilités et potentiel*, Théorie des Martingales, Hermann, Paris, 1980.

[ElK] N. El Karoui, *Les aspects probabilistes du contrôle stochastique*: *Ecole d'été Saint-Flour 1979*, Lecture Notes in Mathematics, Vol 876, 1981, pp. 74–238.

[El-Qu] N. El Karoui and M. C. Quenez, *Programmation dynamique et évaluation des actifs contingents en marché incomplet*, C. R. Acad. Sci. Paris, 331 (1991), pp. 851–854.

[Fö-Sc] H. Föllmer and M. Schweizer, *Hedging of contingent claims under incomplete information*, in Applied Stochastic Analysis, Stochastics Monographs Vol 5, M. H. A. Davis and R. J. Elliot, eds., Gordon and Breach, New York, 1991, pp. 389–414.

[Ha-Kr] J. M. Harrison and D. M. Kreps, *Martingales and arbitrage in multiperiod securities markets*. J. Econom. Theory, 20 (1979), pp. 381–408.

[Ha-Pl] J. M. Harrison and S. Pliska, *Martingales and stochastic integrals in the theory of continuous trading*, Stochastic Process. Appl., 15 (1983), pp. 313–316.

[He-Pe] H. He and N. D. Pearson, *Consumption and portfolio policies with incomplete markets and short-sale constraints*: *the infinite dimensional case*, J. Econom. Theory, 54 (1991), pp. 259–304.

[H-J-M] D. Health, R. Jarrow, and A. Morton, *Bond pricing and the term structure of interest rates*: *a new methodology for contingent claims valuation*, Econometrica, 60 (1992), pp. 77–105.

[Kar] I. Karatzas, *Optimization problems in the theory of continuous trading*, SIAM J. Control Optim., 27 (1989), pp. 1221–1259.

[KLSX] I. Karatzas, J. Lehoczky, S. Shreve, and G. L. Xu, *Martingale and duality methods for utility maximisation in an incomplete market*, SIAM J. Control Optim. 29 (1991) pp. 702–730.

[Kry] N. V. Krylov, *Controlled Diffusion Processes*, Springer-Verlag, Berlin, 1980.

[Mey] P. A. Meyer, *Un cours sur les intégrales stochastiques*, in Seminaire de Probabilités X, Lecture Notes in Mathematics 511, Springer-Verlag, New York, 1976, pp. 245–400.

[Pag] H. Pagès, *Optimal consumption and portfolio policies when markets are incomplete*, MIT mimeo, 1987.

[Pa-Pe] E. Pardoux and S. G. Peng, *Adapted solution of a backward stochastic differential equation*, Systems Control Lett., 14 (1990), pp. 55–61.

[Sch] M. Schweizer, *Mean-variance hedging for general claims*, Ann. Appl. Prob., 2 (1992), pp. 171–179.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 33, No. 1, pp. 67–88, January 1995

© 1995 Society for Industrial and Applied Mathematics

003

FINITE ELEMENT APPROXIMATIONS OF COMPENSATOR DESIGN FOR ANALYTIC GENERATORS WITH FULLY UNBOUNDED CONTROLS/OBSERVATIONS*

I. LASIECKA†

Abstract. An approximation theory leading to a design of a finite-dimensional compensator for control systems generated by analytic semigroups is presented. The novelty of this paper with respect to other results available in the literature is threefold: (i) it treats fully unbounded control/observation operators; (ii) it does not require compactness property of the underlined generator (an assumption that is often violated in practice); and (iii) the design of a finite-dimensional compensator is based on finite element approximation of the original model rather than on modal (eigenfunctions) approximations which, in turn, require the a priori knowledge of the eigenvalues for the system. Applications of the theory to heat equations and plate equations are provided.

Key words. finite element approximations, compensator design, Riccati equations, analytic semigroups, unbounded control/observation operators

AMS subject classifications. 49, 65

1. Introduction.

Consider the following control system:

$$\begin{cases} x_1 = Ax + Bu; \qquad x(0) = x_0 \in H, \\ y = Cx. \end{cases} \tag{1.1}$$

We make the following assumptions on (1.1).

(i) H, U, and Z are Hilbert spaces.

(ii) $A : H \supset \mathcal{D}(A) \to H$ is the generator of a strongly continuous analytic semigroup e^{At} on H; $t > 0$, generally unstable on H so that $\|e^{At}\|_{\mathcal{L}(H)} \leq Me^{wt}$; $t \geq 0$. We then consider throughout the translation $\hat{A} = -A + \hat{\lambda} I$, where $\hat{\lambda} > w$, so that $\hat{A}$ has well-defined fractional powers on H and $-\hat{A}$ is the generator of a strongly continuous analytic semigroup e^{-At} on H satisfying $\|e^{-\hat{A}t}\|_{\mathcal{L}(H)} \leq \hat{M}e^{\lambda t}$; $\lambda = w - \hat{\lambda}$.

(iii) $B : U \to (\mathcal{D}(A^*))'$ where $(\mathcal{D}(A^*))'$ is the dual of $\mathcal{D}(A^*)$ with respect to the H-topology. It is assumed that B is $\hat{A}^{*\gamma}$ bounded, or equivalently, as in the following hypothesis.

Hypothesis 1a. $(\hat{A})^{-\gamma} B \in \mathcal{L}(U;\, H)$; $0 \leq \gamma < 1$.

(iv) $C : \mathcal{D}(C) \subset H \to Z$ is a closed, densely defined operator such that the following hypothesis holds.

Hypothesis 1b. $C(\hat{A})^{-r} \in \mathcal{L}(H;\, Z)$; $0 < r < 1$.

Standard arguments from perturbation theory of analytic semigroups (see [K.1]) yield the fact that with any $F \in \mathcal{L}(H;\, U)$ and $K \in \mathcal{L}(Z;\, H)$, $A + BF$ and $A - KC$ generates an analytic semigroup on H. In what follows we make the following stabilizability-detectability assumptions.

Hypothesis 2. There exists $F \in \mathcal{L}(H;\, U)$ and $K \in \mathcal{L}(Z;\, H)$ such that

$$|e^{(A+BF)t}|_{\mathcal{L}(H)} + |e^{(A-KC)t}|_{\mathcal{L}(H)} \leq Me^{-w_1 t}$$

for some $w_1 > 0$. Under the above assumptions it is straightforward to show that there exists an infinite-dimensional "compensator," i.e.,

$$w_t = (A + BF - KC)w + KCx, \tag{1.2}$$

* Received by the editors June 9, 1992; accepted for publication (in revised form) August 2, 1993. This research was supported by National Science Fundation grant DMS-9204338.

† Department of Applied Mathematics, University of Virginia, Charlottesville, Virginia 22903.

such that the feedback control

$$u = Fw \tag{1.3}$$

exponentially stabilizes system (1.1). Precise statement of this fact is given in the following theorem.

THEOREM 1.1. *Let $\mathcal{H} \equiv H \times H$. The operator $\mathcal{A} : \mathcal{D}(\mathcal{A}) \subset \mathcal{H} \to \mathcal{H}$ given by*

$$\mathcal{A} \equiv \begin{pmatrix} A & BF \\ KC & A + BF - KC \end{pmatrix}$$

with $\mathcal{D}(\mathcal{A}) \equiv \{(x, y) \in \mathcal{H};\ Ax + BFy \in H;\ KCx + (A + BF - KC)\,y \in H\}$ generates an analytic and exponentially stable semigroup on $\mathcal{H}$.

Proof of Theorem 1.1. The proof, being a rather routine exercise in perturbation theory of analytic semigroups, is omitted.

The main goal of this paper is to construct *finite-dimensional* compensator w_h (which will be based on finite element approximations of the original problem) such that the *finite-dimensional* control feedback

$$u_h(t) = F_h w_h(t), \tag{1.4}$$

once inserted into the original system, will produce the solutions that are uniformly (with respect to h) exponentially stable.

The concept of a dynamic compensator is well known in the context of finite-dimensional systems and the idea goes back to Luenberger (see [L.3]). Many of these finite-dimensional ideas have been successfully generalized to infinite-dimensional systems—mainly of parabolic type (or, more generally, analytic semigroups); see [S.1], [C.1], [G.1], [G.2]. However, the techniques employed in these references are restricted to the cases when either the control/observation operators are bounded or, if unbounded (see [C.1]), the degree of unboundedness is severely limited—typically $\gamma + r \leq \frac{1}{2}$. On the other hand, there are many physically significant examples (heat equation with Dirichlet boundary control, strongly damped plate equations, etc.; see §6) where the above restrictions are not met. Also, the mathematical difficulties encountered in dealing with "fully" unbounded control operators (i.e., $\frac{1}{2} < \gamma < 1$) are much greater. Indeed, when $\gamma > \frac{1}{2}$, the basic solution operator $u \to x(t)$ is not defined from $L_2(0T;\ U) \to H$. This fact, recognized earlier in the context of the theory of Riccati equations (see for instance [B-D-D-M] and [L-T]), is a source of substantial technical difficulties. Thus, the main feature that distinguishes this paper from the other works is the fact that we treat *fully unbounded* control/observation operators, i.e., $\gamma + r < 1$. Moreover, in most papers available in the literature, the construction of finite-dimensional compensators is based on the knowledge of the eigenfunctions/eigenvalues of the generator A (see for instance [C.1]). For partial differential equation (PDE) systems in higher dimensions defined on arbitrary domains Ω, the eigenfunctions of the open-loop systems are generally not available. In view of this, our goal is to design finite-dimensional compensators based on *finite element* approximation of the original model and, as such, it would not require any specific knowledge of the spectrum of the original, usually unstable, generator. (Finite element approximations of compensator design with control operator B *bounded* were treated in [G.2] and references cited therein, while the case of *unbounded* operator B for hyperbolic-like dynamics was considered in [L.2].) Another attractive feature of our approach is that it does not require the compactness of the resolvent of the generator A—an assumption that has been used in an essential way in all previous works. In fact, as we will see in §6.3, there are examples of damped plate equations where this assumption is violated.

To establish the appropriate convergence and stability results for the compensator, the techniques recently developed in the context of finite element approximations of Riccati equations with fully unbounded control operators (see [L-T.1]) are used in an essential way. The outline of the paper is as follows. In §2 we introduce approximating subspaces and operators together with their properties. In §3 we formulate the main results of this paper. Sections 4 and 5 are devoted to the proofs of the main results. Section 6 provides several physical examples motivating the theory.

2. Approximating subspaces and operators. We introduce a family of approximating subspaces $V_h \subset H \cap \mathcal{D}(B^*) \cap \mathcal{D}(\hat{A}^r)$, where h is a parameter of discretization that tends to zero, $h \le h_0$. Let π_h be the orthogonal projection of H onto V_h with the usual approximating property, for some $s > 0$,

$$|\pi_h x - x|_H \le M h^s |x|_{\mathcal{D}(A)}. \tag{2.1}$$

Throughout this paper, M denotes a generic constant independent on $0 \le h \le h_0$.

Approximation of A. Let $A_h : V_h \to V_h$ be an approximation of A that satisfies the requirements in Assumptions 1 and 2.

Assumption 1 (uniform analyticity). For $a > w$, there exists $\sum(A, a) =$ closed triangular sector containing the axis $[-\infty, a]$ and delimited by two rays $a + \rho e^{\pm i Q_0}$ for some $\pi/2 < Q_0 < 2\pi$, and there exists h_a such that, if $\sum^c$ denotes the complement of $\sum$ in the complex plane, then for all $0 \le h \le h_a$, we have $\sigma(A_h) =$ spectrum of $A_h \subset \sum(A, a)$,

$$|R(\lambda, A_h)\hat{A}_h^Q \pi_h|_{\mathcal{L}(H)} \le \frac{M}{|\lambda - a|^{1-Q}} \qquad \forall \lambda \in \textstyle\sum^c(A, a). \tag{2.2}$$

Assumption 2. $|\pi_h \hat{A}^{-1} - \hat{A}_h^{-1} \pi_h|_{\mathcal{L}(H)} \le M h^s$.

Approximation assumptions on the operators B and C.

Assumption 3. (i) $|B^* x_h|_U \le M h^{-\gamma s} |x_h|_H$, $x_h \in V_h$;

(ii) $|\hat{A}^r x_h|_H \le M h^{-rs} |x_h|_H$.

Assumption 4. (i) $|B^*(\pi_h - I)x|_U \le M h^{s(1-\gamma)} |x|_{\mathcal{D}(A^*)}$;

(ii) $|\hat{A}^r(\pi_h - I)x|_H \le M h^{s(1-r)} |x|_{\mathcal{D}(A)}$.

Assumption 5 (i) $|B^* \pi_h x|_U \le M |\hat{A}^{*\gamma} x|_H$;

(ii) $|\hat{A}^* \pi_h x|_H \le M |\hat{A}^r x|_H$.

Approximation of F and K.

Assumption 6. The operator $F_h : V_h \to U$ satisfies one of the following conditions:

(i) either $F_h \pi_h \to F$ strongly and $B^* R(\lambda_0, A^*)$ is compact,

(ii) or $|F_h \pi_h - F|_{\mathcal{L}(H; U)} \to 0$ as $h \to 0$.

Assumption 7. Similarly for $K, K_h : Z \to V_h$,

(i) either $K_h \to K$ strongly and $CR(\lambda_0, A)$ is compact,

(ii) or $|K_h - K|_{\mathcal{L}(Z; H)} \to 0$ as $h \to 0$.

Consequences of approximating assumptions on A. From Assumptions 1 and 2, the following "rough" data estimates follow (see [L.1, Appendix], in the form to be used later):

$$\|R(\lambda, A) - R(\lambda, A_h)\pi_h\|_{\mathcal{L}(H)} \le M h^s, \qquad s > 0 \tag{2.3}$$

uniformly in $\lambda \in \sum^c(A, a)$,

$$|B^*[R(\lambda_0, A^*) - R(\lambda_0, A_h^*)\pi_h]|_{\mathcal{L}(H; U)} \le M h^{s(1-\gamma)}, \tag{2.4}$$

$$(2.5)\qquad |R(\lambda, A_h)\pi_h B|_{\mathcal{L}(U;H)} \le \frac{M}{|\lambda - a|^{1-\gamma}},$$

where the last two inequalities follow from Lemma 3.1 in [L-T.1]. From Assumptions 4 and 5, by interpolation, we obtain

$$(2.6)\qquad |\hat{A}^r(\pi_h - I)x|_H \le Mh^{s(Q-r)}\,|\hat{A}^Q x|_H; \qquad r \le Q \le 1.$$

3. Statement of main results. Consider the following finite-dimensional dynamic compensator:

$$(3.1)\qquad w_t = (A_h\pi_h + \pi_h BF_h\pi_h - K_hC)w + K_hCx.^1$$

We associate (3.1) with the control system

$$(3.2)\qquad \begin{cases} x_t = Ax + BF_h\pi_h w, \\ x(0) = x_0 \in H. \end{cases}$$

Our main results are contained in the following theorem.

THEOREM 3.1. *Assume Hypotheses* 1 *and* 2 *and Assumptions* 1–7. *Moreover, assume* $\gamma + r < 1$. *Then system* (3.1), (3.2) *represented by*

$$\mathcal{A}_h \equiv \begin{bmatrix} A & BF_h\pi_h \\ K_hC & A_h\pi_h + \pi_h BF_h\pi_h - K_hC \end{bmatrix}$$

generates a uniformly analytic and uniformly exponentially stable semigroup, i.e.,

$$|e^{\mathcal{A}_h t}|_{\mathcal{L}(\mathcal{H})} \le Me^{-\omega_0 t} \quad \textit{for some } \omega_0 > 0,$$

where the constants M *and* ω_0 *are* uniform *with respect to* $0 < h \le h_0$.

Remark 3.1. In (3.1) one could also approximate the operator B by its approximation $B_h \in \mathcal{L}(U; V_h)$ subject to the usual consistency assumptions (see [L-T.1]). The analysis of the problem is the same.

The problem of practical interest is how to find the stabilizing feedbacks F_h and K_h. If the knowledge of unstable eigenvalues is available then this could be accomplished by a routine pole assignment procedure (see [C.1], [S.1]). Since, in general, a precise knowledge of eigenfunctions is not known, a different approach based on approximation of the algebraic Riccati equation are pursued (see [G.1] for the case of bounded control/observation operators). To accomplish this we recall recent results from [L-T] and [L-T.1] on solvability and approximations of Riccati equations arising in control dynamics with unbounded input/output operators.

We associate the following algebraic Riccati equation with the dynamics (1.1):

$$(3.3)\qquad (A^*Px, y)_H + (PAx, y)_H + (x, y)_H = (B^*Px, B^*Py)_U \quad \text{for } x, y \in \mathcal{D}(A).$$

By virtue of Hypotheses 1 and 2 we know (see [D-I], [F.2], [F.3], [L-T], [B-D-D-M]) that there exists a unique solution $P \in \mathcal{L}(H)$, $P = P^*$ such that

$$(3.4)\qquad B^*P \in \mathcal{L}(H; U),$$

[1] $\pi_h B \subset V_h$ is defined, as usual, by duality: $(\pi_h Bg, v_h)_H = (g, B^*v_h)_U$ for all $v_h \in V_h$, $g \in U$.

(3.5) $e^{(A-BB^*P)t}$ generates an analytic and exponentially stable semigroup.

On the other hand, the approximating properties (Assumptions 1–7) guarantee (see [L-T.1]) that the equation

$$A_h^* P_h + P_h A_h + \pi_h = P_h \pi_h BB^* P_h \tag{3.6}$$

is uniquely solvable with $P_h \in \mathcal{L}(V_h)$. Moreover,

$$B^* P_h \pi_h \in \mathcal{L}(H;\, U) \quad \text{uniformly in } h > 0 \tag{3.7}$$

and (see [L-T.1, Thm. 1.1])

$$|P_h \pi_h - P|_{\mathcal{L}(H)} + |B^*(P_h \pi_h - P)|_{\mathcal{L}(H;\, U)} \to 0 \quad \text{as } h \to 0. \tag{3.8}$$

The same argument can be repeated for the "dual" Riccati equation

$$(AQx, y)_H + (QA^*x, y)_H + (x, y)_H = (C^*Qx, C^*Qy)_H; \qquad x, y \in \mathcal{D}(A^*), \tag{3.9}$$

which, in view of Hypotheses 1 and 2, yields the unique solution $Q \in \mathcal{L}(H)$ such that

$$CQ \in \mathcal{L}(H;\, Z), \tag{3.10}$$

$$e^{(A^* - C^* CQ)t} \quad \text{is exponentially stable.} \tag{3.11}$$

Moreover, in view of Assumptions 1–5, approximation theory of [L-T.1] applies to provide the existence of $Q_h^* = Q_h \in \mathcal{L}(V_h)$ such that

$$A_h Q_h + Q_h A_h^* + \pi_h = Q_h CC^* Q_h \tag{3.12}$$

is satisfied and, moreover,

$$|CQ_h \pi_h|_{\mathcal{L}(H;\, Z)} \le M, \tag{3.13}$$

$$|C(Q_h \pi_h - Q)|_{\mathcal{L}(H;\, Z)} \to 0 \quad \text{as } h \to 0. \tag{3.14}$$

Thus we are in a position to apply the results of Theorem 3.1 with the specific feedbacks

$$\begin{cases} F_h \equiv B^* P_h, \\ K_h = Q_h \pi_h C^*. \end{cases} \tag{3.15}$$

Indeed, parts (ii) of Assumptions 6 and 7 are satisfied by virtue of (3.8) and (3.14), with $F \equiv B^*P$; $K \equiv QC^*$ and F_h, K_h as in (3.15). We have thus obtained the following corollary.

COROLLARY 3.1. *Assume Hypotheses* 1 *and* 2, *and Assumptions* 1–5, $\gamma + r < 1$. *Then the conclusion of Theorem* 3.1 *holds with* F_h *and* K_h *given by* (3.15).

4. Technical lemmas and the proof of Theorem 3.1. We define

$$A_{hB} \equiv A_h \pi_h + \pi_h B F_h \pi_h,$$

$$A_B \equiv A + B F_h \pi_h,$$

$$A_C \equiv A - K_h C.$$

The following result is a consequence of Theorem 4.2 in [L-T.1].

PROPOSITION 4.1 *There exist constants $M > 0$, $w_0 > 0$ such that*

$$|R(\lambda, A_{hB})|_{\mathcal{L}(H)} \leq \frac{M}{|\lambda + w_0|}, \tag{4.1}$$

$$|R(\lambda, A_B)|_{\mathcal{L}(H)} \leq \frac{M}{|\lambda + w_0|}, \tag{4.2}$$

$$|R(\lambda, A_C)|_{\mathcal{L}(H)} \leq \frac{M}{|\lambda + w_0|}, \tag{4.3}$$

where the above estimates hold uniformity in $h < h_0$ and $\lambda \in \sum^c$, where $\sum \equiv$ closed triangular sector containing the axis $(-\infty, -w_0)$ and delimited by two rays $-w_0 + \rho\, e^{\pm Qi}$ for some $\pi/2 < Q < 2\pi$. Here $w_0 < w_1$ where, we recall, w_1 is the margin of stability for $A + BF$ and $A - KC$.

Proof. By virtue of Assumption 6, for some $\lambda_0 \in \rho(A)$,

$$|R(\lambda_0, A) B (F_h \pi_h - F)|_{\mathcal{L}(H)} \to 0, \quad \text{as}\, h \to 0. \tag{4.4}$$

Hence we are in a position to apply Theorem 4.2 of [L-T.1] to conclude (4.1). To assert (4.3) it is enough to note that Assumption 7 implies

$$|(K_h^* - K^*) C R(\lambda_0, A^*)|_{\mathcal{L}(H)} \to 0 \quad \text{as}\, h \to 0. \tag{4.5}$$

Now conclusion of Theorem 4.2 applied to the operator $A^* - C^* K_h^*$ yields the desired inequality in (4.3). The estimate (4.2) holds by virtue of (4.4) and Remark 4.3 in [L-T.1]. □

The proof of Theorem 3.1 is based on the following three lemmas.

LEMMA 4.1. *We have that*

$$|A R(\lambda, A_C)|_{\mathcal{L}(H)} \leq M. \tag{4.6}$$

Let $\gamma + r < 1$. Then

$$|\hat{A}^r R(\lambda, A_C) \hat{A}^\gamma|_{\mathcal{L}(H)} \leq M, \tag{4.7}$$

$$|\hat{A}^r R(\lambda, A_B) \hat{A}^\gamma|_{\mathcal{L}(H)} \leq M, \tag{4.8}$$

where the bounds are uniform in $h \leq h_0$; $\lambda \in \sum^c$.

LEMMA 4.2. *Let $\gamma + r < 1$. Then we have the following.*

(i) *For all* $\lambda \in \sum^c(A,a)$, *the following inequality holds uniformly in* λ *and* $h \le h_0$:

$$|\hat{A}^r[R(\lambda, A_{hB}) - R(\lambda, A_B)]|_{\mathcal{L}(H)} \le \frac{M}{|\lambda - a|^{1-r}}. \tag{4.9}$$

(ii) *For any* $\lambda_0 \in \sum^c(A,a)\,(\lambda_0$ *fixed), the following inequality holds uniformly in* $h \le h_0$:

$$|\hat{A}^r[R(\lambda_0, A_{hB}) - R(\lambda_0, A_B)]|_{\mathcal{L}(H)} \le Mh^{s(1-r-\gamma)}. \tag{4.10}$$

(iii) *The estimate in* (4.10) *can be extended to hold uniformly in* λ_0 *for all* λ_0 *such that*

$$\lambda_0 \in \textstyle\sum^c \cap \{\lambda_0;\ |\lambda_0| \le R_0\} \quad \textit{where } R_0 \textit{ is fixed.} \tag{4.11}$$

The proofs of Lemmas 4.1 and 4.2, being technical, are relegated to §5.

Let us introduce the operator $T_h(\lambda) : \mathcal{D}(C) \to H$ defined by

$$T_h(\lambda) \equiv R(\lambda, A_C)(A + BF_h\pi_h - \lambda I)[R(\lambda, A_{h,B}) - R(\lambda, A_B)]K_hC. \tag{4.12}$$

We prove the following lemma.

LEMMA 4.3. *Assume* $\gamma + r < 1$. *Then*

(i) $|T_h(\lambda)|_{\mathcal{L}(\mathcal{D}(\hat{A}^r))} \le M$,

(ii) $|(I - T_h(\lambda))^{-1}|_{\mathcal{L}(\mathcal{D}(\hat{A}^r))} \le M$, *where the constant* M *is uniform in* $h < h_0$ *and* $\lambda \in \sum^c$.

Proof. To prove the lemma it suffices to show that

$$|T_h(\lambda)|_{\mathcal{L}(\mathcal{D}(\hat{A}^r))} < \frac{1}{2} \quad \text{uniformly in } h < h_0 \text{ and } \lambda \in \textstyle\sum^c. \tag{4.13}$$

From part (i) of Lemma 4.2 we obtain, for $\lambda \in \sum^c(A,a)$,

$$|\hat{A}^r[R(\lambda, A_{hB}) - R(\lambda, A_B)]K_hC|_{\mathcal{L}(H)} \le \frac{M}{|\lambda - a|^{1-r}}|C|_{\mathcal{L}(\mathcal{D}(\hat{A}^r);H)}. \tag{4.14}$$

On the other hand,

$$R(\lambda, A_C)(A + BF_h\pi_h - \lambda I) = -I + R(\lambda, A_C)(BF_h\pi_h + K_hC). \tag{4.15}$$

From (4.6) in Lemma 4.1,

$$|\hat{A}^r R(\lambda, A_C)K_hC|_{\mathcal{L}(\mathcal{D}(\hat{A}^r);H)} \le M|C|_{\mathcal{L}(\mathcal{D}(\hat{A}^r);H)} \le M, \tag{4.16}$$

and by (4.7) together with Hypothesis 1a and Assumption 6,

$$|\hat{A}^r R(\lambda, A_C)\hat{A}^\gamma \hat{A}^{-\gamma} BF_h\pi_h| \le M. \tag{4.17}$$

Combining (4.14)–(4.17) with the definition of $T_h(\lambda)$ given in (4.12) yields

$$|T_h(\lambda)|_{\mathcal{L}(\mathcal{D}(\hat{A}^r))} \le \frac{M}{|\lambda - a|^{1-r}} \tag{4.18}$$

which, in particular, implies

$$|T_h(\lambda)|_{\mathcal{L}(\mathcal{D}(A^r))} < \frac{1}{2} \quad \text{for } \lambda \in \textstyle\sum^c(A,a) \text{ and } |\lambda| > R_0,$$

where R_0 is sufficiently large. Consider next $\lambda \in \sum^c \cap \{|\lambda| \leq R_0\}$. By repeating the same arguments as above, with the only difference that instead of result (i) of Lemma 4.2 we use result (iii), we obtain

$$|T_h(\lambda_0)|_{\mathcal{L}(\mathcal{D}(\hat{A}^r))} < Ch^{s(1-\gamma-r)} < \frac{1}{2} \tag{4.19}$$

for $h < h_0$ where $\lambda_0 \in \sum^c \cap \{|\lambda| \leq R_0\}$, which then leads to the desired conclusions of Lemma 4.3. □

Proof of Theorem 3.1. We use the following transformation introduced in [S.1]:

$$Z \equiv \begin{pmatrix} I & I \\ 0 & I \end{pmatrix}; \qquad Z^{-1} = \begin{pmatrix} I & -I \\ 0 & I \end{pmatrix}, \quad \text{where } I \text{ is an identity on the Hilbert space } H.$$

To prove Theorem 3.1 it is enough to show the estimate

$$|R(\lambda, \mathcal{A}_h)|_{\mathcal{L}(\mathcal{H})} \leq \frac{M}{|\lambda + \omega_0|}, \qquad \lambda \in \sum^c, \tag{4.20}$$

or equivalently,

$$|R(\lambda, \hat{\mathcal{A}}_h)|_{\mathcal{L}(\mathcal{H})} \leq \frac{M}{|\lambda + \omega_0|}, \quad \text{where } \hat{\mathcal{A}}_h \equiv Z^{-1}\mathcal{A}_h Z, \tag{4.21}$$

holds with some $\omega_0 > 0$. It is straightforward to verify that

$$\hat{\mathcal{A}}_h = \begin{pmatrix} A - K_hC, & A - A_h\pi_h + (I - \pi_h)BF_h\pi_h \\ K_hC, & A_h\pi_h + \pi_hBF_h\pi_h \end{pmatrix}.$$

We compute $R(\lambda, \hat{\mathcal{A}}_h)$. Let

$$(\mathcal{A}_h - \lambda I)^{-1}\begin{pmatrix} f \\ g \end{pmatrix} = \begin{pmatrix} x(\lambda) \\ w(\lambda) \end{pmatrix}, \quad \text{hence}$$

$$\begin{cases} (A - K_hC - \lambda I)x(\lambda) + (A - A_h\pi_h + (I - \pi_h)BF_h\pi_h)w(\lambda) = f, \\ K_hCx(\lambda) + (A_h\pi_h + \pi_hBF_h\pi_h - \lambda I)w(\lambda) = g. \end{cases} \tag{4.22}$$

The result of Proposition 4.1 is

$$\begin{aligned} &\text{(i)} \quad w(\lambda) = R(\lambda, A_{hB})[-g + K_hCx(\lambda)], \\ &\text{(ii)} \quad x(\lambda) = R(\lambda, A_C)[-f + (A - A_h\pi_h + (I - \pi_h)BF_h\pi_h)w(\lambda)]. \end{aligned} \tag{4.23}$$

We show that

$$R(\lambda, A_{hB}) - R(\lambda, A_B) = -R(\lambda, A_B)(A - A_h\pi_h + (I - \pi_h)BF_h\pi_h)R(\lambda, A_{hB}). \tag{4.24}$$

Indeed,

$$\begin{aligned} &R(\lambda, A_B)[A - A_h\pi_h + (I - \pi_h)BF_h\pi_h]R(\lambda, A_{hB}) \\ &\quad = R(\lambda, A_B)[A_B - \lambda I - A_{hB} + \lambda I]R(\lambda, A_{hB}) \\ &\quad = R(\lambda, A_B) - R(\lambda, A_{hB}) \end{aligned}$$

as desired for (4.24). Combining (4.23) with (4.24) we arrive at

$$x(\lambda) = R(\lambda, A_C)(A + BF_h\pi_h - \lambda I)[R(\lambda, A_{hB}) - R(\lambda, A_B)]K_hCx(\lambda),$$

$$- R(\lambda, A_C)f - R(\lambda, A_C)(A + BF_h\pi_h - \lambda I)[R(\lambda, A_{hB}) - R(\lambda, A_B)]g, \quad \text{or} \tag{4.25}$$

(4.26)

$$\begin{cases} x(\lambda) = -[I - T_h(\lambda)]^{-1}[R(\lambda, A_C)f \\ \qquad\qquad\qquad + R(\lambda, A_C)(A + BF_h\pi_h - \lambda I)[R(\lambda, A_{hB}) - R(\lambda, A_B)]g], \\ w(\lambda) = R(\lambda, A_{hB})[-g + K_hCx(\lambda)]. \end{cases}$$

Noting that

$$R(\lambda, A_C)(A + BF_h\pi_h - \lambda I) = -I + R(\lambda, A_C)(BF_h\pi_h + K_hC),$$

we obtain, by Lemma 4.2 and Lemma 4.1,

(4.27)

$$\begin{aligned} &|\hat{A}^r R(\lambda, A_C)(A + BF_h\pi_h - \lambda I)[R(\lambda, A_{hB}) - R(\lambda, A_B)]g|_H \\ &\quad \leq |\hat{A}^r[R(\lambda, A_{hB}) - R(\lambda, A_B)]g|_H \\ &\qquad + |\hat{A}^r R(\lambda, A_C)(BF_h\pi_h + K_hC)[R(\lambda, A_{hB}) - R(\lambda, A_B)]g|_H \\ &\quad \leq \{M + |\hat{A}^r R(\lambda, A_C)\hat{A}^\gamma \hat{A}^{-\gamma} BF_h\pi_h|_{\mathcal{L}(H)} + |\hat{A}^r R(\lambda, A_C)K_hC\hat{A}^{-r}|_{\mathcal{L}(H)}\}|g|_H \\ &\quad \leq M|g|_H. \end{aligned}$$

From (4.27), (4.26), and Lemma 4.3 we then infer that

$$|\hat{A}^r x(\lambda)|_H \leq M[|g|_H + |f|_H], \tag{4.28}$$

and by Proposition 4.1 with (4.26),

$$|w(\lambda)|_H \leq \frac{M}{|\lambda + w_0|}[|g|_H + |f|_H] \quad \text{for } \lambda \in \sum\nolimits^c. \tag{4.29}$$

Going back to (4.25) and recalling (4.15), we obtain

(4.30)

$$\begin{aligned} &|x(\lambda)|_H \\ &\quad \leq M[|R(\lambda, A_C)|_{\mathcal{L}(H)}|f|_H + |[I - R(\lambda, A_C)\cdot(BF_h\pi_h + K_hC)](R(\lambda, A_{hB}) \\ &\qquad - R(\lambda, A_B))[g - K_hCx(\lambda)]|_H] \\ &\quad \leq M\{|R(\lambda, A_C)|_{\mathcal{L}(H)}|f|_H + |R(\lambda, A_C)\hat{A}^\gamma|_{\mathcal{L}(H)}\cdot|\hat{A}^{-\gamma}BF_h\pi_h|_{\mathcal{L}(H)}(|R(\lambda, A_{hB})|_{\mathcal{L}(H)} \\ &\qquad + |R(\lambda, A_B)|)_{\mathcal{L}(H)}\cdot|g - K_hCx(\lambda)|_H \\ &\qquad + |R(\lambda, A_C)|_{\mathcal{L}(H)}|K_hC\hat{A}^{-r}|_{\mathcal{L}(H)}\cdot|\hat{A}^r(R(\lambda, A_{hB}) \\ &\qquad - R(\lambda, A_B))|_{\mathcal{L}(H)}|g - K_hCx(\lambda)|_H \\ &\qquad + |R(\lambda, A_{hB}) - R(\lambda, A_B)|_H|g - K_hCx(\lambda)|_H\} \\ &\quad \leq \frac{M}{|\lambda + w_0|}[|f|_H + |g|_H], \end{aligned}$$

where in the last inequality, we have used once more the result of Proposition 4.1, Lemma 4.2, Lemma 4.1, and (4.28).

Inequalities (4.29) and (4.30) yield the desired conclusion in (4.21) with $\omega_0 = w_0 > 0$. □

5. Proofs of Lemmas 4.1 and 4.2. We begin with a sequence of preliminary estimates.

PROPOSITION 5.1. *Let $T : H \to (\mathcal{D}(A^*))'$ be such that $\hat{A}^{-\gamma}T \in \mathcal{L}(H)$. Then for any $\bar{\gamma} > \gamma$ there exists a constant $M > 0$ such that*

$$(5.1) \qquad |\hat{A}^{1-\bar{\gamma}}(\lambda_0 I - A - T)^{-(1-\gamma)}|_{\mathcal{L}(H)} \leq M$$

for all $\lambda_0 \in \sum^c(A + T)$, where $\sum(A + T)$ is a closed triangular sector containing in its interior the spectrum of $A + T$.

Proof. By the assumption and analyticity of A, $A + T$ generates an analytic semigroup [K.1].

Step 1. For $\mu \in \rho(A + T) \cap \rho(A)$,

$$(5.2) \qquad (\mu I - A - T)^{-1} = (I - R(\mu, A)T)^{-1}R(\mu, A).$$

On the other hand, by the definition of fractional powers of the operators (see [P.1], [K.1]),

$$(5.3) \qquad (\lambda_0 I - A - T)^{-(1-\gamma)} = \frac{1}{2\pi i}\int_\Gamma z^{-(1-\gamma)}[(\lambda_0 - z)I - A - T]^{-1}\,dz,$$

where Γ denotes the triangular path that runs in $\rho(\lambda_0 I - A - T) \cap \rho(\lambda_0 I - A)$, avoiding the real negative axis and the origin. Hence

$$(5.4) \qquad \begin{aligned} &\hat{A}^{1-\bar{\gamma}}(\lambda_0 I - A - T)^{-(1-\gamma)} \\ &= \frac{1}{2\pi i}\hat{A}^{1-\bar{\gamma}}\int_\Gamma [I - R(\lambda_0 - z;\, A)T]^{-1}R(\lambda_0 - z, A)z^{-(1-\gamma)}\,dz \\ &= \frac{1}{2\pi i}\int_\Gamma \hat{A}^{1-\bar{\gamma}}[I - R(\lambda_0 - z, A)T]^{-1}\hat{A}^{-(1-\bar{\gamma})}\hat{A}^{1-\bar{\gamma}}R(\lambda_0 - zA)z^{-(1-\gamma)}\,dz. \end{aligned}$$

Step 2. We show that for $\mu \in \rho(A) \cap \rho(A + T)$ and for $\bar{\gamma} \geq \gamma$, the following bound holds:

$$(5.5) \qquad |\hat{A}^{1-\bar{\gamma}}[I - R(\mu, A)T]^{-1}\hat{A}^{-(1-\bar{\gamma})}|_{\mathcal{L}(H)} \leq M$$

uniformly in $\mu \in \mathcal{D}$, where the set $\mathcal{D}$ is any closed set contained in $\rho(A) \cap \rho(A + T)$.

Since

$$|R(\mu, A)T|_{\mathcal{L}(H)} \leq \frac{M}{|\mu - a|^{1-\gamma}}$$

for $|\mu|$ large enough, say $|\mu| > R$, we have

$$(5.6) \qquad |[I - R(\mu, A)T]^{-1}|_{\mathcal{L}(H)} \leq M_R.$$

The above estimate can easily be extended to hold for all $\mu \in \mathcal{D}$ (with the bound uniform in μ). Indeed, for $\mu \in \rho(A + T) \cap \rho(A)$, $I - R(\mu, A)T$ is injective. Thus, in order to show (5.6) it suffices to prove that the range of $I - R(\mu, A)T$ is all of H. But this can be accomplished by routine computations involving the resolvent equation. Thus for all $\mu \in \mathcal{D}$ we can solve

$$(5.7) \qquad \hat{A}^{1-\bar{\gamma}}[I - R(\mu, A)T]^{-1}\hat{A}^{-(1-\bar{\gamma})}x = y$$

with the estimate

$$(5.8) \qquad |\hat{A}^{-(1-\bar{\gamma})}y|_H \leq M|\hat{A}^{-(1-\bar{\gamma})}x|_H.$$

On the other hand, from (5.7),

$$x = y - \hat{A}^{(1-\bar{\gamma})} R(\mu, A) T \hat{A}^{-(1-\bar{\gamma})} y; \tag{5.9}$$

and by (5.8), the analyticity of A, and the assumption assumed on T,

$$\begin{aligned} |\hat{A}^{(1-\bar{\gamma})} R(\mu, A) T \hat{A}^{-(1-\bar{\gamma})} y|_H &\le M |\hat{A}^{(1-\bar{\gamma})} R(\mu, A) \hat{A}^{\bar{\gamma}} \hat{A}^{-\bar{\gamma}} T|_{\mathcal{L}(H)} |\hat{A}^{-(1-\bar{\gamma})} x|_H \\ &\le M |\hat{A}^{-(1-\bar{\gamma})} x|_H. \end{aligned} \tag{5.10}$$

Combining (5.9) and (5.10) yields the desired conclusion in (5.5).

Step 3. By combining (5.4) with (5.5) and noting that $\lambda_0 - z \in \mathcal{D}$ for $z \in \Gamma$ we obtain

$$\begin{aligned} |\hat{A}^{1-\bar{\gamma}} (\lambda_0 I - A - T)^{-(1-\gamma)}|_{\mathcal{L}(H)} &\le M \int_\Gamma |\hat{A}^{1-\bar{\gamma}} R(\lambda_0 - z, A)|_{\mathcal{L}(H)} z^{-(1-\gamma)}\, dz \\ &\le M \int_\Gamma \frac{dz}{|-z + \lambda_0 - a|^{\bar{\gamma}} |z|^{1-\gamma}} \le M. \qquad \square \end{aligned}$$

PROPOSITION 5.2. *We have that*

$$|\hat{A}^r R(\lambda, A_h) \pi_h|_{\mathcal{L}(H)} \le \frac{M}{|\lambda - a|^{1-r}}, \tag{5.11}$$

$$|\hat{A}^r R(\lambda, A_h) \pi_h B F_h \pi_h|_{\mathcal{L}(H)} \le M \quad \textit{for } \gamma + r < 1, \tag{5.12}$$

where the estimates are uniform for all $h \le h_0$ *and* $\lambda \in \sum^c(A, a)$.

Proof. Inequality (5.11) is a consequence of Assumptions 1 and 3, and (2.3), and it can be proved by routine arguments (see [L-T.1]).

To prove (5.12), we use duality

$$\begin{aligned} |B^* R(\lambda, A_h^*) \pi_h \hat{A}^{*r}|_{\mathcal{L}(H;U)} &\le |B^* [R(\lambda, A_h^*) \pi_h - \pi_h R(\lambda, A^*)] \hat{A}^{*r}|_{\mathcal{L}(H;U)} \\ &\quad + |B^* \pi_h R(\lambda, A^*) \hat{A}^{*r}|_{\mathcal{L}(H;U)}. \end{aligned} \tag{5.13}$$

From Assumption 5 and the analyticity of A^*,

$$|B^* \pi_h R(\lambda, A^*) \hat{A}^{*r}|_{\mathcal{L}(H;U)} \le C |R(\lambda, A^*) \hat{A}^{*r+\gamma}|_{\mathcal{L}(H)} \le \frac{M}{|\lambda - a|^{1-(\gamma+r)}} \le M. \tag{5.14}$$

From Assumption 3,

$$\begin{aligned} &|B^* [R(\lambda, A_h^*) \pi_h - \pi_h R(\lambda, A^*)] \hat{A}^{*r}|_{\mathcal{L}(H;U)} \\ &\quad \le M h^{-\gamma s} |\hat{A}^r [R(\lambda, A_h) \pi_h - R(\lambda, A) \pi_h]|_{\mathcal{L}(H)} \\ &\quad \le M h^{-\gamma s} |\hat{A}^r [R(\lambda, A_h) \pi_h - \pi_h R(\lambda, A) \pi_h]|_{\mathcal{L}(H)} \\ &\quad\quad + M h^{-\gamma s} |\hat{A}^r (I - \pi_h) R(\lambda, A) \pi_h|_{\mathcal{L}(H)}. \end{aligned} \tag{5.15}$$

As a consequence of Assumption 3, (2.3), and Assumption 1, we obtain

$$|\hat{A}^r [R(\lambda, A_h) \pi_h - \pi_h R(\lambda, A)]|_{\mathcal{L}(H)} \le M h^{s(1-r)}. \tag{5.16}$$

By (5.16), (2.6) applied with $Q = 1$ and (5.15),

$$\begin{aligned} &|B^* [R(\lambda, A_h^*) \pi_h - \pi_h R(\lambda, A^*)] \hat{A}^{*r}|_{\mathcal{L}(H,U)} \\ &\quad \le M h^{s(1-\gamma-r)} + C h^{s(1-\gamma-r)},|A R(\lambda, A)|_{\mathcal{L}(H)} \le M. \end{aligned} \tag{5.17}$$

Combining (5.13) with (5.14) and (5.17) and recalling the assumptions imposed on F_h leads to conclusion (5.12). □

PROPOSITION 5.3. *Let* $\gamma + r < 1$. *Then the following inequalities hold uniformly in* $\lambda \in \sum^c(A, a)$ *and* $h \leq h_0$:

(i) $|\hat{A}^r R(\lambda, A_B)|_{\mathcal{L}(H)} \leq M/|\lambda - a|^{1-r}$;

(ii) $|\hat{A}^r R(\lambda, A_{hB})|_{\mathcal{L}(H)} \leq M/|\lambda - a|^{1-r}$.

Proof. (i) Since, by virtue of Hypothesis 1a,

$$|\hat{A}^{-\gamma} B F_h \pi_h|_{\mathcal{L}(H)} \leq M \quad \text{uniformly in } h, \tag{5.18}$$

we can apply (5.5) in Proposition 5.1 with $T \equiv BF_h\pi_h$, $\bar{\gamma} = 1 - r \geq \gamma$. This yields

$$|\hat{A}^r [I - R(\lambda, A) B F_h \pi_h] \hat{A}^{-r}|_{\mathcal{L}(H)} \leq M \quad \text{uniformly in } \lambda \in \textstyle\sum^c(A, a). \tag{5.19}$$

On the other hand, by using the perturbation formula

$$\hat{A}^r R(\lambda, A_B) = \hat{A}^r [I - R(\lambda, A) B F_h \pi_h]^{-1} R(\lambda, A) \tag{5.20}$$

and (5.19) with (5.20), we obtain

$$|\hat{A}^r R(\lambda, A_B)|_{\mathcal{L}(H)} \leq C |\hat{A}^r R(\lambda, A)|_{\mathcal{L}(H)} \leq \frac{M}{|\lambda - a|^{1-r}}$$

desired for part (i).

(ii) We start with

$$\hat{A}^r R(\lambda, A_{hB}) = \hat{A}^r (I - R(\lambda A_h) \pi_h B F_h \pi_h)^{-1} R(\lambda, A_h). \tag*{5.21)}$$

From (2.5) for $|\lambda|$ large enough,

$$|R(\lambda, A_h) \pi_h B F_h \pi_h|_{\mathcal{L}(H)} \leq \frac{M}{|\lambda|^{1-\gamma}} < \frac{1}{2}.$$

Hence

$$|[I + R(\lambda, A_h) \pi_h B F_h \pi_h]^{-1}|_{\mathcal{L}(H)} \leq M. \tag{5.22}$$

By the uniform analyticity of A_h and the resolvent equation, the above estimate can be extended to hold for all $\lambda \in \sum^c(A, a)$. Recalling (5.12) and repeating the same arguments as in Step 2 of Proposition 5.1 (where we use $\bar{\gamma} \equiv 1 - r \geq \gamma$) yields

$$|\hat{A}^r [I - R(\lambda, A_h) \pi_h B F_h \pi_h)^{-1} \hat{A}^{-r}|_{\mathcal{L}(H)} \leq M \quad \text{uniformly in } \lambda \in \textstyle\sum^c(A, a). \tag{5.23}$$

Combining (5.21) and (5.23) with (5.11) gives

$$\begin{aligned} |\hat{A}^r R(\lambda, A_{hB})|_{\mathcal{L}(H)} &\leq M |\hat{A}^r (I - R(\lambda, A_h) B F_h)^{-1} \hat{A}^{-r}|_{\mathcal{L}(H)} |\hat{A}^r R(\lambda, A_h)|_{\mathcal{L}(H)} \\ &\leq \frac{M}{|\lambda - a|^{1-r}} \end{aligned}$$

as desired. □

Proof of Lemma 4.1.

Proof of (4.6). We first show that for $|\lambda_0| > R_0$ and R_0 being sufficiently large, we have

$$|A(A - K_h C - \lambda_0 I)^{-1}|_{\mathcal{L}(H)} \leq M. \tag{5.24}$$

Indeed, using the formula

$$A(A - K_hC - \lambda_0 I)^{-1} = -AR(\lambda_0, A)[I + K_hCR(\lambda_0, A)]^{-1}$$

and noting that for R_0 sufficiently large

$$|K_hCR(\lambda_0, A)| = \frac{M}{|\lambda_0|^{1-r}} < \frac{1}{2},$$

we obtain (5.24). The resolvent equation

(5.25)

$$A(A - K_hC)^{-1} = A(A - K_hC - \lambda_0 I)^{-1} - \lambda_0 A(A - K_hC - \lambda_0 I)^{-1}(A - K_hC)^{-1}$$

together with (5.24) and (4.3) gives

(5.26) $$|A(A - K_hC)^{-1}|_{\mathcal{L}(H)} < M \quad \text{uniformly in } h \le h_0.$$

Writing

$$|AR(\lambda, A_C)|_{\mathcal{L}(H)} \le |A(A - K_hC)^{-1}|_{\mathcal{L}(H)} |A_C R(\lambda, A_C)|_{\mathcal{L}(H)}$$

yields the result in (4.6).

Proof of (4.7). For any $\varepsilon > 0$,

(5.27)

$$\hat{A}^r R(\lambda, A_C)\hat{A}^\gamma = \hat{A}^r(-A + K_hC)^{-r-\varepsilon}(-A + K_hC)R(\lambda, A_C)(-A + K_hC)^{-(1-r-\varepsilon)}\hat{A}^\gamma.$$

By using (5.24) together with Lemma 7.3 in [K.2, p. 144], we obtain

(5.28) $$|\hat{A}^r(-A + K_hC)^{(-r-\varepsilon)}|_{\mathcal{L}(H)} \le M.$$

Uniform analyticity of A_C (see (4.3)) yields

(5.29) $$|(A - K_hC)R(\lambda, A_C)|_{\mathcal{L}(H)} \le M \quad \text{uniformly in } \textstyle\sum^c \text{ and } h \le h_0.$$

To estimate the last two terms in (5.27) we apply the result of Proposition 5.1 with $A = A^*$, $\lambda_0 = 0$, $T = -C^*K_h^*$, and $\gamma = r + \varepsilon$. This gives

(5.30) $$|\hat{A}^{*(1-\bar{r})}(-A^* + C^*K_h^*)^{-(1-r-\varepsilon)}|_{\mathcal{L}(H)} \le M$$

for all $\bar{r} > r + \varepsilon$. By duality,

(5.31) $$|(-A + K_hC)^{-(1-r-\varepsilon)}\hat{A}^{(1-\bar{r})}|_{\mathcal{L}(H)} \le M.$$

Since $\gamma + r < 1$, we take ε small so $\gamma + r + \varepsilon < 1$ and we replace $1 - \bar{r}$ by γ to obtain

(5.32) $$|(A - K_hC)^{-(1-r)}\hat{A}^\gamma|_{\mathcal{L}(H)} \le M.$$

Combining (5.27) with (5.28), (5.29), and (5.32) yields the desired result in (4.7).

Proof of 4.8. The proof is the same as that of (4.7) after noting that (4.8) is equivalent to

$$|\hat{A}^{*\gamma}(A^* + F_h^*B^* + \lambda I)^{-1}\hat{A}^{*r}|_{\mathcal{L}(H)} \le M.$$

Since $|F_h^* B^* \hat{A}^{*-\gamma}|_{\mathcal{L}(H)} \leq M$, it suffices to reverse the role of r and γ. □

Proof of Lemma 4.2. (i) Part (i) is a direct consequence of Proposition 5.3.

(ii) For $\lambda \in \sum^c(A, a)$,

$$\begin{aligned} & R(\lambda, A_B) - R(\lambda, A_{hB}) \\ & \quad = (I - R(\lambda, A)BF_h\pi_h)^{-1} \cdot [R(\lambda, A) - R(\lambda, A_h)\pi_h] \\ & \qquad + [(I - R(\lambda, A)BF_h\pi_h)^{-1} - (I - R(\lambda, A_h)\pi_h BF_h\pi_h)^{-1}]R(\lambda, A_h)\pi_h \\ & \quad \equiv I + II, \end{aligned} \tag{5.33}$$

$$|\hat{A}^r I|_{\mathcal{L}(H)} \leq |\hat{A}^r (I - R(\lambda, A)BF_h\pi_h)^{-1}\hat{A}^{-r}|_{\mathcal{L}(H)} \cdot |\hat{A}^r[R(\lambda, A) - R(\lambda, A_h)\pi_h]|_{\mathcal{L}(H)}$$

and by (5.23),

$$\leq M|\hat{A}^r[R(\lambda, A) - R(\lambda, A_h)\pi_h]|_{\mathcal{L}(H)}. \tag{5.34}$$

On the other hand, by Assumptions 3–5 and (2.3) we obtain

$$\begin{aligned} & |\hat{A}^r[R(\lambda, A) - R(\lambda, A_h)\pi_h]|_{\mathcal{L}(H)} \\ & \quad \leq |\hat{A}^r[R(\lambda, A_h)\pi_h - \pi_h R(\lambda, A)]|_{\mathcal{L}(H)} + |\hat{A}^r(I - \pi_h)R(\lambda, A)|_{\mathcal{L}(H)} \\ & \quad \leq Mh^{-sr}h^s + h^{s(1-r)}|AR(\lambda, A)|_{\mathcal{L}(H)} \leq Mh^{s(1-r)}. \end{aligned} \tag{5.35}$$

Equation (5.34) combined with (5.35) gives

$$|\hat{A}^r I|_{\mathcal{L}(H)} \leq Mh^{s(1-r)}. \tag{5.36}$$

As for term II we have

$$\begin{aligned} |\hat{A}^r \mathrm{II}|_{\mathcal{L}(H)} \leq & |\hat{A}^r(I - R(\lambda, A)BF_h\pi_h)^{-1}[R(\lambda, A) - R(\lambda, A_h)\pi_h] \\ & \cdot BF_h\pi_h(I - R(\lambda, A_h)\pi_h BF_h\pi_h)^{-1}R(\lambda, A_h)\pi_h|_H \end{aligned}$$

and by (5.23),

$$\leq M|\hat{A}^r(R(\lambda, A) - R(\lambda, A_h)\pi_h)BF_h\pi_h|_{\mathcal{L}(H)}|R(\lambda, A_{hB})|_{\mathcal{L}(H)}. \tag{5.37}$$

We prove that for $\lambda_0 \in \sum^c(A, a)$,

$$|\hat{A}^r[R(\lambda_0, A) - R(\lambda_0, A_h)\pi_h]BF_h\pi_h|_{\mathcal{L}(H)} \leq Mh^{s(1-r-\gamma)}. \tag{5.38}$$

Indeed,

$$\begin{aligned} |\hat{A}^r[R(\lambda_0, A) - R(\lambda_0, A_h)\pi_h]BF_h\pi_h|_{\mathcal{L}(H)} \leq & |\hat{A}^r[\pi_h R(\lambda_0, A) - R(\lambda_0, A_h)\pi_h]BF_h\pi_h|_{\mathcal{L}(H)} \\ & + |\hat{A}^r(\pi_h - I)R(\lambda_0, A)BF_h\pi_h|_{\mathcal{L}(H)} \end{aligned}$$

and by (ii) of Assumption 3 and (2.6) applied with $Q = 1 - \gamma > r$ and (2.4),

$$\begin{aligned} & \leq Mh^{-sr}|B^*(R(\lambda_0, A^*) - R(\lambda_0, A_h^*))|_{\mathcal{L}(H)} + Mh^{s(1-\gamma-r)}|\hat{A}^{1-\gamma}R(\lambda_0, A)BF_h\pi_h|_{\mathcal{L}(H)} \\ & \leq Mh^{-sr}h^{s(1-\gamma)} \end{aligned}$$

as desired for (5.38).

Now we are in a position to complete the proof of part (ii) of Lemma 4.2. Indeed, from (5.37), (5.38), and the uniform analyticity of $R(\lambda, A_{hB})$ (Proposition 4.1) we obtain

$$|\hat{A}^r \Pi|_{\mathcal{L}(H)} \le M h^{s(1-r-\gamma)},$$

which, combined with (5.36) and (5.37), leads to result (ii) of Lemma 4.2.

(iii) To prove part (iii) we use the resolvent equation together with the compactness of the set in (4.11). Indeed, let λ_0 be such that (4.10) holds. Then for any $\lambda \in \sum^c$ and $\lambda_0 \in \sum^c(A, a)$ we have

$$\begin{aligned}
&|\hat{A}^r[R(\lambda, A_{hB}) - R(\lambda, A_B)]|_{\mathcal{L}(H)} \\
&\quad \le |\hat{A}^r R(\lambda_0, A_{hB}) - R(\lambda_0, A_B)|_{\mathcal{L}(H)} \\
&\qquad + |\lambda - \lambda_0||\hat{A}^r R(\lambda, A_B)|_{\mathcal{L}(H)}|R(\lambda_0, A_{hB}) - R(\lambda_0, A_B)|_{\mathcal{L}(H)} \\
&\qquad + |\lambda - \lambda_0||\hat{A}^r[R(\lambda, A_{hB}) - R(\lambda, A_B)]|_{\mathcal{L}(H)}|(R(\lambda_0, A_{hB})|_{\mathcal{L}(H)}
\end{aligned}$$

and by (4.8), (4.1), and (4.10),

$$(5.39) \quad \le M_0 h^{s(1-r-\gamma)}[1 + |\lambda - \lambda_0|] + M_1|\lambda - \lambda_0||\hat{A}^r[R(\lambda, A_{hB}) - R(\lambda, A_B)]|_{\mathcal{L}(H)},$$

where the constant $M_1 \equiv |R(\lambda_0, A_{hB})|_{\mathcal{L}(H)}$ is uniform in λ_0 (and in λ). Taking λ such that $|\lambda - \lambda_0| \le 1/2M_1$ we obtain $|\hat{A}^r[R(\lambda, A_{hB}) - R(\lambda, A_B)]|_{\mathcal{L}(H)} \le M h^{s(1-r-\gamma)}$.

Repeating the same argument finitely many times leads to the desired result of part (iii). □

6. Examples. We provide several examples illustrating the theory presented. All the examples presented here refer to the situation when $\gamma + r > \frac{1}{2}$. Moreover, Example 6.3 deals with the case when the resolvent operator is *not compact*.

Let Ω be an open bounded domain in R^n with a boundary Γ. We assume that Ω is either smooth or convex.

Example 6.1. Heat equation with Dirichlet boundary control. In Ω, we consider the Dirichlet mixed problem for the heat equation in the unknown $x(t, \xi)$,

$$(6.1) \qquad \begin{cases} x_t = (\Delta + c^2)x & \text{in } (0, \infty) \times \Omega \equiv Q, \\ x(0, \cdot) = x_0 & \text{in } \Omega, \\ x|_{\Sigma} = u & \text{in } (0, \infty) \times \Gamma \equiv \sum, \end{cases}$$

with a boundary control $u \in L_2(\sum)$ and the initial data $x_0 \in L_2(\Omega)$. The solution $x(t)$ is "observed" via a finite-dimensional observation operator given by

$$(6.2) \qquad y(t, \xi) = \int_{\Omega} x(s) w(s) ds\, g(\xi), \qquad \xi \in \Gamma,$$

where $g \in L_2(\Gamma)$ and $w \in H^{-\alpha}(\Omega)$; $0 \le \alpha < \frac{1}{2}$. Here w satisfies

$$(6.3) \qquad \int_{\Omega} w\phi_j \ne 0,$$

where ϕ_j are the eigenvectors corresponding to unstable eigenvalues of $(\Delta + c^2)$, λ_j; $j = 1 \cdots N$, which (without loss of generality) are assumed to have simple multiplicity. To put problem (6.1), (6.2) into the abstract setting of the preceding sections, we introduce the following operators and spaces:

$$(6.4) \qquad Ah \equiv \Delta h + c^2 h, \qquad \mathcal{D}(A) = H^2(\Omega) \cap H_0^1(\Omega);$$

$$H = L_2(\Omega), \quad U = L_2(\Gamma), \quad Z = L_2(\Gamma); \tag{6.5}$$

$$Bu = -A\,Du; \tag{6.6}$$

where D (Dirichlet map) is defined by $h = Dg$ if and only if $(\Delta + c^2)h = 0$ in Ω; $h|_\Gamma = g$. By elliptic regularity and the identification of fractional powers of elliptic operators (see [F.1], [G.3]) we obtain

$$D : \text{continuous } L_2(\Gamma) \to H^{1/2}(\Omega) \subset \mathcal{D}(\hat{A}^{1/4-\varepsilon}); \qquad \varepsilon > 0. \tag{6.7}$$

From (6.7) and (6.6) it follows that Hypothesis 1a is satisfied with $\gamma = \frac{3}{4} + \varepsilon$, where ε is arbitrarily small,

$$\begin{aligned} Cx &\equiv \int_\Omega x(s)w(s)ds\, g(\xi), \\ C &\in \mathcal{L}(H^\alpha(\Omega), L_2(\Gamma)). \end{aligned} \tag{6.8}$$

Since $H^\alpha(\Omega) \subset \mathcal{D}(\hat{A}^{\alpha/2})$ (see [F.1] and [G.3]), for $\alpha < \frac{1}{2}$, Hypothesis 1b holds with $r = \frac{1}{4} - \rho$, where ρ is any positive constant. Hence $r + \gamma < 1$. To verify Hypothesis 2 we recall some stabilizability results proved in [T.2] for the heat equation with boundary controls. In fact, in [T.2] it was shown that the pair (A, B) (A given by (6.4) and B given by (6.6)) is stabilizable with boundary feedback $F \in \mathcal{L}(L_2(\Omega);\ L_2(\Gamma))$ given by

$$(Fx)(\xi) = \int_\Omega xz\, d\Omega\, g_1(\xi), \qquad \xi \in \Gamma, \tag{6.9}$$

where $z \in L_2(\Gamma)$; $g_1 \in L_2(\Gamma)$ and z satisfies condition (6.3) (with w replaced by z). Hence there exists $w_F > 0$ such that

$$|e^{(A+BF)t}|_{\mathcal{L}(L_2(\Omega))} \leq Ce^{-w_F t}. \tag{6.10}$$

As to the detectability condition, note that $C^*v(\xi) = \int_\Gamma vg\, d\Gamma\, w(\xi)$, $\xi \in \Omega$. Thus selecting the operator $K \in \mathcal{L}(L_2(\Gamma);\ L_2(\Omega))$ as $(Ku)(\xi) = \int_\Gamma u \cdot g\, d\Gamma\, z(\xi)$; $\xi \in \Omega$ for a suitable vector $z \in L_2(\Omega)$, we obtain

$$(C^*K^*x)(\xi) = |g|^2_{L_2(\Omega)} \int_\Omega xz\, d\Omega\, w(\xi). \tag{6.11}$$

Now the theory of [T.2] applies again to assert that under condition (6.3) there exists a vector $z \in L_2(\Omega)$ such that

$$|e^{(A^*-C^*K^*)t}|_{\mathcal{L}(L_2(\Omega))} \leq Ce^{-w_K t} \tag{6.12}$$

for some $w_K > 0$. Taking $w_1 = \min(w_K, w_F)$ yields the desired conclusion in Hypothesis 2. Thus we have verified all the continuous hypothesis. As to the approximation framework, we start by introducing the following spaces and operators:

$$V_h \subset H_0^1(\Omega) \tag{6.13}$$

a space of splines (linear, cubic, etc.) that comply with the usual approximation properties

$$|\pi_h y - y|_{H^l(\Omega)} \leq Ch^{s-l}|y|_{H^s(\Omega)}, \quad s \leq 2, \quad s - l \geq 0, \quad 0 \leq l \leq 1; \tag{6.14}$$

$$|y_h|_{H^\alpha(\Omega)} \leq Ch^{-\alpha}|y_h|_{L_2(\Omega)}, \qquad 0 \leq \alpha < 1; \tag{6.15}$$

$$(6.16) \qquad \text{(i)}\ |D^l(y-\pi_h y)|_{L_2(\Gamma)} \le Ch^{s-l-1/2}|y|_{H^s(\Omega)}, \quad \frac{3}{2} < s \le 2, \quad l=0,\,1;$$
$$\text{(ii)}\ |D^l y_h|_{L_2(\Gamma)} \le Ch^{-l-1/2}|y_h|_{L_2(\Omega)}, \qquad l=0,\,1;$$

where D^l stands for the differential operator of order l. Standard formulation of approximation properties (6.14), (6.16) (see [T.1]) involves an interpolation operator rather than an orthogonal projection. However, since inverse approximation properties (6.15), (6.16(ii)) are assumed (always satisfied on a quasi-uniform mesh), we can easily show that these properties hold with the orthogonal projection operator as well. Also, in the case of property (6.16(i)), the usual formulation requires that $s \ge 2$. However, it was pointed out by Peterson [P.2] that it suffices to assume $s > \frac{3}{2}$.

(6.17) $\qquad A_h : V_h \to V_h$ is defined as the usual Galerkin approximation,

$$(A_h, x_h, y_h)_\Omega \equiv \int_\Omega A_h x_h y_h \, d\Omega \equiv -\int_\Omega [\nabla x_h \nabla y_h - c^2 x_h y_h]\, d\Omega \quad \text{for all } x_h, y_h \in V_h.$$

It is well known (see [T.1], [B-S-T-W], and [L.1]) that the condition of uniform analyticity, i.e., Assumption 1, is satisfied. Also, classical results on approximation of elliptic problems (see [T.1]) imply that Assumption 2 holds with $s = 2$. Since

$$(6.18) \qquad B^* x = -\frac{\partial}{\partial \nu} x|_\Gamma,$$

Assumptions 3(i) and 4(i) are direct consequences of (6.16) applied with $l = 1$. Assumption 5(i) follows at once from the trace theory. Indeed by (6.16) and (6.18),

$$|B^*\pi_h x|_{L_2(\Gamma)} = \left|\frac{\partial}{\partial\nu}\pi_h x\right|_{L_2(\Gamma)} \le \left|\frac{\partial}{\partial\nu}(\pi_h - I)x\right|_{L_2(\Gamma)} + \left|\frac{\partial}{\partial\nu}x\right|_{L_2(\Gamma)}$$
$$\le ch^\varepsilon |y|_{H^{3/2+\varepsilon}(\Omega)} + C|y|_{H^{3/2+\varepsilon}(\Omega)} \le C|y|_{\mathcal{D}(\hat{A}^{3/4+\varepsilon})}.$$

As for Assumptions 3(ii), 4(ii), and 5(ii), it is enough to note that

$$|\hat{A}^r x|_{L_2(\Omega)} \simeq |x|_{H^{1/2-2\rho}(\Omega)}, \qquad \rho > 0$$

and validity of these assumptions follows directly from (6.14) and (6.15). Finally, to comply with Assumptions 6 and 7, since $B^*R(\lambda_0, A)$ and $CR(\lambda_0, A)$ are compact, it suffices to take

$$F_h \equiv F\pi_h \quad \text{and} \quad K_h \equiv \pi_h K.$$

We conclude that all the assumptions of Theorems 1.1 and 3.1 are satisfied.

Remark 6.1. Note that actual design of a finite-dimensional compensator for system 6.1 *does not* require any knowledge of the eigenfunctions/eigenvalues of A. Indeed, we can construct stabilizing feedbacks F_h and K_h from the appropriate Riccati equations (3.6) and (3.11) that involve only A_h, B_h, and C_h.

Example 6.2. Heat equation with Neumann boundary control and boundary observation. In Ω we consider

$$(6.19) \qquad \begin{cases} x_t = (\Delta + c^2)x & \text{in } Q, \\ x(0) = x_0 & \text{in } \Omega, \\ \left.\dfrac{\partial x}{\partial \nu}\right|_\Gamma = u & \text{in } \Sigma. \end{cases}$$

We observe on the boundary

$$y = x|_\Gamma. \tag{6.20}$$

To put problem (6.19), (6.20) into an abstract framework we set

$$Ah \equiv \Delta h + c^2 h, \qquad \mathcal{D}(A) = \left\{ h \in H^2(\Omega);\ \frac{\partial}{\partial\nu} h|_\Gamma = 0 \right\}; \tag{6.21}$$

$$H = L_2(\Omega), \quad U = L_2(\Gamma), \quad Z = L_2(\Gamma); \tag{6.22}$$

$$Bu = -ANu; \tag{6.23}$$

where N (Neumann map) is defined by

$$h = Ng \quad \text{iff}\, (\Delta + c^2)h = 0 \quad \text{in}\, \Omega, \qquad \frac{\partial}{\partial\nu} h|_\Gamma = g.$$

By the elliptic regularity and [F.1], [G.3],

$$N : \text{continuous}\, L_2(\Gamma) \to H^{3/2-\varepsilon}(\Omega) \subset \mathcal{D}(\hat{A}^{3/4-\varepsilon}), \tag{6.24}$$

$$Cx \equiv x|_\Gamma. \tag{6.25}$$

From (6.23) and (6.24) it follows that Hypothesis 1a is satisfied with $\gamma = \frac{1}{4} + \varepsilon$. Equation (6.25) together with trace theory and the inclusion $\mathcal{D}(\hat{A}^{1/4+\rho}) \subset H^{1/2+2\rho}(\Omega)$ (see [F.1], [G.3]) imply that Hypothesis 1b holds with $r = \frac{1}{4} + \rho$, where ρ is arbitrarily small. Hence $\gamma + r < 1$, but $\gamma + r > \frac{1}{2}$. As to the stabilizability condition (Hypothesis 2) we recall that in [T.2] it was shown that the same feedback F as in (6.9) (with appropriate choices of vectors g_1 and z) uniformly stabilizes A, i.e., there exist constants C and w_F such that

$$|e^{(A+BF)t}|_{\mathcal{L}(L_2(\Omega))} \le Ce^{-w_F t}, \qquad w_F > 0. \tag{6.26}$$

Detectability condition in Hypothesis 2 amounts to the same thing, since

$$C^* u = ANu \quad \text{and}$$
$$(A^* - C^*K^*)x = x_t \quad \text{is equivalent to}\, x_t = (\Delta + c^2)x; \quad \frac{\partial}{\partial\nu} x = -K^* x.$$

Thus, it is enough to select $K^* = F$ to obtain

$$|e^{(A^* - C^*K^*)t}|_{\mathcal{L}(L_2(\Omega))} \le Ce^{-w_F t} \tag{6.27}$$

which, in turn, implies Hypothesis 2. Approximation framework is very similar to the Dirichlet case considered in Example 6.1. Indeed, the space $V_h \subset H^1(\Omega)$ is defined as a space of splines complying with properties (6.14)–(6.16). The approximating generator $A_h : V_h \to V_h$ is defined by the same formula (6.17), and Assumptions 1 and 2 hold with $s = 2$. Since

$$B^* x = x|_\Gamma, \tag{6.28}$$

Assumptions 3(i) and 4(i) are direct consequences of (6.16) applied with $l = 0$. As for Assumption 5(i),

$$|B^* \pi_h x|_{L_2(\Gamma)} \le |\pi_h x|_{L_2(\Gamma)} \le C|\pi_h x|_{H^{1/2+2\varepsilon}(\Omega)} \le C|x|_{H^{1/2+2\varepsilon}(\Omega)} \le C|x|_{\mathcal{D}(\hat{A}^{1/4+\varepsilon})},$$

where we have used (6.14) and trace theory. Since

$$|\hat{A}^r x|_{L_2(\Omega)} \simeq |x|_{H^{2r}(\Omega)}$$

(see [F.1], [G.3]), the validity of Assumption 3(ii), 4(ii), and 5(ii) follows from (6.14) and (6.15). Assumptions 6 and 7 hold (in view of compactness of $R(\lambda, A)$) with F_h and K_h taken as the projections of F and K, respectively, i.e., $F_h \equiv F\pi_h$, $K_h \equiv \pi_h K$.

We conclude that all the hypotheses of Theorems 1.1 and 3.1 are satisfied.

Example 6.3. Structurally damped plate equation with point control and boundary observation. Consider the following model of a plate equation in the deflection $w(t,\xi)$:

$$(6.29)\qquad \begin{cases} w_{tt} + \Delta^2 w + \Delta^2 w_t = \sum_{i=1}^{3} \delta(\xi - \xi^i) u_i(t) & \text{in } Q, \\ w(0,\cdot) = w_o;\ w_t(0,\cdot) = w_1 & \text{in } \Omega, \\ \Delta w|_\Sigma + (1-\mu)B_1 w = 0 & \text{in } \sum, \\ \dfrac{\partial}{\partial \nu}\Delta w|_\Sigma + (1-\mu)B_2 w \equiv 0 & \text{on } \sum, \end{cases}$$

with load control concentrated at the interior points ξ^i of an open bounded domain Ω of R^2. Here $0 < \mu < \frac{1}{2}$ is the Poisson modulus and the boundary operators B_1 and B_2 are

$$B_1 w \equiv 2\nu_1\nu_2 w_{xy} - \nu_1^2 w_{yy} - \nu_2^2 u_{xx},$$

$$B_2 w = \frac{\partial}{\partial \tau}[(\nu_1^2 - \nu_2^2) w_{xy} + \nu_1\nu_2(w_{yy} - w_{xx})],$$

where $\partial/\partial\tau$ stands for the tangential derivative. The observation operator is given by

$$(6.30)\qquad y = \frac{\partial}{\partial \nu} w_t|_\Sigma.$$

To put problem (6.29), (6.30) into an abstract setting we introduce

$$(6.31)\qquad \begin{aligned} \mathcal{A}h &= \Delta^2 h, \\ \mathcal{D}(\mathcal{A}) &= \left\{ h \in H^4(\Omega); \Delta h + (1-\mu)B_1 h = 0; \frac{\partial}{\partial\nu}\Delta h + (1-\mu)B_2 h = 0 \text{ on } \Gamma \right\}; \end{aligned}$$

$$(6.32)\qquad H = \mathcal{D}(\mathcal{A}^{1/2}) \times L_2(\Omega) = H^2(\Omega) \times L_2(\Omega), \quad U = R^3, \quad Z = L_2(\Gamma);$$

$$(6.33)\qquad A = \begin{bmatrix} 0 & I \\ -\mathcal{A} & -\mathcal{A} \end{bmatrix};$$

$$(6.34)\qquad Bu = \begin{bmatrix} 0 \\ \sum_{i=1}^{3} \delta(\xi - \xi^i) u_i \end{bmatrix};$$

$$(6.35)\qquad Cx = C\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \frac{\partial}{\partial\nu} x_2|_\Gamma.$$

It is known (see [L-T.1]) that A generates an analytic semigroup on H and, moreover, computations in [L-T.1, p. 536] (see also [Ch-T.2]) show that Hypothesis 1a is satisfied with $\gamma = \frac{1}{4} + \varepsilon$. As to Hypothesis 1b, we note that with $x = (x_1, x_2) \subset H$,

$$\hat{A}^{-1/2}x = \frac{\partial}{\partial\nu}[-\mathcal{A}^{1/4}[2I + \mathcal{A}^{1/2}]^{-1/2}x_1 + \mathcal{A}^{-1/4}[2I + \mathcal{A}^{1/2}]^{-1/2}x_2],$$

and by trace theory and the equivalence $\mathcal{D}(\mathcal{A}^Q) = H^{4Q}$, $0 \leq Q < \frac{1}{2}$,

$$|C\hat{A}^{-1/2}x|_Z \leq M|x|_H. \tag{6.36}$$

Hence, the value of r in Hypothesis 1b is equal to $\frac{1}{2}$ and $\gamma + r = \frac{1}{4} + \varepsilon + \frac{1}{2} < 1$. As for the stabilizability condition (Hypothesis 2), note that $\lambda = 0$ is the only unstable eigenvalue of $\mathcal{A}$ (hence of A) with the multiplicity equal to three. In fact, functions 1, ξ_1, ξ_2 are three linearly independent eigenfunctions (corresponding to the zero eigenvalue of $\mathcal{A}$). Thus, the results of [T.2] apply, provided that the points ξ^i, $i = 1 - 3$ satisfy the following rank condition:

$$\det \begin{vmatrix} 1 & \xi_1^1 & \xi_2^1 \\ 1 & \xi_1^2 & \xi_2^2 \\ 1 & \xi_2^3 & \xi_2^3 \end{vmatrix} \neq 0. \tag{6.37}$$

If (6.37) holds, then the feedback operator $F : H \to R^3$ given by

$$Fx = \int_\Omega x_1 \vec{z}\, d\Omega \quad \text{with } \vec{z} \in (L_2(\Omega))^3, \tag{6.38}$$

suitably selected vector, uniformly stabilizes the dynamics in (6.29) with $u = Fx$. This is to say that

$$|e^{(A+BF)t}|_{\mathcal{L}(H)} \leq Ce^{-w_F t}; \qquad w_F > 0.$$

To establish the detectability condition (Hypothesis 2), it suffices to note that the problem of finding $K \in \mathcal{L}(Z;\, H)$ such that $|e^{(A-KC)t}|_{\mathcal{L}(H)} \leq Ce^{-w_K t}$, $w_K > 0$ is equivalent, in our case, to finding $K^* \in \mathcal{L}(H;\, L_2(\Gamma))$ such that the system

$$\begin{cases} w_{tt} + \Delta^2 w + \Delta^2 w_t = 0; \\ w(0) = w_0, \qquad w_t(0) = w_1; \\ \Delta w + (1-\mu)B_1 w = K^*(w, w_t) \quad \text{on } \sum; \\ \dfrac{\partial}{\partial\nu}\Delta w + (1-\mu)B_2 w = 0 \quad \text{on } \sum \end{cases} \tag{6.39}$$

is exponentially stable on $\mathcal{D}(\mathcal{A}^{1/2}) \times L_2(\Omega)$. Problem (6.39) is a classical damped plate problem with boundary feedback and with one unstable pole ($\lambda = 0$). Thus the results of [T.2] apply again. Indeed, we can easily show that the feedback .

$$K^*(w) = \sum_{i=1}^{3} \int_\Omega w z_i \, dz\, g_i,$$

where $z_i \in L_2(\Omega)$ and $g_i \in L_2(\Gamma)$ are suitably selected (to guarantee the appropriate rank condition to be satisfied) exponentially stabilizes (6.39).

For the approximation part we introduce the following spaces and operators. $V_h \equiv \mathcal{V}_h \times \mathcal{V}_h$, where $\mathcal{V}_h \subset H^2(\Omega)$ is a space of splines (of order r_0) complying with the following properties:

$$|Q_h x - x|_{H^l(\Omega)} \leq Ch^{s-l}|x|_{H^s(\Omega)}, \quad 0 \leq l \leq 2, \quad l \leq s \leq r_0; \tag{6.40}$$

$$|x_h|_{H^\alpha(\Omega)} \leq Ch^{-s}|x_h|_{H^{\alpha-s}(\Omega)}, \qquad 0 \leq \alpha < 2; \tag{6.41}$$

where Q_h is the orthogonal projection of $L_2(\Omega)$ onto $\mathcal{V}_h$.

$$\begin{aligned} &\mathcal{A}_h : \mathcal{V}_h \to \mathcal{V}_h \quad \text{is given by} \\ &\mathcal{A}_h = Q_h \mathcal{A} Q_h, \quad \text{i.e., } (\mathcal{A}_h x_h, y_h)_{L_2(\Omega)} = (\Delta x_h, \Delta y_h)_{L_2(\Omega)} = (\mathcal{A} x_h, y_h)_{L_2(\Omega)}, \\ &A_h : V_h \to V_h \quad \text{is defined as} \end{aligned} \tag{6.42}$$

$$A_h = \begin{bmatrix} 0 & Q_h \\ -\mathcal{A}_h & -\mathcal{A}_h \end{bmatrix}, \tag{6.43}$$

$$\pi_h \equiv Q_h \times Q_h, \tag{6.44}$$

$$F_h \equiv F\pi_h, \qquad K_h \equiv \pi_h K. \tag{6.45}$$

Validity of the uniform analyticity hypothesis (Assumption 1) follows by applying verbatim the argument of [Ch-T.1] of the continuous case to the finite-dimensional operator A_h (note that in this example the bilinear form associated with A is *not* coercive). Assumption 2 with $s = 2$ is a consequence of

$$\|(\pi_h A^{-1} - A_h^{-1}\pi_h)x\|_H = \|(Q_h A^{-1} - \mathcal{A}_h^{-1} Q_h)x_2\|_{H^2(\Omega)} \leq Ch^2\|x_2\|_{L_2(\Omega)} \leq Ch^2\|x\|_H,$$

where we have used standard approximation properties of the biharmonic operator. To verify part (i) of Assumptions 3–5, we note first that in our case,

$$B^*x = [x_2(\xi^1), x_2(\xi^2);\ x_2(\xi^3)]. \tag{6.46}$$

Hence, by Sobolev imbeddings and (6.41),

$$|B^*x_h|_U \leq \sup_{i=1-3} |x_{2h}(\xi^i)| \leq C|x_{2h}|_{H^{1+\varepsilon}(\Omega)} \leq Ch^{-1-\varepsilon}|x_{2h}|_{L_2(\Omega)} \leq Ch^{-\gamma \cdot s}|x_h|_H$$

as $\gamma s = \frac{1}{2} + 2\varepsilon$, where ε can be taken arbitrarily small. Assumptions 4(i) and 5(i) are verified in a similar manner (see [L-T.1, p. 537]). Parts (ii) of Assumptions 3–5 follow in a direct way from the trace theory combined with approximation properties (6.40) and (6.41). Details, being straightforward, are omitted.

Finally, Assumption 6 is satisfied by the virtue of the compactness of $B^*R(\lambda_0, A^*)$ (note that $R(\lambda_0, A^*)$ is *not* compact). Indeed, from (6.46),

$$|B^*R(\lambda_0, A^*)x|_U \leq |(R(\lambda_0, A^*)x)_2|_{H^{1+\varepsilon}(\Omega)}. \tag{6.47}$$

Inequality (6.42) combined with the fact that $x \to (R(\lambda_0, A^*x))_2$ is bounded from $H \to H^2(\Omega)$, hence compact $H \to H^{1+\varepsilon}(\Omega)$; $\varepsilon < 1$ yields the desired conclusion. Similarly, validity of Assumption 7 follows from (6.45) and the compactness of $CR(\lambda_0, A)$. Indeed, note that $CR(\lambda_0, A) : H \to H^{1/2}(\Gamma)$ is bounded, hence compact $H \to L_2(\Gamma)$.

We conclude that all the hypotheses of Theorems 1.1 and 3.1 are satisfied.

REFERENCES

[B-S-T-W] J. H. Bramble, A. H. Schatz, V. Thomee, and L. B. Wahlbin, *Some convergence estimates for semidiscrete Galerkin-type approximations for parabolic equations*, SIAM J. Numer. Anal., 14 (1977), pp. 218–241.

[B-D-D-M] A. Bensoussan, G. Da Prato, M. Delfour, and S. Mitter, *Representation and Control of Infinite Dimensional Systems*, Birkhäuser, Basel, 1993.

[Ch-T.1] S. Chen and R. Triggiani, *Proof of extensions of two conjectures on structural damping for elastic systems*, Pacific J. Math., 136 (1989), pp. 15–55.

[Ch-T.2] ———, *Characterization of domains of fractional powers of certain operators arising in elastic systems, and applications*, J. Differential Equations, 88 (1990), pp. 279–293.

[C.1] R. Curtain, *Finite dimensional compensators for parabolic distributed systems with unbounded control and observation*, SIAM J. Control, 22 (1984), pp. 255–277.

[D-I] G. Da Prato and A. Ichikawa, *Riccati equations with unbounded coefficients*, Ann. Mat. Pura Appl., 140 (1985), pp. 209–221.

[F.1] D. Fujiwara, *Concrete characterizations of the domains of fractional powers of same elliptic differential operators of the second order*, Proc. Japan Acad., 48 (1967), pp. 82–86.

[F.2] F. Flandoli, *Algebraic Riccati equations arising in boundary control problems*, SIAM J. Control Optim., 25 (1987), pp. 612–636.

[F.3] ———, *Riccati equation arising in a boundary control problem with distributed parameters*, SIAM J. Control Optim., 22 (1984), pp. 76–86.

[G.1] J. S. Gibson, *An analysis of optimal model regulation: Convergence and stability*, SIAM J. Control Optim., 19 (1981), pp. 686–707.

[G.2] ———, Approximation theory for linear quadratic Gaussian control of flexible structures, SIAM J. Control Optim. 29 (1990), 1–38.

[G.3] P. Grisvard, *Characterization de quelques espaces d'interpolation*, Arch. Rational Mech. Anal., 25 (1967), pp. 40–63.

[K.1] T. Kato, *Perturbations Theory for Linear Operators*, Springer-Verlag, New York, Berlin, 1966.

[K.2] S. G. Krejin, *Linear Differential Equations in Banach Space, American Mathematical Society*, Providence, RI, 1971.

[L.1] I. Lasiecka, *Convergence estimates for semidiscrete approximations of nonselfadjoint parabolic equations*, SIAM J. Numer. Anal., 21 (1984), pp. 894–909.

[L.2] I. Lasiecka, *Galerkin approximations of infinite dimensional compensators for flexible structures with unbounded control action*, Acta Appl. Math., 28 (1992), pp. 101–133.

[L.3] D. G. Luenberger, *An introduction to observers*, IEEE Trans. Automat. Control, AC-16 (1971), pp. 596–602.

[L-T] I. Lasiecka and R. Triggiani, *Algebraic Riccati Equations with Applications to Boundary/Point Control Problems: Continuous Theory and Approximation Theory*, Springer-Verlag, L N C I S (1991).

[L-T.1] ———, *Numerical approximations of algebraic Riccati equations for abstract systems modelled by analytic semigroups, and applications*, Mathematics of Computations Vol. 57 (1991), pp. 639–662 and 513–537.

[P.1] A. Pazy, *Semigroups of Operators and Applications to Partial Differential Equations*, Springer-Verlag, 1983.

[P.2] T. Peterson, Private communication.

[S.1] J. M. Schumacher, *A direct approach to compensator design for distributed parameter systems*, SIAM J. Control, 21 (1983), pp. 823–837.

[T.1] V. Thomée, *Galerkin Finite Element Methods for Parabolic Problems*, Lecture Notes in Math., Vol. 1054, Springer, Berlin, 1984.

[T.2] R. Triggiani, *Boundary feedback stabilization of parabolic equations*, Appl. Math. Optim., 6 (1980), pp. 201–220.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 33, No. 1, pp. 89–106, January 1995

© 1995 Society for Industrial and Applied Mathematics
004

WEIGHTED ESTIMATION AND TRACKING FOR ARMAX MODELS [*]

B. BERCU[†]

Abstract. For complex multivariate ARMAX models, the author studies the weighted least squares algorithm which offers, by the choice of suitable weightings, the advantages of both the extended least squares and the stochastic gradient algorithms. Concerning adaptive tracking problems, the strong consistency of the estimator and control optimality are both ensured. Almost sure rates of convergence are also provided.

Résumé. Pour les modèles ARMAX vectoriels complexes, on étudie l'algorithme des moindres carrés pondérés qui conjugue, par le choix de pondérations convenables, à la fois les avantages des algorithmes des moindres carrés généralisés et du gradient. Concernant les problèmes de poursuite adaptative, on assure la consistance de l'estimateur et l'optimalité du contrôle. On précise également les vitesses de convergence presque sûre.

Key words. linear systems, estimation, strong consistency, adaptive tracking

AMS subject classifications. 60G42, 93E12, 62M20, 93E20

1. Introduction. In the study of recursive identification and adaptive tracking for ARMAX linear systems, the major goal is to find a stochastic algorithm that ensures both strong consistency of the estimator and control optimality. On one hand, if we focus our attention on the strong consistency, we choose the extended least squares (ELS) algorithm [15], [19], [22], [23], [24]. On the other hand, if we are interested in adaptive tracking, we should use the stochastic gradient (SG) algorithm [8], [18]. Therefore, a natural question is: Can we find a stochastic algorithm that combines both advantages of the ELS for strong consistency and of the SG for adaptive tracking? A positive answer was recently given by Bercu and Duflo [4] when they proposed a new weighted least squares (WLS) algorithm. In this paper we complete their work, giving a solution to the twenty-year-old adaptive tracking problem proposed by Aström and Wittenmark [1] for ARMAX models.

The paper is organized as follows. In §2, we describe the WLS algorithm. The main difference from the ELS algorithm is the introduction of a random weighting sequence $a = (a_n)$. Section 3 is devoted to the crucial choice of $a = (a_n)$. The main results of the paper are given in §4. We can see that the WLS algorithm equals the performance of the ELS for the strong consistency and matches the best result of the SG for the adaptive tracking. More precisely, the relation (24) is similar to the one obtained by Lai and Wei [24], [25] or Chen and Guo [12] for the ELS estimator. Moreover, concerning the prediction errors sequence, the relation (26) is exactly the same as the one proved by Goodwin, Ramadge, and Caines [18] for the SG algorithm. Finally, in §§5 and 6, we solve, in a simple way, the adaptive tracking problem. We prove both strong consistency of the WLS estimator and control optimality. We also provide almost sure rates of convergence. Section 7 is devoted to a survey on earlier related works on adaptive tracking. Comparing our work with previous similar results, we show how the WLS algorithm is well suited for adaptive tracking problems. A short conclusion is given in §8. All technical proofs are collected in the Appendices.

Notations. In the following sections, for any matrix A, ${}^tA$ denotes the transpose of A, ${}^*A$ represents the Hermitian adjoint of A and we set $\|A\|^2 = \text{tr}(A^*A)$. Moreover, if A is a square matrix, $\text{tr}(A)$ denotes the trace of A, $\det(A)$ the determinant of A, and $\lambda_{\min}(A)$, $\lambda_{\max}(A)$ the minimum and the maximum eigenvalues of A, respectively. In addition, if A and B are two positive definite Hermitian matrices, then $A \leq B$ if $B - A$ is positive definite. Finally, for any positive integer d, I_d is the identity matrix of order d.

[*] Received by the editors August 31, 1992; accepted for publication (in revised form) August 5, 1993.

[†] Laboratoire de Statistiques, Batiment 425 Mathématiques, Université de Paris-Sud, 91 405 Orsay Cedex, France.

2. Weighted estimation. Let $(\Omega, \mathcal{A}, P)$ be a probability space with a filtration $\mathbf{F} = (\mathcal{F}_n)_{n\geq 0}$, where $\mathcal{F}_n$ is the σ-algebra generated by events occurring up to time n. We consider the following complex multivariate ARMAX model of order (p, q, r):

$$A(R)Y_n = B(R)U_n + C(R)\varepsilon_n, \tag{1}$$

where Y, U, and ε are the d_1-dimensional output, d_2-dimensional input, and d_1-dimensional driven noise, respectively. Set for the shift-back operator R,

$$A(R) = I_{d_1} - A_1 R - \cdots - A_p R^p, \tag{2}$$

$$B(R) = B_1 R + \cdots + B_q R^q, \tag{3}$$

$$C(R) = I_{d_1} + C_1 R + \cdots + C_r R^r, \tag{4}$$

where A_i, B_j, C_k are unknown matrices. Assume that the control $U = (U_n)$ and the noise $\varepsilon = (\varepsilon_n)$ are adapted to **F** and that ε is a martingale difference sequence with

$$\sup_{n\geq 0} E[\|\varepsilon_{n+1}\|^2 \mid \mathcal{F}_n] < \sigma^2 \quad \text{a.s.}, \tag{5}$$

where σ^2 is deterministic. The initial state ${}^t\Psi_0 = ({}^tY_0^p, {}^tU_0^q, {}^t\varepsilon_0^r)$ is $\mathcal{F}_0$ measurable and, for $n \geq 0$,

$${}^t\Psi_n = ({}^tY_n^p, {}^tU_n^q, {}^t\varepsilon_n^r), \tag{6}$$

where ${}^tY_n^p = ({}^tY_n, \ldots, {}^tY_{n-p+1})$, ${}^tU_n^q = ({}^tU_n, \ldots, {}^tU_{n-q+1})$ and ${}^t\varepsilon_n^r = ({}^t\varepsilon_n, \ldots, {}^t\varepsilon_{n-r+1})$. Let $\hat{\theta}_n$ be an estimator of θ where

$${}^*\theta = (A_1, \ldots, A_p, B_1, \ldots, B_q, C_1, \ldots, C_r). \tag{7}$$

The noise ε is predicted by the a posteriori error $\hat{\varepsilon}$ with $\hat{\varepsilon}_0^r = 0$ and, for $n \geq 0$,

$$\hat{\varepsilon}_{n+1} = Y_{n+1} - {}^*\hat{\theta}_{n+1}\Phi_n, \tag{8}$$

where

$${}^t\Phi_n = ({}^tY_n^p, {}^tU_n^q, {}^t\hat{\varepsilon}_n^r). \tag{9}$$

Let $a = (a_n)$ be a sequence of random variables adapted to **F**, positive, nonincreasing, and $\leq 1. a = (a_n)$ is called a weighting sequence. We propose, in order to estimate θ, the WLS estimator $\hat{\theta}_n$ introduced by Bercu and Duflo [4] and given, for $n \geq 0$, by

$$\hat{\theta}_{n+1} = \hat{\theta}_n + a_n S_n^{-1}(a)\Phi_n{}^*(Y_{n+1} - {}^*\hat{\theta}_n\Phi_n), \tag{10}$$

where the initial value $\hat{\theta}_0$ is arbitrarily chosen and, for $n \geq 0$,

$$S_n(a) = \sum_{k=0}^{n} a_k \Phi_k{}^*\Phi_k + S \tag{11}$$

with any positive definite Hermitian and deterministic matrix S. We also write, for such a matrix Q,

$$Q_n(a) = \sum_{k=0}^{n} a_k \Psi_k{}^*\Psi_k + Q. \tag{12}$$

The inverses of the matrices $S_n(a)$ and $Q_n(a)$ are recursively generated by the matrix inversion formula of Riccati. Denote by $s_n(a)$ and $q_n(a)$ the traces of $S_n(a)$ and $Q_n(a)$, respectively. We also make use of

$$s_n = \sum_{k=0}^{n} \|\Phi_k\|^2 + s, \tag{13}$$

where $s = \text{tr}(S)$, and of the prediction errors sequence $\pi = (\pi_n)$ defined by

$$\pi_n = {}^*\theta\Psi_n - {}^*\hat{\theta}_n\Phi_n. \tag{14}$$

3. Admissibility. Consider a weighting sequence $a = (a_n)$ and set, for $n \geq 0$,

$$f_n(a) = a_n\, {}^*\Phi_n S_n^{-1}(a)\Phi_n, \qquad \Delta = \sum_{n=0}^{\infty} a_n f_n(a). \tag{15}$$

a is said to be admissible if Δ is integrable. Let F be the family of continuous and nonincreasing functions f from $\mathbf{R}^+$ to $\mathbf{R}^+$ such that $xf(x)$ converges to 0 as x goes to infinity and

$$\int_c^{+\infty} f(x)dx < +\infty \tag{16}$$

for any constant $c > 0$. We have the classical inequality

$$f_n(a) \leq \inf\{1, \log(\det S_n(a)) - \log(\det S_{n-1}(a))\}. \tag{17}$$

Using (17), Bercu and Duflo [4] have shown that a weighting sequence $a = (a_n)$ such that $a_n = f(n)$ or $a_n = f(\log s_n)$ with $f \in F$ is admissible. More precisely, they have proved that Δ is always almost surely bounded.

Throughout the following, we always assume that the weighting sequence a is admissible.

Remark. It is important to see that the WLS algorithm does not include the ELS as a special case since the weighting sequence with general constant term equal to 1 is not admissible.

4. Strong consistency. We make use of the following traditional assumption of passivity:

(A_1) $C^{-1} - \frac{1}{2}I_{d_1}$ is strictly positive real.

THEOREM 1. *For the model and the* WLS *algorithm* (1) *to* (14), *assume that* (A_1) *is satisfied. Then we have*

$$E\left[\sup_{n\geq 0} \|S_n^{1/2}(a)(\hat{\theta}_{n+1} - \theta)\|^2\right] < +\infty, \tag{18}$$

$$E\left[\sum_{n=0}^{\infty} a_n \|{}^*(\hat{\theta}_{n+1} - \theta)\Phi_n\|^2\right] < +\infty, \tag{19}$$

$$E\left[\sum_{n=0}^{\infty} \lambda_{\min} S_{n-1}(a)\|\hat{\theta}_{n+1} - \hat{\theta}_n\|^2\right] < +\infty, \tag{20}$$

$$E\left[\sum_{n=0}^{\infty} a_n \|\Phi_n - \Psi_n\|^2\right] < +\infty, \tag{21}$$

$$E\left[\sum_{n=0}^{\infty} a_n(1 - f_n(a))\|\pi_n\|^2\right] < +\infty. \tag{22}$$

Proof. The proof is given in Appendix A.

COROLLARY 1. If (A_1) *holds, the* WLS *estimator given by* (10) *is strongly consistent on*

$$I = \{\lim_{n\to+\infty} \lambda_{\min} S_n(a) = +\infty\} \tag{23}$$

and, on I, we have

$$\|\hat{\theta}_{n+1} - \theta\|^2 = O(\{\lambda_{\min} S_n(a)\}^{-1}) \quad a.s. \tag{24}$$

Remark. As $S_n(a) \geq S$, the WLS estimator is always almost surely bounded. By (24), we can conclude that the WLS algorithm behaves as well as the ELS for ARMAX parameter estimation [12], [24], [25].

COROLLARY 2. *If* (A_1) *holds, the prediction errors sequence satisfies*

$$E\left[\sum_{n=0}^{\infty} (a_n^{-1} + \|\Phi_n\|^2)^{-1}\|\pi_n\|^2\right] < +\infty. \tag{25}$$

More particularly, if, for $n \geq 0$, $s_n^{-1} \leq ca_n$ with c deterministic > 0, then

$$E\left[\sum_{n=0}^{\infty} \frac{\|\pi_n\|^2}{s_n}\right] < +\infty. \tag{26}$$

Finally, we also have

$$E\left[\sum_{n=0}^{\infty} \left(a_n^{-1} + \frac{\|\Phi_n\|^2}{\lambda_{\min} S_{n-1}(a)}\right)^{-1} \|\pi_n\|^2\right] < +\infty. \tag{27}$$

Remark. By (26), we can conclude that the WLS algorithm behaves as well as the SG for adaptive tracking [8], [18].

Proof. $\pi_n = {}^*\theta\Psi_n - {}^*\hat{\theta}_n\Phi_n$, so we can rewrite $\pi_n = \tau_{n+1} + {}^*(\hat{\theta}_{n+1} - \hat{\theta}_n)\Phi_n$ where $\tau_{n+1} = {}^*\theta\Psi_n - {}^*\hat{\theta}_{n+1}\Phi_n$. Then, by use of (19) together with (21),

$$E\left[\sum_{n=0}^{\infty} a_n\|\tau_{n+1}\|^2\right] < +\infty. \tag{28}$$

Hence, (25) and (27) immediately follow from (20) and (28). Moreover, if $s_n^{-1} \leq ca_n$ with c deterministic > 0, (20) and (28) imply (26). □

COROLLARY 3. *Assume that the driven noise ε satisfies the strong law of large numbers* (LN) *with, for $n \geq 0$,*

$$E[\varepsilon_{n+1}{}^*\varepsilon_{n+1} \mid \mathcal{F}_n] = \Gamma, \tag{29}$$

where Γ is a deterministic covariance matrix. To estimate Γ, we propose the two empirical estimators

- $\hat{\Gamma}_n = \frac{1}{n}\sum_{k=1}^{n} (Y_k - {}^*\hat{\theta}_{k-1}\Phi_{k-1})\,{}^*(Y_k - {}^*\hat{\theta}_{k-1}\Phi_{k-1})$,
- $\tilde{\Gamma}_n = \frac{1}{n}\sum_{k=1}^{n} \hat{\varepsilon}_k{}^*\hat{\varepsilon}_k$.

Suppose that (A_1) *is satisfied and that $s_n^{-1} = O(a_n)$. Then, on the set $\{s_n = O(n)$ and $s_n \to +\infty\}$, $\hat{\Gamma}_n$ and $\tilde{\Gamma}_n$ areboth strongly consistent estimators of Γ.*

Proof. The proof is obvious using (21) and (26) with Kronecker's lemma. □

5. Adaptive tracking. We still consider the model and the WLS algorithm (1) to (14). The goal of adaptive tracking is to find a control sequence $U = (U_n)$ that forces the output $Y = (Y_n)$ to follow a given reference trajectory $y = (y_n)$. We first use the traditional adaptive tracking control (ATC) introduced by Aström and Wittenmark [1] such that, for $n \geq 0$,

$$y_{n+1} = {}^*\hat{\theta}_n \Phi_n. \tag{30}$$

It is well known that the ATC is almost surely defined if the following assumption is satisfied:

(A_2) For $n \geq 0$, the $\mathcal{F}_n$ conditional distribution of ε_{n+1} is absolutely continuous with respect to the Lebesgue measure.

Remark. If (A_2) is satisfied, Caines [8] has shown how to solve the zero divisor problem for the ATC. We will see in the next section how to avoid this assumption.

Throughout the following, we assume that the driven noise ε has constant conditional covariance matrix Γ given by (29). We also make use of the two classical assumptions about ε:

(N_1) ε has finite conditional moment of order >2;

(N_2) ε is independent and identically distributed with mean 0 and covariance matrix Γ. Therefore, if (N_1) or (N_2) are fulfilled, ε satisfies LN, i.e., if

$$\Gamma_n = \frac{1}{n}\sum_{k=1}^{n} \varepsilon_k {}^*\varepsilon_k, \tag{31}$$

Γ_n converges to Γ almost surely. Finally, we need the following usual assumption of causality:

(A_3) $d_2 \leq d_1$ and the matrix B_1 is of full rank d_2. Moreover, if B_+ denotes the left inverse of B_1 and if $D(R) = B_+ R^{-1} B(R)$ for the shift-back operator R, then D is causal.

Throughout this section, we use a similar approach as that of Bercu and Duflo [4] in the ARX framework. Let $C = (C_n)$ be the average cost matrix sequence defined by

$$C_n = \frac{1}{n}\sum_{k=1}^{n}(Y_k - y_k)^*(Y_k - y_k). \tag{32}$$

The ATC is said to be optimal if C_n converges almost surely to Γ. If we use the ATC, we have from (30)

$$Y_{n+1} - y_{n+1} = \pi_n + \varepsilon_{n+1}. \tag{33}$$

Then, we can easily prove that

$$\|C_n - \Gamma_n\| = O\left(\frac{1}{n}\sum_{k=1}^{n} \|\pi_{k-1}\|^2\right) \quad \text{a.s.} \tag{34}$$

Therefore, in order to show the ATC optimality, we only have to prove that the prediction errors sequence satisfies

$$\sum_{k=1}^{n} \|\pi_k\|^2 = o(n) \quad \text{a.s.} \tag{35}$$

THEOREM 2. *For the model and the* WLS *algorithm* (1) *to* (14), *assume that* (A_1)–(A_3) *and* (N_1) *or* (N_2) *are satisfied. For the tracking trajectory* $y = (y_n)$, *suppose that, for* $n \geq 0$, y_{n+1} *is measurable with respect to* $\mathcal{F}_n$ *and*

$$\sum_{k=1}^{n} \|y_k\|^2 = O(n) \quad \textit{a.s.} \tag{36}$$

If $a_n = f(\log s_n)$ with $f \in F$ and if $s_n^{-1} = O(a_n)$, then the ATC *is optimal. Moreover, we have*

$$\sum_{n=1}^{\infty} \frac{1}{n} \|Y_n - y_n - \varepsilon_n\|^2 < +\infty \quad a.s. \tag{37}$$

Finally, $\hat{\Gamma}_n$ and $\tilde{\Gamma}_n$ are both strongly consistent estimators of Γ.

Remark. For the SG algorithm, the ATC optimality was established by Goodwin, Ramadge, and Caines [18]. Such a theorem was never proven for the ELS algorithm.

Proof. The proof is given in Appendix B.

We now give a useful excitation transfer (ET) lemma similar to the one established by Lai and Wei [25]. We begin by stating the following assumption of irreducibility which uses the same notation as (A_3):

(A_4) The matrix B_+B_q is regular and the polynomials of matrices B_+A, B_+C and D are left coprime.

EXCITATION TRANSFER LEMMA. *Suppose that (A_3) and (A_4) are satisfied. Then we can find a constant $M > 0$ such that, for $n \geq s$,*

$$\lambda_{\min}\left(\sum_{k=0}^{n} \Psi_k {}^*\Psi_k\right) \geq M \lambda_{\min}\left(\sum_{k=s}^{n} H_{k+1} {}^*H_{k+1}\right) \quad a.s., \tag{38}$$

where ${}^tH_n = ({}^tY_n^{p+s+1}, {}^t\varepsilon_n^{r+s+1})$ and $s = d_2(q-1)$.

Proof. A proof can be found in Lai and Wei [25] or Duflo [17].

THEOREM 3. *For the model and* WLS *algorithm* (1) *to* (14), *assume that* (A_1)–(A_4) *and* (N_1) *or* (N_2) *are satisfied. Assume that the covariance matrix Γ is regular. For the tracking trajectory $y = (y_n)$, suppose that y_{n+1} is $\mathcal{F}_{n-p-s} \cap \mathcal{F}_{n-r-s}$-measurable with $\|y_n\|^2 = o(n)$ and*

$$\sum_{k=1}^{n} \|y_k\|^2 = O(n) \quad a.s. \tag{39}$$

Moreover, suppose that y is exciting with order $p + s + 1$, i.e.,

$$\liminf \lambda_{\min}\left(\frac{1}{n} \sum_{k=p+s}^{n} y_k^{p+s+1*} y_k^{p+s+1}\right) > 0 \quad a.s. \tag{40}$$

If $a_n = f(\log s_n)$ with $f \in F$ and if $s_n^{-1} = O(a_n^2)$, then the ATC *is optimal*

$$\|\Phi_n\|^2 = o(n), \quad f_n(a) = o(1) \quad a.s., \tag{41}$$

$$\|C_n - \Gamma_n\| = o\left(\frac{1}{nf(\log n)}\right) \quad a.s., \tag{42}$$

$$\sum_{n=1}^{\infty} f(\log n) \|Y_n - y_n - \varepsilon_n\|^2 < +\infty \quad a.s. \tag{43}$$

Moreover, the WLS *estimator $\hat{\theta}_n$ converges almost surely to θ and we obtain*

$$\|\hat{\theta}_{n+1} - \theta\|^2 = O\left(\frac{1}{nf(\log n)}\right) \quad a.s. \tag{44}$$

Finally, $\hat{\Gamma}_n$ and $\tilde{\Gamma}_n$ are both strongly consistent estimators of Γ and we find relations similar to (42) *with $\hat{\Gamma}_n$ or $\tilde{\Gamma}_n$ instead of C_n.*

Proof. The proof is given in Appendix C.

Remark. We can prove the ATC optimality and the strong consistency with a condition less restrictive than (40) for the tracking trajectory. More precisely, let $\lambda = (\lambda_n)$ be a deterministic positive sequence, increasing to infinity, such that (λ_n) has the same behavior as (λ_{n-1}) and $\lambda_n = O(n)$. Assume that y is λ-exciting with order $p + s + 1$, i.e.,

$$\liminf \lambda_{\min}\left(\frac{1}{\lambda_n}\sum_{k=p+s}^{n} y_k^{p+s+1*}y_k^{p+s+1}\right) > 0 \quad \text{a.s.} \tag{45}$$

Then, if $\lambda_n^{-1} = O(a_n^2)$, the ATC is optimal and the WLS estimator is strongly consistent with

$$\|\hat{\theta}_{n+1} - \theta\|^2 = O\left(\frac{1}{\lambda_n f(\log n)}\right) \quad \text{a.s.} \tag{46}$$

We next consider the continually disturbed control (CDC) introduced by Caines [6], [8] such that, for $n \geq 0$,

$$y_{n+1} + \xi_{n+1} = {}^*\hat{\theta}_n\Phi_n, \tag{47}$$

where ξ is a d_1-dimensional exogenous noise, adapted to $\mathbf{F}$, with mean 0 and covariance matrix Λ. The CDC is said to be residually optimal if C_n converges almost surely to $\Gamma + \Lambda$.

THEOREM 4. *For the model and* WLS *algorithm* (1) *to* (14), *take the same assumptions as in Theorem* 3 *except condition* (40) *for the tracking trajectory y. Moreover, for the exogenous noise ξ, assume that the* LN *is satisfied with Λ regular. In addition, assume that ξ is independent of ε, of y, and of the initial state Ψ_0. If $a_n = f(\log s_n)$ with $f \in F$ and if $s_n^{-1} = O(a_n^2)$, then the* CDC *excited by ξ is residually optimal*

$$\|\Phi_n\|^2 = o(n), \quad f_n(a) = o(1) \quad \textit{a.s.}, \tag{48}$$

$$\sum_{n=1}^{\infty} f(\log n)\|Y_n - y_n - \varepsilon_n - \xi_n\|^2 < +\infty \quad \textit{a.s.} \tag{49}$$

Moreover, the WLS *estimator $\hat{\theta}_n$ converges almost surely to θ and we obtain*

$$\|\hat{\theta}_{n+1} - \theta\|^2 = O\left(\frac{1}{nf(\log n)}\right) \quad \textit{a.s.} \tag{50}$$

Proof. The proof is similar to that of Theorem 3.

6. Modified adaptive tracking. We now use a similar approach as that of Guo and Chen [19], [15] in the ELS framework. Assume that (A_1) and (A_3) are satisfied. Without assumption (A_2), to avoid the zero divisor problem with the ATC, we propose a modified WLS estimator. Throughout the following, the major restriction is that the noise is supposed to satisfy (N_1). All the results of this section are also true, without modification, if we assume that (A_2) is satisfied.

Let $\hat{B}_n^1$ be the matrix component of $\hat{\theta}_n$ that estimates B_1. P_n and Q_n are the orthogonal matrices associated with the singular value decomposition of $\hat{B}_n^1$. The columns of P_n are eigenvectors of $\hat{B}_n^{1*}\hat{B}_n^1$ and the columns of Q_n are eigenvectors of ${}^*\hat{B}_n^1\hat{B}_n^1$ [20]. We set

$$\tilde{B}_n^1 = \begin{cases} \hat{B}_n^1 & \text{if } \lambda_{\min}({}^*\hat{B}_n^1\hat{B}_n^1) > 0, \\ \hat{B}_n^1 + \sqrt{\nu_n}P_n{}^*Q_n & \text{otherwise,} \end{cases} \tag{51}$$

for any positive, deterministic and summable sequence $\nu = (\nu_n)$. By (51), $\tilde{B}_n^1$ is clearly of full rank d_2. Denote by $\tilde{\theta}_n$ the modified WLS estimator of θ where $\hat{B}_n^1$ is replaced by $\tilde{B}_n^1$ in $\hat{\theta}_n$. It immediately follows from (51) that

$$\|\hat{\theta}_n - \tilde{\theta}_n\|^2 \leq \nu_n \quad \text{a.s.} \tag{52}$$

Hence, since $\nu_n = o(1)$, the WLS algorithm is not modified for parameter estimation. We first consider the modified ATC such that, for $n \geq 0$,

$$y_{n+1} = {}^*\tilde{\theta}_n \Phi_n. \tag{53}$$

We clearly have from (53)

$$Y_{n+1} - y_{n+1} = \tilde{\pi}_n + \varepsilon_{n+1}, \tag{54}$$

where

$$\tilde{\pi}_n = \pi_n + (\hat{B}_n^1 - \tilde{B}_n^1)U_n. \tag{55}$$

Therefore, the modified ATC is optimal if

$$\sum_{k=1}^{n} \|\tilde{\pi}_k\|^2 = o(n) \quad \text{a.s} \tag{56}$$

THEOREM 5. *For the model and the* WLS *algorithm* (1) *to* (14), *assume that* (A_1), (A_3), *and* (N_1) *are satisfied. For a positive, nonincreasing and deterministic sequence* $\alpha = (\alpha_n)$ *such that* $\alpha_n = O(n)$, *assume that we have* $\|\varepsilon_n\|^2 = O(\alpha_n)$. *For the tracking trajectory* $y = (y_n)$, *suppose that, for* $n \geq 0$, y_{n+1} *is* $\mathcal{F}_n$*-measurable with* $\|y_n\|^2 = O(\alpha_n)$ *and*

$$\sum_{k=1}^{n} \|y_k\|^2 = O(n) \quad a.s. \tag{57}$$

If $a_n = f(\log s_n)$ *with* $f \in F$ *and if* $s_n^{-1} = O(a_n)$, *then the modified* ATC *is optimal. Moreover, consider the positive random sequence* $v = (v_n)$ *such that* $v_n = \alpha_n + a_n^{-1}$. *Then we also have*

$$\|\Phi_n\|^2 = O(v_{n+1}) \quad a.s., \tag{58}$$

$$\|C_n - \Gamma_n\| = o\left(\frac{v_n}{n}\right) \quad a.s., \tag{59}$$

$$\sum_{n=1}^{\infty} \frac{1}{v_n} \|Y_n - y_n - \varepsilon_n\|^2 < +\infty \quad a.s. \tag{60}$$

Finally, $\hat{\Gamma}_n$ *and* $\tilde{\Gamma}_n$ *are both strongly consistent estimators of* Γ *and we find relations similar to* (59) *with* $\hat{\Gamma}_n$ *or* $\tilde{\Gamma}_n$ *instead of* C_n.

Proof. The proof is given in Appendix D.

Concerning adaptive tracking, we now give the last but most important theorem of this paper. It ensures both strong consistency for the WLS estimator and modified continually disturbed control (CDC) optimality. We recall here that the following theorem is also true, without modification, if we assume that (A_2) is satisfied. Before stating it, we assume in (51) that the sequence $(n\nu_n)$ is summable.

Consider a deterministic positive sequence $\lambda = (\lambda_n)$, increasing to infinity, such that, for $n \geq 1$, $\lambda_n - \lambda_{n-1} \leq 1$, (λ_n) has the same behavior as (λ_{n-1}) and $\lambda_n = O(n)$. Let ξ be a d_1-dimensional exogenous noise adapted to **F** with mean 0 and covariance matrix Λ. Set, for $n \geq 1$, $\chi_n = \sqrt{\lambda_n - \lambda_{n-1}}\xi_n$. We use the CDC introduced by Bercu and Duflo [4] such that, for $n \geq 0$,

$$y_{n+1} + \chi_{n+1} =^* \tilde{\theta}_n \Phi_n. \tag{61}$$

It follows immediately from (61) that

$$Y_{n+1} - y_{n+1} - \chi_{n+1} = \tilde{\pi}_n + \varepsilon_{n+1}. \tag{62}$$

THEOREM 6. *For the model and the* WLS *algorithm* (1) *to* (14), *assume that* (A_1), (A_3), (A_4), *and* (N_1) *are satisfied and that* Γ *is regular. For a positive, nonincreasing and deterministic sequence* $\alpha = (\alpha_n)$ *such that* $\alpha_n = O(n)$, *assume that* $\|\varepsilon_n\|^2 = O(\alpha_n)$. *For the tracking trajectory* $y = (y_n)$, *suppose that* y_{n+1} *is* $\mathcal{F}_{n-p-s} \cap \mathcal{F}_{n-r-s}$*-measurable with* $\|y_n\|^2 = O(\alpha_n)$ *and*

$$\sum_{k=1}^{n} \|y_k\|^2 = O(n) \quad a.s. \tag{63}$$

Moreover, assume that the exogenous noise ξ *satisfies* (N_1) *with* Λ *regular. In addition, suppose that* ξ *is independent of* ε, *of the initial state* Ψ_0, *and of the tracking trajectory* y, *and that* $\|\xi_n\|^2 = O(\alpha_n)$. *Consider the positive random sequence* $v = (v_n)$ *such that* $v_n = \alpha_n + a_n^{-1}$. *Assume that* $v_n = o(\lambda_n)$ *and* $\lambda_n^{-1} = O(a_n^2)$. *If* $a_n = f(\log s_n)$ *with* $f \in F$ *and if* $s_n^{-1} = O(a_n^2)$, *then*

$$\|\Phi_n\|^2 = o(\lambda_n), \quad f_n(a) = o(1) \quad a.s., \tag{64}$$

$$\sum_{n=1}^{\infty} f(\log n)\|Y_n - y_n - \chi_n - \varepsilon_n\|^2 < +\infty \quad a.s., \tag{65}$$

$$\frac{1}{\lambda_n}\sum_{k=1}^{n}(Y_k - y_k - \varepsilon_k)^*(Y_k - y_k - \varepsilon_k) \to \Lambda \quad a.s., \tag{66}$$

$$\|C_n - \Gamma_n\| = O\left(\frac{\lambda_n}{n}\right) \quad a.s. \tag{67}$$

Therefore, if $\lambda_n = o(n)$, *the modified* CDC *excited by* χ *is optimal. Moreover, the* WLS *estimator* $\hat{\theta}_n$ *converges almost surely to* θ *and*

$$\|\hat{\theta}_{n+1} - \theta\|^2 = O\left(\frac{1}{\lambda_n f(\log n)}\right) \quad a.s. \tag{68}$$

Finally, $\hat{\Gamma}_n$ *and* $\tilde{\Gamma}_n$ *are both strongly consistent estimators of* Γ *and we find relations similar to* (67) *with* $\hat{\Gamma}_n$ *or* $\tilde{\Gamma}_n$ *instead of* C_n.

Proof. The proof is given in Appendix E.

Remark. If assumption (N_1) is satisfied with $\alpha > 2$, then, by use of the conditional Borel–Cantelli lemma, we can take $\alpha_n = n^{\beta}$ with $2\alpha^{-1} < \beta < 1$. We can also choose $a_n = (\log s_n)^{-1-\gamma}$ with $\gamma > 0$. Therefore, if we take $\lambda_n = n^{\delta}$ with $\beta < \delta < 1$, we obtain the convergence rates $n^{-\delta}(\log n)^{1+\gamma}$ for the strong consistency and $n^{\delta-1}$ for the optimality. In addition, if ε and ξ are Gaussian white noises, then, using again the Borel–Cantelli lemma, we can take $\alpha_n = \log n$. On one hand, if we focus our attention on the strong consistency, we can use the same choice as above and find the convergence rate $n^{-\delta}(\log n)^{1+\gamma}$. On the other hand, if we are interested in the optimality, we can take $\lambda_n = (\log n)^{\delta}$ with $2(1+\gamma) \leq \delta$ and we obtain the convergence rate $n^{-1}(\log n)^{\delta}$. One can realize that the attenuation $\lambda = (\lambda_n)$ plays a prominent part, reducing the role of the weighting sequence $a = (a_n)$.

7. Survey on adaptive tracking. We now give a short survey on earlier related works on adaptive tracking. We complete this section by comparing our work with previous similar results.

Concerning the SG algorithm, Goodwin, Ramadge, and Caines [18] proved global convergence and adaptive tracking control (ATC) optimality. In the scalar tracking problem, Becker, Kumar, and Wei [2] established convergence to a random multiple of the parameter to be estimated. If the tracking trajectory is sufficiently rich, Kumar and Praly [21] showed, in the scalar case, strong consistency and ATC optimality. Caines [6], [8] realized that, in order to enforce strong consistency, it is necessary to modify the ATC of Aström and Wittenmark [1] and he introduced the CDC. In the scalar case, Caines and Lafortune [7] obtained the first results of CDC optimality and persistent excitation. For the same purpose, Chen [9], [11] chose a weak hypothesis of excitation and using this assumption, Chen and Caines [10] proved, in the scalar case, strong consistency and CDC residual optimality. In a multidimensional framework and with a restrictive assumption on the noise, Chen and Guo [13] established both strong consistency and CDC optimality.

Concerning the ELS algorithm, Solo [27] gave, in the scalar case, a persistent excitation condition in order to guarantee strong consistency. Lai and Wei [23], [24] proposed a weaker excitation condition to obtain strong consistency. For bounded noise, they used a rather complicated control to obtain both strong consistency and CDC optimality. Under the same condition but in a multidimensional framework, Lai and Wei [25] showed strong consistency and gave an excitation transfer theorem useful in obtaining persistent excitation results. Analogously, in a multidimensional framework, Chen and Guo [12] gave conditions to obtain strong consistency. Then they used a rather complex control to prove both strong consistency and CDC optimality with almost sure rates of convergence. Chen and Zhang [14] established similar results in the multi-delay case. Recently Kumar [22] showed, for white Gaussian noise and in a regression framework, the existence of an almost sure limit for the least squares estimator, for almost all parameter values. Strong consistency and optimality results followed. Sin and Goodwin [26] introduced the modified least squares (MLS) algorithm and obtained results similar to those of Goodwin, Ramadge, and Caines [18]. Chen [9], [11] also introduced an algorithm similar to the MLS and in a multidimensional framework he proved strong consistency and CDC residual optimality.

Recently, Guo and Chen [19], [15] established the most important result concerning adaptive tracking for ARMAX models. They found a solution to the twenty-year-old adaptive tracking problem proposed by Aström and Wittenmark [1]. In a multidimensional framework, they proved both strong consistency of ELS estimator and CDC optimality. They also provided almost sure rates of convergence. The key idea was an over-estimation of the ARMAX regression vector norm.

With the WLS algorithm, we have also given a solution to the Aström and Wittenmark [1] adaptive tracking problem. We now compare our work to the results of Guo and Chen [19].

- One can remark that the WLS algorithm is similar to the ELS. The main difference is the easy introduction of a random weighting sequence $a = (a_n)$ in relation (10).
- To obtain strong consistency results, Guo and Chen [19] always required that the driven noise ε had finite conditional moment of order > 2. In §4, we showed how to avoid this assumption by the choice of an admissible weighting sequence. Moreover, it is easy to see via (24) that the WLS estimator performs as well as that of the ELS for ARMAX parameter estimation.
- Furthermore, to obtain adaptive tracking results, Guo and Chen [19] proposed a modified ELS estimator. One can realize that they established CDC optimality by use of a rather technical procedure. Our modification (51) is really simple. Moreover, via (26), we can easily prove the CDC optimality. In addition, our results are also true without modification if the continuity assumption (A_2) on the distribution of ε is satisfied. One can remark that such a result has not been proved by Guo and Chen [19] with pure ELS estimator.
- Finally, we have shown that the WLS algorithm is really easy to handle. We can choose the weighting sequence or the attenuation as we want to privilege the strong consistency or the optimality. One can also realize that our convergence rates are more precise. For example, suppose that we focus our attention on the control optimality. If ε is a Gaussian white noise, we can take the attenuation $\lambda_n = (\log n)^4$. Then, we obtain from (66) a convergence rate in power of log n. It improves the result of Guo and Chen [19] as they founded a convergence rate in power of n.

8. Conclusion. Finally, as it was done for the ELS algorithm, we have shown that the WLS algorithm has rather attractive properties. Under classical assumptions, we have proved both strong consistency of the WLS estimator and CDC optimality. We have also established almost sure rates of convergence. We can easily guess that the weighted estimation can be used in many other frameworks. For instance, the adaptive tracking problems for linear ARMAX models with time varying parameters or for functional ARMAX models remain to be studied, following the choice of suitable weighting sequences.

Appendix A. We make use of the following two lemmas.

LEMMA 1. *Set* $f_n(a) = a_n {}^*\Phi_n S_n^{-1}(a)\Phi_n$ *and* $g_n(a) = a_n {}^*\Phi_n S_{n-1}^{-1}(a)\Phi_n$. *Then*

$$(1 - f_n(a)) = (1 + g_n(a))^{-1} \quad \textit{so} \quad 0 \leq f_n(a) \leq 1;$$

$$\hat{\varepsilon}_{n+1} = (1 - f_n(a))(\pi_n + \varepsilon_{n+1}), \qquad \hat{\theta}_{n+1} = \hat{\theta}_n + a_n S_{n-1}^{-1}(a)\Phi_n {}^*\hat{\varepsilon}_{n+1};$$

$$f_n(a) \leq \inf\{1, \log(\det S_n(a)) - \log(\det S_{n-1}(a))\}.$$

LEMMA 2. *Assume that* ε *is a martingale difference sequence satisfying* (5). *For a vectorial random sequence* $\varphi = (\varphi_n)$ *adapted to* $\mathbf{F}$, *set*

$$M_{n+1} = \sum_{k=0}^{n*} \varphi_k \varepsilon_{k+1}.$$

Then we always have

$$E\left[\sup_{k \leq n+1} |M_k|^2 = O\left(E\left[\sum_{k=0}^{n} \|\varphi_k\|^2\right]\right).\right.$$

We can easily prove Lemma 1 using the same arguments as without the weighting sequence $a = (a_n)$. Lemma 2 can be established by the use of a stopping time argument together with the Kolmogorov's inequality [17]. It can also be proved via the Burkholder, Davis, and Gundy inequality [16], [28].

Proof of Theorem 1. For $n \geq 0$, set $\breve{\theta}_n = \hat{\theta}_n - \theta, \breve{\varepsilon}_n = \hat{\varepsilon}_n - \varepsilon_n$, and $v_n = \operatorname{tr}({}^*\breve{\theta}_n S_{n-1}(a)\breve{\theta}_n)$, where $S_{-1}(a) = S$. By Lemma 1, we can find the following relation similar, without the weighting sequence $a = (a_n)$, to the well-known equality due to Caines [8], Chen [11], Duflo [17], Guo and Chen [11], [15], or Lai and Wei [25]:

$$\begin{aligned} v_{n+1} + P_{n+1} \; = \; & v_0 + \sum_{k=0}^{n} a_k \|\alpha_{k+1}\|^2 + 2\sum_{k=0}^{n} a_n f_k(a) \|\varepsilon_{k+1}\|^2 \\ & +2\mathrm{Re}(M_{n+1}) - 2\mathrm{Re}(L_{n+1}), \end{aligned} \tag{A.1}$$

with $\alpha_n = -{}^*\breve{\theta}_n \Phi_{n-1}$; $\beta_n = {}^*\breve{\theta}_n \Phi_n + f_n(a)\pi_n$; and

- $P_{n+1} = \sum_{k=0}^{n} a_k f_k(a)(1 - f_k(a))\|\pi_k + \varepsilon_{k+1}\|^2$,
- $M_{n+1} = \sum_{k=0}^{n} a_k {}^*\beta_k \varepsilon_{k+1}$,
- $L_{n+1} = \sum_{k=0}^{n} a_k {}^*\alpha_{k+1} \breve{\varepsilon}_{k+1}$.

Moreover, since (A_1) is satisfied, $\alpha_n = C(R)\hat{\varepsilon}_n$ and $a = (a_n)$ is positive and nonincreasing, we can find a positive constant l and an integrable random variable L such that

$$2\mathrm{Re}(L_{n+1}) + L \geq (1 + l)\sum_{k=0}^{n} a_k \|\alpha_{k+1}\|^2. \tag{A.2}$$

In addition, it immediately follows from Lemma 2 that

$$E\left[\sup_{k \leq n+1} |M_k|^2\right] = O\left(E\left[\sum_{k=0}^{n} a_k^2 \|\beta_k\|^2\right]\right). \tag{A.3}$$

Therefore, recalling that $\beta_n = -\alpha_{n+1} - f_n(a)\varepsilon_{n+1}$, we obtain that either

$$E\left[\sup_{n \geq 1} |M_n|\right] < +\infty \tag{A.4}$$

or

$$E\left[\sup_{k \leq n+1} |M_k|\right] = o\left(E\left[\sum_{k=0}^{n} a_k(\|\alpha_{k+1}\|^2 + f_k(a)\|\varepsilon_{k+1}\|^2)\right]\right). \tag{A.5}$$

Finally, by (A.1) and (A.2), we find that

$$E\left[\sup_{k \leq n} v_{k+1}\right] = O\left(E\left[\sum_{k=0}^{n} a_k f_k(a) \|\varepsilon_{k+1}\|^2\right]\right). \tag{A.6}$$

Now, from (5) and (15) together with the monotone convergence theorem,

$$E\left[\sum_{n=0}^{\infty} a_n f_n(a) \|\varepsilon_{n+1}\|^2\right] < +\infty. \tag{A.7}$$

Then we obtain (18) from (A.6) and (A.7). Next, we also obtain (19) and (21) from (A.1), (A.2), (A.7) and the passivity assumption. It remains to show (20) and (22). Let $\pi_n = {}^*\theta\Psi_n - {}^*\hat{\theta}_n\Phi_n$

be the prediction error at time n. By use of Lemma 1, we have $(1 - f_n(a))\pi_n = \check{\varepsilon}_{n+1} + f_n(a)\varepsilon_{n+1}$. Hence, (21) and (A.7) imply

$$E\left[\sum_{n=0}^{\infty} a_n(1 - f_n(a))^2\|\pi_n\|^2\right] < +\infty. \tag{A.8}$$

Recalling (A.1), we also find that

$$E\left[\sup_{n\geq 1} P_n\right] < +\infty, \tag{A.9}$$

$$E\left[\sum_{n=0}^{\infty} a_n f_n(a)(1 - f_n(a))\|\pi_n\|^2\right] < +\infty, \tag{A.10}$$

and we clearly deduce (22) from (A.8) and (A.10). Furthermore, by the matrix inversion formula of Riccati, we obtain

$${}^*\Phi_n S_n^{-2}(a)\Phi_n = (1 - f_n(a))^{2*}\Phi_n S_{n-1}^{-2}(a)\Phi_n. \tag{A.11}$$

So if we set $d = d_1 p + d_2 q + d_1 r$, we obtain, by Lemma 1,

$$a_n\lambda_{\min}S_{n-1}(a)^*\Phi_n S_n^{-2}(a)\Phi_n \leq df_n(a)(1 - f_n(a)). \tag{A.12}$$

Finally, (10) and (14), together with (A.12) and (22), imply (20), completing the proof of Theorem 1. □

Appendix B.

Proof of Theorem 2. Denote by s_∞ and a_∞ the limits of the sequences $s = (s_n)$ and $a = (a_n)$, respectively. To use relation (26) together with Kronecker's lemma, we first have to show that $s_\infty = +\infty$. If we assume that $s_\infty < +\infty$, it follows from the assumption $s_n^{-1} = O(a_n)$ that necessarily $a_\infty > 0$. Hence, using (21), we have

$$\sum_{n=0}^{\infty} \|\Phi_n - \Psi_n\|^2 < \infty. \tag{B.1}$$

If we set $q_n = \sum_{k=0}^{n} \|\Psi_k\|^2 + s$, we can easily see that

$$q_n \leq 2\sum_{k=0}^{n} \|\Phi_k - \Psi_k\|^2 + 2s_n, \tag{B.2}$$

so $q_\infty < +\infty$ where q_∞ denotes the limit of the sequence $q = (q_n)$. But we also have from (6) that

$$q_n \geq \sum_{k=0}^{n} \|\varepsilon_k\|^2 \tag{B.3}$$

and as ε satisfies the LN, we get $n = O(q_n)$. Finally, we lead to a contradiction so that $s_\infty = +\infty$, $a_\infty = 0$. Moreover, we also have

$$s_n \leq 2\sum_{k=0}^{n} \|\Phi_k - \Psi_k\|^2 + 2q_n. \tag{B.4}$$

Then, by use of (21) together with Kronecker's lemma, we find that

$$\sum_{k=0}^{n} \|\Phi_k - \Psi_k\|^2 = o(a_n^{-1}) \tag{B.5}$$

and, as $a_n^{-1} = O(s_n)$, we obtain $s_n = O(q_n)$. In addition, from (6), we have

$$q_n = O\left(\sum_{k=0}^{n} \|Y_k\|^2 + \sum_{k=0}^{n} \|U_k\|^2 + \sum_{k=0}^{n} \|\varepsilon_k\|^2\right). \tag{B.6}$$

Hence, assumptions (N_1) or (N_2) imply that

$$q_n = O\left(n + \sum_{k=0}^{n} \|Y_k\|^2 + \sum_{k=0}^{n} \|U_k\|^2\right). \tag{B.7}$$

Recalling (1), we have $U_{n-1} = D^{-1}(R)B_+A(R)Y_n - D^{-1}(R)B_+C(R)\varepsilon_n$, where R is the shift-back operator. Then, using (B.7), we find that

$$q_{n-1} = O\left(n + \sum_{k=1}^{n} \|Y_k\|^2\right) \tag{B.8}$$

and so, as $s_n = O(q_n)$, we prove that

$$s_{n-1} = O\left(n + \sum_{k=1}^{n} \|Y_k\|^2\right). \tag{B.9}$$

By (26) and the assumption (36) for the tracking trajectory, we have

$$\sum_{k=1}^{n} \|Y_k\|^2 = o(s_{n-1}) + O(n). \tag{B.10}$$

Finally, using (B.9), we obtain that

$$\sum_{k=1}^{n} \|Y_k\|^2 = O(n), \qquad s_n = O(n). \tag{B.11}$$

From (26) together with Kronecker's lemma, we conclude that the ATC is optimal. Moreover, (26) immediately implies (37). We complete the proof using Corollary 3. □

Appendix C.

Proof of Theorem 3. Using the same ideas developed by Bercu and Duflo [4] in the ARX framework, we now prove the Theorem 3. We have already seen in Theorem 2 the ATC optimality with

$$\sum_{k=1}^{n} \|\pi_k\|^2 = o(n). \tag{C.1}$$

Set, for $n \geq 0$, ${}^tL_r = ({}^ty_n^{p+s+1} + {}^t\varepsilon_n^{p+s+1}, {}^t\varepsilon_n^{r+s+1})$. Using the notation of the ET Lemma, we obtain from (33),

$$\|H_{n+1} - L_{n+1}\|^2 = \sum_{k=1}^{p+s+1} \|\pi_{n-k+1}\|^2. \tag{C.2}$$

Then it follows from (C.1) and (C.2) that

$$\sum_{k=0}^{n} \|H_{k+1} - L_{k+1}\|^2 = o(n). \tag{C.3}$$

Moreover, as y is strongly exciting with order $p + s + 1$, we have

$$\liminf \lambda_{\min} \left(\frac{1}{n} \sum_{k=0}^{n} y_{k+1}^{p+s+1*} y_{k+1}^{p+s+1} \right) > 0. \tag{C.4}$$

Furthermore, Γ is regular. Hence, if we assume that y_{n+1} is $\mathcal{F}_{n-p-s} \cap \mathcal{F}_{n-r-s}$-measurable, we find, by use of a classical excitation transfer property proved by Duflo [17] or Lai and Wei [25], that

$$\liminf \lambda_{\min} \left(\frac{1}{n} \sum_{k=0}^{n} L_{k+1}{}^* L_{k+1} \right) > 0. \tag{C.5}$$

Then (C.3) together with (C.5) imply

$$\liminf \lambda_{\min} \left(\frac{1}{n} \sum_{k=0}^{n} H_{k+1}{}^* H_{k+1} \right) > 0. \tag{C.6}$$

Now, if we set

$$Q_n = \sum_{k=0}^{n} \Psi_k{}^* \Psi_k + Q, \qquad a_n Q_n \leq Q_n(a) \leq Q_n. \tag{C.7}$$

Hence, by use of the ET Lemma, we find that $n = O(\lambda_{\min} Q_n)$, which implies $na_n = O(\lambda_{\min} Q_n(a))$. In addition, we have already proved that $s_n = O(n)$. Consequently, from the assumption $(a_n s_n)^{-1} = O(a_n)$, we find that $\lambda_{\min} Q_n(a) \to +\infty$. Next, $q_n = O(n)$, so $\log(\lambda_{\max} Q_n(a)) = O(\log(n))$ and $\log(\lambda_{\max} Q_n(a)) = o(\lambda_{\min} Q_n(a))$. Therefore, via a well-known transfer property, we can conclude that $na_n = O(\lambda_{\min} S_n(a))$. Finally, the WLS estimator is strongly consistent and from (24) we obtain the convergence rate given in (44). Moreover, by use of (27), we also find that

$$\sum_{n=0}^{\infty} a_n \|\pi_n\|^2 < +\infty. \tag{C.8}$$

Hence, we obtain, from Kronecker's lemma,

$$\sum_{k=1}^{n} \|\pi_k\|^2 = o(a_n^{-1}). \tag{C.9}$$

Finally, from (34) and (C.9), we obtain the convergence rate given in (42). In addition, we immediately obtain (43) from (C.8). To complete the proof of Theorem 3, we now show that $\|\Phi_n\|^2 = o(n)$. It will clearly lead to $f_n(a) = o(1)_-$. From (C.8), we have $\|\pi_n\|^2 = o(n)$. Then, (33) together with the assumption $\|y_n\|^2 = o(n)$ imply $\|Y_{n+1}\|^2 = o(n)$. Recalling (1) and the causality assumption (A_3), we have

$$U_n = D^{-1}(R) B_+ A(R) Y_{n+1} - D^{-1}(R) B_+ C(R) \varepsilon_{n+1}, \tag{C.10}$$

where R is the shift-back operator. Hence, we can see that $\|U_n\|^2 = o(n)$. Finally $\|\Psi_n\|^2 = o(n)$ and from (21), $\|\Phi_n\|^2 = o(n)$, completing the proof of Theorem 3. □

Appendix D.

Proof of Theorem 5. From the causality assumption (A3), Caines [8] or Guo and Chen [19] proved that we can find a positive constant $\lambda < 1$ such that

$$\|U_{n-1}\|^2 = O(F_n + \alpha_n), \tag{D.1}$$

where

$$F_n = \sum_{k=0}^{n} \lambda^{n-k} \|Y_k\|^2. \tag{D.2}$$

Furthermore, we have from (25) that

$$\|\pi_n\|^2 = o(a_n^{-1} + \|\Phi_n\|^2). \tag{D.3}$$

In addition, (54) and (55) together with (D.1) imply

$$\|Y_{n+1}\|^2 \leq O(\alpha_{n+1}) + O(\|\pi_n\|^2) + \nu_n O(F_{n+1}). \tag{D.4}$$

Therefore, it follows from (21) and (D.1) – (D.4) that

$$\|Y_{n+1}\|^2 \leq O(v_{n+1}) + o(F_{n+1}), \tag{D.5}$$

where $v_n = \alpha_n + a_n^{-1}$. Moreover, as $F_{n+1} = \lambda F_n + \|Y_{n+1}\|^2$, we obtain

$$F_{n+1} \leq \mu F_n + O(v_{n+1}) \tag{D.6}$$

for some positive constant $\mu < 1$. Finally $F_n = O(v_n)$ and we obtain that

$$\|\Phi_n\|^2 = O(v_{n+1}). \tag{D.7}$$

Recalling (55), we also have

$$\|\tilde{\pi}_n\|^2 \leq 2\|\pi_n\|^2 + 2\nu_n \|\Phi_n\|^2. \tag{D.8}$$

Hence, by use of (26) and (D.8), since $\nu = (\nu_n)$ is summable, we find that

$$\sum_{n=0}^{\infty} \frac{\|\tilde{\pi}_n\|^2}{s_n} < +\infty. \tag{D.9}$$

Therefore, as in the proof of Theorem 2, it follows from (D.9) that $s_n = O(n)$. Finally, (56) and (D.9) imply the modified ATC optimality. In addition, it immediately follows from (D.7) that $a_n^{-1} + \|\Phi_n\|^2 = O(v_{n+1})$. Then, we establish from (25) that

$$\sum_{n=0}^{\infty} \frac{\|\tilde{\pi}_n\|^2}{v_{n+1}} < +\infty. \tag{D.10}$$

Finally, (54) and (D.10) together with Kronecker's lemma imply (59) and (60), completing the proof of Theorem 5. □

Appendix E.

Proof of Theorem 6. We prove Theorem 6 using the same approach as Bercu and Duflo [4] in the ARX framework. The exogenous noise ξ satisfies assumption (N_1). Then, as $\lambda_n - \lambda_{n-1} \leq 1$, we obtain, by use of Chow's lemma,

$$\sum_{n=1}^{\infty} \lambda_n^{-1}(\lambda_n - \lambda_{n-1})(\xi_n {}^*\xi_n - \Lambda) < +\infty. \tag{E.1}$$

But $\chi_n = \sqrt{\lambda_n - \lambda_{n-1}}\,\xi_n$, so (E.1) implies immediately that

$$\sum_{n=1}^{\infty} \lambda_n^{-1}(\chi_n {}^*\chi_n - (\lambda_n - \lambda_{n-1})\Lambda) < +\infty. \tag{E.2}$$

Then, as $\lambda = (\lambda_n)$ increases to infinity, we obtain, by Kronecker's lemma,

$$\frac{1}{\lambda_n}\sum_{k=1}^{n} \chi_k {}^*\chi_k \to \Lambda. \tag{E.3}$$

Since Λ is regular, we immediately obtain from (E.3) that

$$\liminf \lambda_{\min}\left(\frac{1}{\lambda_n}\sum_{k=1}^{n} \chi_k {}^*\chi_k\right) > 0. \tag{E.4}$$

Moreover, as $\|\xi_n\|^2 = O(\alpha_n)$, we also have $\|\chi_n\|^2 = O(\alpha_n)$. Consequently, by use of (62) together with the proof of Theorem 5, we obtain the first relation of (64). Therefore, since $v_n = o(\lambda_n)$, we obtain, from (25),

$$\sum_{k=1}^{n} \|\tilde{\pi}_k\|^2 = o(\lambda_n). \tag{E.5}$$

Set, for $n \geq 0$, ${}^tL_n = ({}^ty_n^{p+s+1} + {}^t\varepsilon_n^{p+s+1} + {}^t\chi_n^{p+s+1}, {}^t\varepsilon_n^{r+s+1})$. Using the notation of the ET Lemma, we obtain, from (62),

$$\|H_{n+1} - L_{n+1}\|^2 = \sum_{k=1}^{p+s+1} \|\tilde{\pi}_{n-k+1}\|^2. \tag{E.6}$$

Then, from (E.5) and (E.6), we obtain

$$\sum_{k=0}^{n} \|H_k - L_k\|^2 = o(\lambda_n). \tag{E.7}$$

In addition, by (E.4), we also have

$$\liminf \lambda_{\min}\left(\frac{1}{\lambda_n}\sum_{k=1}^{n} L_k {}^*L_k\right) > 0. \tag{E.8}$$

Finally, (E.7), (E.8), and the ET Lemma imply $\lambda_n = O(\lambda_{\min}Q_n)$. Therefore, as in the proof of Theorem 3, the assumption $(a_n\lambda_n)^{-1} = O(a_n)$ implies $\lambda_n a_n = O(\lambda_{\min}S_n(a))$. Hence, we clearly obtain the second relation of (64). Moreover (68) follows immediately from (24). Recalling (27), (55), and (64), as $v_n = o(\lambda_n)$ and the sequence $(n\nu_n)$ is summable, we also find that

$$\sum_{n=0}^{\infty} a_n\|\tilde{\pi}_n\|^2 < +\infty. \tag{E.9}$$

Then we clearly obtain (65) from (E.9). Finally, we obtain (66) and (67) from (62), (E.3), and (E.5), completing the proof of Theorem 6. □

Acknowledgment. The author expresses his gratitude to Prof. Marie Duflo. This work was completed thanks to her constant attention.

REFERENCES

[1] K. J. Aström and B. Wittenmark, *On self-tuning regulators*, Automatica, 9 (1973), pp. 185–199.

[2] A. H. Becker, P. R. Kumar, and C. Z. Wei, *Adaptive control with the stochastic approximation algorithm: Geometry and convergence*, IEEE Trans. Automat. Control, AC-30 (1985), pp. 330–338.

[3] B. Bercu, *Sur l'estimateur des moindres carrés généralisé d'un modèle* ARMAX, Application à l'identification des modèles ARMA, Ann. Inst. Henri Poincare, 27 (1991), pp. 425–443.

[4] B. Bercu and M. Duflo, *Moindres carrés pondérés et poursuite*, Ann. Inst. Henri Poincare, 28 (1992), pp. 403–430.

[5] B. Bercu, *Estimation pondérée et poursuite pour les modèles* ARMAX, Note au CRAS, Série 1, 314 (1992), pp. 403–406.

[6] P. E. Caines, *Stochastic adaptive control: Randomly varying parameters and continually disturbed controls*, in Control Science and Technology for the Progress of Society, H. Akashi, ed., Pergamon Press, Oxford, 1981, pp. 925–930.

[7] P. E. Caines and S. Lafortune, *Adaptive control with recursive identification for stochastic linear systems*, IEEE Trans. Automat. Control, AC-29 (1984), pp. 312–321.

[8] P. E. Caines, *Linear Stochastic Systems*, John Wiley, New York, 1988.

[9] H. F. Chen, *Recursive system identification and adaptive control by use of the modified least squares algorithm*, SIAM J. Control Optim., 22 (1984), pp. 758–776.

[10] H. F. Chen and P. E. Caines, *The strong consistency of the stochastic gradient algorithm of adaptive control*, IEEE Trans. Automat. Control, AC-30 (1985), pp. 189–192.

[11] H. F. Chen, *Recursive Estimation and Control for Stochastic Systems*, John Wiley, New York, 1985.

[12] H. F. Chen and L. Guo, *Convergence rate of least squares identification and adaptive control for stochastic systems*, Internat. J. Control, 44 (1986), pp. 1459–1476.

[13] ———, *Asymptotically optimal adaptive control with consistent parameter estimates*, SIAM J. Control Optim., 25 (1987), pp. 558–575.

[14] H. F. Chen and J. F. Zhang, *Convergence rates in stochastic adaptive tracking*, Internat. J. Control, 49 (1989), pp. 1915–1935.

[15] H. F. Chen and L. Guo, *Identification and stochastic adaptive control*, Birkhäuser, Boston, 1991.

[16] Y. S. Chow and H. Teicher, *Probability theory: Independence interchangeability and martingales*, Springer-Verlag, Berlin, 1978.

[17] M. Duflo, *Méthodes récursives aléatoires*, Masson, Paris, 1990.

[18] G. C. Goodwin, P. J. Ramadge, and P. E. Caines, *Discrete time stochastic adaptive control*, SIAM J. Control Optim., 19 (1981), pp. 829–853.

[19] L. Guo and H. F. Chen, *The Aström–Wittenmark self-tuning regulator revisited and* ELS*-based adaptive trackers*, IEEE Trans. Automat. Control, 36 (1991), pp. 802–812.

[20] R. Horn and C. Johnson, *Matrix Analysis*, Cambridge University Press, Cambridge, UK, 1985.

[21] P. R. Kumar and L. Praly, *Self-tuning trackers*, SIAM J. Control Optim., 25 (1987), pp. 1053–1071.

[22] P. R. Kumar, *Convergence of adaptive control schemes using least squares parameter estimates*, IEEE Trans. Automat. Control, 35 (1990), pp. 416–424.

[23] T. L. Lai and C. Z. Wei, *Least squares estimates in stochastic regression models with applications to identification and control of dynamic systems*, Ann. Statist., 10 (1982), pp. 154–166.

[24] T. L. Lai and C. Z. Wei, *Extended least squares and their applications to adaptive control and prediction in linear systems*, IEEE Trans. Automat. Control, AC-31 (1986), pp. 898–906.

[25] T. L. Lai and C. Z. Wei, *On the concept of excitation in least squares identification and adaptive control*, Stochastics, 16 (1986), pp. 227–254.

[26] K. S. Sin and G. C. Goodwin, *Stochastic adaptive control using a modified least squares algorithm*, Automatica, 18 (1982), pp. 315–321.

[27] V. Solo, *The convergence of* AML, IEEE Trans. Automat. Control, AC-24 (1979), pp. 958–962.

[28] C. Z. Wei, *Adaptive prediction by least squares predictors in stochastic regression models with applications to time series*, Ann. Statist., 15 (1987), pp. 1667–1682.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 33, No. 1, pp. 107–125, January 1995

© 1995 Society for Industrial and Applied Mathematics
005

NONINTERACTION AND STABILITY VIA INVERTIBLE FEEDBACK LAWS AND SOME EXISTENCE CONDITIONS *

S. BATTILOTTI†

Abstract. For a wide class of nonlinear systems with more inputs than outputs, the authors show that the asymptotic stability of suitable dynamics and the asymptotic stabilizability via dynamic state feedback of suitable systems are necessary to achieve noninteraction and stability via *invertible* feedback laws. These conditions generalize some recent results obtained for the same class of systems. Some interesting existence conditions are also given, and relationships with the above necessary conditions are identified.

Key words. noninteraction, stability, invertible feedback laws

AMS subject classification. 93

1. Formulation of the problem.

We consider nonlinear systems

$$\begin{aligned} \dot{x} &= f(x) + \sum_{j=1}^{m} g_j(x)u_j, \\ y_i &= h_i(x), \qquad i = 1, \ldots, p, \end{aligned} \tag{1}$$

where f and g_j are smooth vector fields and h_i are smooth real-valued functions. Moreover, let $x_0 = 0$ and assume that $f(x_0) = 0$ and $dh_1^T(x_0), \ldots, dh_p^T(x_0)$ are linearly independent vectors. By $G(x)$ we denote the matrix $(g_1(x) \ldots g_m(x))$, by u we denote $(u_1 \ldots u_m)^T$ and by $\mathcal{G}$ we denote the distribution span$\{g_1, \ldots, g_m\}$ or, equivalently, span$\{G\}$. Let us denote by kerdh_i the codistribution that assigns at each x the subspace of $\mathbb{R}^n$ $\{v \in \mathbb{R}^n : \langle w, v \rangle = 0 : w \in \text{span}\{dh_i(x)\}\}$, where $\langle \cdot, \cdot \rangle$ is the standard inner product of $\mathbb{R}^n$, and let $\mathcal{K}_i = \text{ker}dh_i$. Moreover, we denote by $L_\tau \varphi$ the Lie derivative of the smooth function φ along the vector field τ. Given smooth vector fields τ_1 and τ_2, we denote by $[\tau_1, \tau_2]$ the Lie bracket of τ_1 with τ_2. For a given vector field τ and covector field ω we denote by $L_\tau \omega$ the Lie derivative of τ along ω. The reader is referred to [1] for an easy introduction to the above concepts. Also, in what follows, we implicitly restrict our analysis to a neighbourhood of x_0.

A distribution Δ is *involutive* if $[\tau_1, \tau_2] \subset \Delta$ for any $\tau_1, \tau_2 \in \Delta$. A distribution Δ is *invariant* under a vector field τ if $[\Delta, \tau] \subset \Delta$, where $[\Delta, \tau]$ is the distribution spanned by the vector fields $[v, \tau]$ and v, with $v \in \Delta$. Let $\langle f, g_1, \ldots, g_m | \text{span}\{g_i : i \in I\} \rangle$, with $I \subset \{1, \ldots, m\}$, be the smallest distribution that is invariant under $f, g_1, \ldots, g_m$ and contains span$\{g_i : i \in I\}$. Moreover, let $\mathcal{R}_0 = \langle f, g_1, \ldots, g_m | \text{span}\{G\} \rangle$. It is easy to show that $\mathcal{R}_0$ is invariant under feedback laws $u = \alpha(x) + \beta(x)v$, with $\beta(x_0)$ invertible.

A distribution Δ is *controlled invariant* for (1) if $[f, \Delta] \subset \Delta + \mathcal{G}$ and $[g_j, \Delta] \subset \Delta + \mathcal{G}$, $j = 1, \ldots, m$. A distribution Δ is a *controllability distribution* for (1) if it is involutive and there exist a feedback law $u = \alpha(x) + \beta(x)v$, with $\beta(x)$ nonsingular, and a subset $I \subset \{1, \ldots, m\}$ such that $\Delta = \langle f + G\alpha, G\beta_1, \ldots, G\beta_m | \text{span}\{g\beta_j : j \in I\} \rangle$, where β_j is the jth column of the matrix β.

The system (1) is said to be *noninteractive* if there exists a partition $u_{I_1}, \ldots, u_{I_{p+1}}$ of the input vector u, with $I_i \subset \{1, \ldots, m\}$ and $I_i \cap I_j = \{\phi\}$ for $j \neq i$, $i = 1, \ldots, p+1$, such that the ith output is influenced only by u_s with $s \in I_i$. Note that the inputs u_s, $s \in I_{p+1}$, do not influence any output. The system (1) is said to be *noninteractive with stability* if it is

* Received by the editors November 9, 1992; accepted for publication (in revised form) September 10, 1993. This research was supported by Ministero dell'Università e della Ricerca Scientifica e Tecnologica.

† Dipartimento di Informatica e Sistemistica, Università degli Studi di Roma "La Sapienza," Via Eudossiana 18, 00184 Roma, Italy.

noninteractive and locally asymptotically stable in x_0 (in the sense of Lyapunov). If (1) is not noninteractive with stability, we can try to modify its behaviour through suitable feedback laws so as to achieve these properties, while preserving the affine structure of the system and its equilibrium point. Toward this end, we consider dynamic feedback laws of the form

$$\begin{aligned} u =&\ \alpha(x,w)+\beta(x,w)v, \\ \dot{w} =&\ \delta(x,w)+\gamma(x,w)v, \qquad w\in\mathbb{R}^{n^w}, \end{aligned} \tag{2}$$

with $\alpha(0,0)=0$, $\delta(0,0)=0$. Moreover, for any (block) partition $v_1,\dots,v_{p+1}$ of the input vector v let $(\beta_1(x,w)\dots\beta_{p+1}(x,w))$ and $(\gamma_1(x,w)\dots\gamma_{p+1}(x,w))$ be the corresponding partition of $\beta(x,w)$ and $\gamma(x,w)$, respectively. If $n^w=0$, we obtain *static* feedback laws. We say that a static feedback law $u=\alpha(x)+\beta(x)v$ is *regular* at x_0 if the matrix $\beta(x_0)$ is invertible. Also, we say that a feedback law (2) is *invertible* if the system

$$\begin{aligned} &u=\alpha(x,w)+\beta(x,w)v, \\ &\dot{x}=f(x)+G(x)\alpha(x,w)+G(x)\beta(x,w)v, \\ &\dot{w}=\delta(x,w)+\gamma(x,w)v, \qquad w\in\mathbb{R}^{n^w} \end{aligned} \tag{3}$$

is invertible in the sense of Singh [2]. In the linear case, this amounts to requiring that the transfer function matrix associated with (3) has rank m. This class of feedback laws has been considered for the first time in the case of nonlinear systems in [3]. In what follows, we assume that the open and dense set, on which a given feedback law is invertible, contains the origin of the state space.

Next, we formulate the problem of achieving noninteracting control with stability via dynamic feedback. To this end, first we consider the additional dynamics (the clever trick was first introduced in [4] for linear systems)

$$\dot{w}=u^w, \qquad w\in\mathbb{R}^{n^w},$$

where u^w are auxiliary inputs, and define the extended system

$$\begin{aligned} &\dot{x}=f(x)+G(x)u, \\ &\dot{w}=u^w, \\ &y_i=h_i(x), \qquad i=1,\dots,p. \end{aligned} \tag{4}$$

If we set

$$x^e=\begin{pmatrix}x\\w\end{pmatrix},\quad x^e_0=\begin{pmatrix}x_0\\w_0\end{pmatrix}=\begin{pmatrix}0\\0\end{pmatrix},\quad u^e=\begin{pmatrix}u\\u^w\end{pmatrix},$$

$$f^e(x^e)=\begin{pmatrix}f(x)\\0_{n^w\times 1}\end{pmatrix},\quad G^e(x^e)=\begin{pmatrix}G(x) & 0_{n\times n^w}\\ 0_{n^w\times m} & I_{n^w\times n^w}\end{pmatrix},\qquad h^e_i(x^e)=h_i(x),$$
$$i=1,\dots,p,$$

where 0 and I are, respectively, the zero and identity matrices and the subscripts denote the dimensions of these matrices, then (4) can be rewritten in the following form:

$$\begin{aligned} \dot{x}^e =&\ f^e(x^e)+G^e(x^e)u^e, \\ y_i =&\ h^e_i(x^e), \qquad i=1,\dots,p. \end{aligned} \tag{5}$$

Moreover,

$$\mathcal{G}^e = \operatorname{span}\{G^e\}, \quad \mathcal{G}^w = \operatorname{span}\{\partial/\partial w\}, \quad \mathcal{K}_i^e = \ker dh_i^e.$$

According to our notation, let

$$\begin{aligned}\tilde{f}^e(x^e) &= f^e(x^e) + G^e(x^e)\begin{pmatrix}\alpha(x^e)\\ \delta(x^e)\end{pmatrix},\\ \tilde{g}_j^e(x^e) &= G^e(x^e)\begin{pmatrix}\beta_j(x^e)\\ \gamma_j(x^e)\end{pmatrix}, \qquad j = 1,\ldots,p+1,\\ \tilde{G}^e(x^e) &= (\tilde{g}_1^e(x^e)\ldots\tilde{g}_{p+1}^e(x^e)),\end{aligned}$$

and denote by $\tilde{\Sigma}^e$ the corresponding closed-loop system (1), (2). If $n^w = 0$, i.e., the feedback law is *static*, the superscript "*e*" in the above notation will be omitted. Moreover, by stability or stabilizability of a given system we mean stability and, respectively, stabilizability at the origin of its state space.

LND (local noninteracting control via dynamic state feedback). *Find* n^w *and a feedback law* (2) *such that* $\tilde{\Sigma}^e$ *is noninteractive and has vector relative degree at* x_0^e.

LNSD (local noninteracting control with stability via dynamic state feedback). *Find* n^w *and a feedback law* (2) *such that* $\tilde{\Sigma}^e$ *is noninterative, locally asymptotically stable in* x_0^e, *and has vector relative degree at* x_0^e.

The requirement of vector relative degree at x_0^e ensures that at least one input u_s for some $s \in I_i$ *does* influence the ith output (for a definition of vector relative degree see [1]).

In [5] and, more completely, in [6], using a necessary condition proved in [7], the problem of noninteracting control with stability has been completely solved for a wide class of nonlinear systems (1) with $m = p$. This class consists of the systems (1), with $m = p$, such that *noninteraction* can be achieved via static state feedback laws, *regular* at x_0, or, equivalently, the systems that have vector relative degree at x_0. A necessary and sufficient condition to achieve *noninteraction* and *stability* for this class of systems via *invertible* dynamic state feedback is that a suitable dynamics Σ_0 be asymptotically stable and certain systems Σ_i, $i = 1,\ldots,m$, be asymptotically stabilizable via dynamic state feedback law. As shown in [7], the dynamics Σ_0 is the *nonlinear* obstruction to achieve noninteraction and stability via dynamic feedback. In the case of linear systems it is indeed a trivial dynamics. By combining the results contained in [8] and [6], it is not hard to show that for a given nonlinear system (1), with $m = p$, *noninteraction* and *stability* can be achieved via dynamic state feedback law if and only if *noninteraction* can be achieved for (1) via dynamic state feedback law and *noninteraction* and *stability* can be achieved via dynamic state feedback law for a *canonical* system, which has vector relative degree at the origin of its state space and it is obtained from (1) through a *canonical* dynamic extension. This extension is *canonical* in the sense that any other system, obtained from (1) and having vector relative degree at the origin of its state space, can be obtained from the canonical extension of (1) through invertible feedback laws. This impressively generalizes a well-known result for linear systems ([4] and [9]), which states that for a given linear system *noninteraction* and *stability* can be achieved via dynamic state feedback law if and only if *noninteraction* can be achieved via dynamic state feedback law for the system itself. Indeed, as can be easily proved, *noninteraction* and *stability* can always be achieved for the canonical dynamic extension, as defined in [8], of a given linear system.

In [10] the definition of Σ_0 and Σ_i (in this case, the number of these systems is p) has been extended to the case of nonlinear systems with block-partitioned outputs, for which *noninteraction* can be achieved by means of static state feedback laws, which are *regular* at x_0. These systems are such that the controllability distributions $\mathcal{R}_1^*,\ldots,\mathcal{R}_p^*$, where $\mathcal{R}_i^*$ is the maximal controllability distribution for (1) contained in $\cap_{j\neq i}\mathcal{K}_j$ (see [1]), are *compatible* or, equivalently, there is a static feedback law $u=\alpha(x)+\beta(x)v$, regular at x_0, and a (block) partition $\beta_1(x),\dots,\beta_{p+1}(x)$ of $\beta(x)$ such that $\mathcal{R}_i^*=\langle f+G\alpha, G\beta_1,\dots,G\beta_{p+1}|\mathrm{span}\{G\beta_i,G\beta_{p+1}\}\rangle$ and $\mathrm{span}\{G\beta_i,G\beta_{p+1}\}=\mathcal{R}_i^*\cap\mathcal{G}$ for $i=1,\dots,p$ (note that here we abused the notation, since $G\beta_i$ is, in general, a matrix). This class of systems has been characterized from a geometric point of view in [11]. If $m=p$, the set of distributions $\mathcal{R}_1^*,\dots,\mathcal{R}_p^*$ is the *only* set of controllability distributions that is compatible. Moreover, in this case it is not hard to show that the systems Σ_0 and Σ_i, $i=1,\dots,p$, can be *uniquely* (up to static feedback transformation, regular at x_0, and local changes of coordinates) associated to (1). If $m>p$ this uniqueness property is lost (for an idea of the proof, see [11]) and Σ_0 and Σ_i, $i=1,\dots,p$, are *uniquely* (up to static feedback transformation, regular at x_0, and local changes of coordinates) associated to (1), once a set of compatible controllability distributions $\mathcal{R}_1,\dots,\mathcal{R}_p$ is chosen (not necessarily the *maximal* one). Many interesting interrelated properties can be found among the systems Σ_0 and Σ_i, $i=1,\dots,p$, associated to the maximal set $\mathcal{R}_1^*,\dots,\mathcal{R}_p^*$, and those correspondingly associated to a given set of compatible controllability distributions $\mathcal{R}_1,\dots,\mathcal{R}_p$, but this will be pursued elsewhere. For clarity of exposition, we denote by Σ_0^* and Σ_i^*, $i=1,\dots,p$, the systems (uniquely in the above sense) associated to the maximal set $\mathcal{R}_1^*,\dots,\mathcal{R}_p^*$.

In [10] it is also proved that for the above class of systems the asymptotic stability of Σ_0^*, plus the *exponential* stabilizability of the systems Σ_i^*, $i=1,\dots,p$, and of a suitable system Σ_{p+1}^* (which is trivial in the case $m=p$), are sufficient to achieve *noninteraction* and *stability* via dynamic state feedback law. More generally, for the same class of systems with block-partitioned outputs, it is not hard to show that the asymptotic stability of Σ_0^* and the *asymptotic* stabilizability via dynamic state feedback laws of the systems Σ_i^*, $i=1,\dots,p$, are *necessary* to achieve noninteraction and stability via dynamic feedback laws, *regular* in the sense of [13]. There is no apparent relation between this class of feedback laws and the well-known class of invertible feedback laws. Conversely, it is easy to see that the asymptotic stability of Σ_0^* and the *asymptotic* stabilizability of the systems Σ_i^*, $i=1,\dots,p+1$, are also *sufficient* to achieve *noninteraction* and *stability* via regular dynamic state feedback law. Unfortunately, as shown by a counterexample in [13], the *asymptotic* stabilizability of Σ_{p+1}^* via dynamic feedback law is *not* necessary to achieve noninteraction and stability even via *invertible* dynamic feedback laws. This fact, together with the generic nontriviality of Σ_0^*, gives an estimate of how the linear case may differ from the nonlinear case. Indeed, for linear systems, the asymptotic stabilizability of Σ_{p+1}^* is automatically guaranteed by its very definition. Yet, for the class of systems for which Σ_{p+1}^* is trivial, we obtain a set of necessary and sufficient conditions, which generalize the ones given in [6].

In this paper, for the class of nonlinear systems (1), for which *noninteraction* can be achieved via static state feedback laws, regular at x_0, we show that the asymptotic stability of Σ_0^* and the *asymptotic* stabilizability via dynamic state feedback of the systems Σ_i^*, $i=1,\dots,p$, are *necessary* to achieve noninteraction and stability via *invertible* feedback laws (Theorem 4). This generalizes the result given in [7] (and, partially, the results given in [4] and [9] for linear systems). Moreover, our result shows that (a) for the class of systems here considered any invertible feedback law, which solves LND, is also *regular* in the sense of [13], filling the gap between the class of invertible feedback laws and the class of feedback laws, regular in the sense of [13]; (b) any invertible feedback law that solves LND can be expressed as the cascade of a *static* feedback law, regular at x_0, together with an *invertible noninteraction* feedback law (see Definition 9).

A main difficulty of the problem solved in this paper is to show that the controllability distributions $\mathcal{R}_i^e=\langle\tilde{f}^e,\tilde{g}_1^e,\dots,\tilde{g}_{p+1}^e|\mathrm{span}\{\tilde{g}_i^e,\tilde{g}_{p+1}^e\}\rangle$, $i=1,\dots,p$, defined on $\mathbb{R}^{n+n^w}$, project down to well-defined controllability distributions defined on $\mathbb{R}^n$ (see the proof of Theorem 4). This is actually not needed when studying the problem of achieving noninteraction *without stability* (in this case our result is apparently trivial, since decoupling can be achieved via *static* state feedback by assumption), while it becomes essential when addressing the issue of stability. Indeed, this fact allows us to derive stabilizability properties of systems defined on the original state space from stability properties of systems defined on the extended state space. The above property of projection is guaranteed by Proposition 2 and Lemma 3, which is based on the algorithm (18). Among others, one consequence of our result is to connect explicitly the invertibility of the feedback laws, which solve NLD, with a geometric property, given in terms of the algorithm (18). Moreover, this algorithm gives a more general and intuitively geometric version of the one given in [13]. We note once again that the class of systems considered here can always be rendered noninteractive but not necessarily with internal stability. The main result of this paper is to give necessary conditions for also achieving internal stability.

Some interesting existence conditions are also given (§4). Among others, one result is that if noninteraction and stability can be achieved via dynamic feedback laws, which are possibly *noninvertible*, for each $i \in \{1, \dots, p\}$ there must exist a dynamic feedback law that asymptotically stabilizes Σ_i^* on a suitable *invariant submanifold* $\mathcal{M}^i$ of the state space of the system, resulting from Σ_i^* after applying this feedback law, and transversal to the submanifold $h_i^{-1}(0)$. This manifold depends, in general, on the decoupling feedback law, except in some significant cases (one is the case in which we consider *invertible* feedback laws).

2. Some preliminary results.

Before going through the technical details, we give some basic assumptions. In a natural way, as long as we are interested in necessary conditions, we will assume that (1) is locally asymptotically stabilizable at the origin. Let $\mathcal{V}_i^*$ be the maximal controlled invariant distribution for (1) which is contained in $\cap_{j \neq i} \mathcal{K}_j$, $\mathcal{R}_i^*$ be the maximal controllability distribution for (1) contained in $\cap_{j \neq i} \mathcal{K}_j$, $\mathcal{Q}^*$ be the maximal controllability distribution for (1) contained in $\mathcal{R}^* = \cap_{i=1}^{p} \sum_{j \neq i} \mathcal{R}_j^*$ and $\mathcal{R}_0$ be as above. These distributions can be computed by means of standard algorithms [1], which give a sequence of distributions. If these distributions have constant dimension, we will say that $\mathcal{V}_i^*$ (respectively, $\mathcal{R}_i^*$, $\mathcal{Q}^*$ or $\mathcal{R}_0$) are *regularly computable* at x_0. We make the following assumptions.

(H1) $\dim \mathcal{R}_0 = n$;

(H2) the distributions $\mathcal{R}_0$, $\mathcal{V}_i^*$, $\mathcal{R}_i^*$, $i = 1, \dots, p$, and $\mathcal{Q}^*$ are regularly computable at x_0;

(H3) the distributions $\mathcal{R}^*$, $\sum_j \mathcal{R}_j^*$, $(\sum_j \mathcal{R}_j^*) \cap \mathcal{G}$ and $\sum_j (\mathcal{R}_j^* \cap \mathcal{G})$ have constant dimension;

(H4) $\sum_{i=1}^{p} (\mathcal{R}_i^* \cap \mathcal{G}) = \mathcal{G}$.

Assumption (H1) is standard, and, under assumptions (H1)–(H3), assumption (H4) is a well-known necessary and sufficient condition to solve LND via static state feedback laws, regular at x_0 ([11]). It can be easily shown that (H4) is equivalent, under the assumption of constant dimensions of the distributions considered, to the fact that (1) has vector relative degree at x_0.

As in [12], we will use the following terminology. Let $\mathcal{R}_1, \dots, \mathcal{R}_p$ be a set of controllability distributions for (1), such that $\mathcal{R}_i \subset \cap_{j \neq i} \mathcal{K}_j$. Moreover, let $\mathcal{Q}$ be the maximal controllability distribution for (1) contained in $\mathcal{R} = \cap_{i=1}^{p} \sum_{j \neq i} \mathcal{R}_i$ (supposed to exist). We will say that $\mathcal{R}_1, \dots, \mathcal{R}_p$ is a *regular set* (in x_0) if (a) the distributions $\mathcal{Q}$, $\mathcal{Q} \cap \mathcal{G}$, $\sum_j \mathcal{R}_j$, $\mathcal{R}$, $(\sum_j \mathcal{R}_j) \cap \mathcal{G}$ and $\sum_j (\mathcal{R}_j \cap \mathcal{G})$ have constant dimension. If a regular set $\mathcal{R}_1, \dots, \mathcal{R}_p$ is such that (b) $\sum_{i=1}^{p} (\mathcal{R}_i \cap \mathcal{G}) = \mathcal{G}$, we will say that $\mathcal{R}_1, \dots, \mathcal{R}_p$ is a *regular solution* (of LND). Indeed, if the regular set $\mathcal{R}_1, \dots, \mathcal{R}_p$ satisfies (b), by standard arguments it can be shown that there exists a static feedback law $u = \alpha(x) + \beta(x)v$, regular at x_0, such that

$$\mathcal{R}_i = \langle \tilde{f}, \tilde{g}_1, \ldots, \tilde{g}_{p+1} | \mathrm{span}\{\tilde{g}_i, \tilde{g}_{p+1}\}\rangle, \qquad i = 1, \ldots, p,$$

$$\sum_{j \neq i} \mathcal{R}_j = \langle \tilde{f}, \tilde{g}_1, \ldots, \tilde{g}_{p+1} | \mathrm{span}\{\tilde{g}_j : j \neq i\}\rangle, \qquad i = 1, \ldots, p,$$

$$\sum_{j=1}^{p} \mathcal{R}_j = \mathcal{R}_0 = \langle \tilde{f}, \tilde{g}_1, \ldots, \tilde{g}_{p+1} | \mathrm{span}\{\tilde{g}_j : j = 1, \ldots, p+1\}\rangle,$$

$$\mathrm{span}\{\tilde{g}_i, \tilde{g}_{p+1}\} = \mathcal{R}_i \cap \mathcal{G}, \qquad i = 1, \ldots, p$$

(see [1], [11]–[13]). By Frobenius' theorem, since $\sum_{j \neq i} \mathcal{R}_j$, $i = 1, \ldots, p$, is involutive and invariant under $\tilde{f}$ and $\tilde{g}_j$, $j = 1, \ldots, p+1$, it is always possible to find a change of coordinates $z(x) = ((z_1(x))^T \ldots (z_{p+2}(x))^T)^T$ such that the system $\tilde{\Sigma}$, obtained from (1) with $u = \alpha(x) + \beta(x)v$, in the new coordinates has the form

$$\begin{aligned} \dot{z}_i &= \tilde{f}_i(z_i) + \tilde{g}_{ii}(z_i)v_i, \qquad i = 1, \ldots, p, \\ \dot{z}_{p+1} &= \tilde{f}_{p+1}(z_1, \ldots, z_p, z_{p+1}) + \sum_{j=1}^{p} \tilde{g}_{p+1,j}(z_1, \ldots, z_p, z_{p+1})v_j, \\ \dot{z}_{p+2} &= \tilde{f}_{p+2}(z_1, \ldots, z_{p+2}) + \sum_{j=1}^{p+1} \tilde{g}_{p+2,j}(z_1, \ldots, z_{p+2})v_j, \\ y_i &= h_i(z_i), \qquad i = 1, \ldots, p, \end{aligned} \tag{6}$$

with $\sum_{j \neq i} \mathcal{R}_j = \mathrm{span}\{\partial/\partial z_j : j \neq i\}$ and $\mathcal{Q} = \mathrm{span}\{\partial/\partial z_{p+2}\}$. In what follows, for simplicity let $z = x$ and $v = u$. Moreover, we assume that (1) is already in the form (6).

Next, we define some suitable dynamics that are crucial in all the subsequent analysis. First, let us introduce the following distribution. Let $\mathcal{I}$ be the Lie ideal generated by the vector fields $\{[\tilde{g}_i, ad_{\tilde{f}}^k \tilde{g}_j] : k \geq 0,\ i, j = 1, \ldots, p;\ i \neq j\}$ in the Lie algebra generated by $\tilde{f}, \tilde{g}_1, \ldots, \tilde{g}_{p+1}$ and let

$$\Delta_{MIX} = \mathrm{span}\{\tau : \tau \in \mathcal{I}\}. \tag{7}$$

The distribution (7) was first introduced in [10]. It is not hard to show that $\Delta_{MIX} \subset \mathcal{R} \subset \mathcal{R}_i$. Moreover, for linear systems $\Delta_{MIX} = 0$. The distribution $\Delta_{MIX} + \mathcal{Q}$ is *uniquely* associated to (1) and to the regular solution $\mathcal{R}_1, \ldots, \mathcal{R}_p$, in the sense that its definition is invariant under local change of coordinates and static feedback transformations, regular at x_0 (see [10]). Assume that the distributions $\Delta_{MIX} + \mathcal{Q}$, $\mathcal{R}_i + \Delta_{MIX}$ and $\mathcal{R} \cap (\mathcal{R}_i + \Delta_{MIX})$ have constant dimension. In what follows, we will assume that a regular set satisfies these additional assumptions.

By Frobenius' theorem, for each $i \in \{1, \ldots, p\}$ there exists a change of coordinates

$$z^i(x) = (x_1^T \ldots x_{i-1}^T\, x_{i+1}^T \ldots x_p^T\, (z_1^i(x))^T \ldots (z_5^i(x))^T)^T,$$

such that

$$\begin{aligned} \mathcal{R}_i + \mathcal{R} &= \mathrm{span}\{\partial/\partial z_1^i, \partial/\partial z_2^i, \partial/\partial z_3^i, \partial/\partial z_4^i, \partial/\partial z_5^i\}, \\ \mathcal{R}_i + \Delta_{MIX} &= \mathrm{span}\{\partial/\partial z_1^i, \partial/\partial z_3^i, \partial/\partial z_4^i, \partial/\partial z_5^i\}, \\ \mathcal{R} &= \mathrm{span}\{\partial/\partial z_2^i, \partial/\partial z_3^i, \partial/\partial z_4^i, \partial/\partial z_5^i\}, \\ \mathcal{R} \cap (\mathcal{R}_i + \Delta_{MIX}) &= \mathrm{span}\{\partial/\partial z_3^i, \partial/\partial z_4^i, \partial/\partial z_5^i\}, \\ \Delta_{MIX} + \mathcal{Q} &= \mathrm{span}\{\partial/\partial z_4^i, \partial/\partial z_5^i\}, \\ \mathcal{Q} &= \mathrm{span}\{\partial/\partial z_5^i\}. \end{aligned}$$

Note that in these coordinates the ith output function h_i depends only on the coordinate(s) z_1^i.

Since $\Delta_{MIX} + \mathcal{Q}$ is involutive and invariant under $\tilde{f}, \tilde{g}_1, \ldots, \tilde{g}_{p+1}$ (see [10] for a proof) and $\tilde{f}(x_0) = 0$, it makes sense to consider the restriction $\tilde{f}|\mathcal{L}_{x_0}^{\Delta_{MIX}+\mathcal{Q}}$, where $\mathcal{L}_{x_0}^{\Delta_{MIX}+\mathcal{Q}}$ is the leaf (or, equivalently, the maximal integral manifold) of $\Delta_{MIX} + \mathcal{Q}$ passing through x_0. This vector field defines a dynamics evolving on $\mathcal{L}_{x_0}^{\Delta_{MIX}+\mathcal{Q}}$, which in z^i-coordinates is given by

$$\begin{aligned} \dot{z}_4^i &= \tilde{f}_{z_4^i}(0, \ldots, 0, z_4^i), \\ \dot{z}_5^i &= \tilde{f}_{z_5^i}(0, \ldots, 0, z_4^i, z_5^i). \end{aligned}$$

Let us consider the subdynamics

$$\dot{z}_4^i = \tilde{f}_{z_4^i}(0, \ldots, 0, z_4^i). \tag{8}$$

The dynamics (8) is trivial in the case of linear systems, since $\Delta_{MIX} = 0$. The dynamics (8), corresponding to the maximal solution $\mathcal{R}_1^*, \ldots, \mathcal{R}_p^*$, has been referred to as Σ_0^* in the introduction. If $m = p$, under assumptions (H1)–(H4), $\mathcal{Q}^* = 0$, where $\mathcal{Q}^*$ is the maximal controllability distribution contained in $\mathcal{R}^*$, since otherwise the decoupling matrix of (1) would not be invertible at x_0 and, as a consequence, (1) would not have vector relative degree at x_0. Thus, if $m = p$, the dynamics (8) coincides with Σ_0^*, as defined in [7] or, equivalently, in [6].

Similarly, for each $i = 1, \ldots, p$, it makes sense to consider also the restrictions $\tilde{f}|\mathcal{L}_{x_0}^{\mathcal{R}_i+\Delta_{MIX}}$ and $\tilde{g}_i|\mathcal{L}_{x_0}^{\mathcal{R}_i+\Delta_{MIX}}$ (note that by definition $\tilde{g}_i \in \mathcal{R}_i$). These vector fields define a control system evolving on $\mathcal{L}_{x_0}^{\mathcal{R}_i+\Delta_{MIX}}$, expressed in z^i-coordinates by

$$\begin{pmatrix} \dot{z}_1^i \\ \dot{z}_3^i \\ \dot{z}_4^i \\ \dot{z}_5^i \end{pmatrix} = \begin{pmatrix} \tilde{f}_{z_1^i}(0, \ldots, 0, z_1^i) \\ \tilde{f}_{z_3^i}(0, \ldots, 0, z_1^i, 0, z_3^i) \\ \tilde{f}_{z_4^i}(0, \ldots, 0, z_1^i, 0, z_3^i, z_4^i) \\ \tilde{f}_{z_5^i}(0, \ldots, 0, z_1^i, 0, z_3^i, z_4^i, z_5^i) \end{pmatrix} + \begin{pmatrix} \tilde{g}_{z_1^i,i}(0, \ldots, 0, z_1^i) \\ \tilde{g}_{z_3^i,i}(0, \ldots, 0, z_1^i, 0, z_3^i) \\ \tilde{g}_{z_4^i,i}(0, \ldots, 0, z_1^i, 0, z_3^i, z_4^i) \\ \tilde{g}_{z_5^i,i}(0, \ldots, 0, z_1^i, 0, z_3^i, z_4^i, z_5^i) \end{pmatrix} u_i.$$

Let us consider the subsystem

$$\begin{pmatrix} \dot{z}_1^i \\ \dot{z}_3^i \end{pmatrix} = \begin{pmatrix} \tilde{f}_{z_1^i}(0, \ldots, 0, z_1^i) \\ \tilde{f}_{z_3^i}(0, \ldots, 0, z_1^i, 0, z_3^i) \end{pmatrix} + \begin{pmatrix} \tilde{g}_{z_1^u,i}(0, \ldots, 0, z_1^i) \\ \tilde{g}_{z_3^i,i}(0, \ldots, 0, z_1^i, 0, z_3^i) \end{pmatrix} u_i. \tag{9}$$

In the case of linear systems, (9) is a controllable system, since $\Delta_{MIX} = 0$ and $\mathcal{R}_i$ is a controllability subspace. Each system (9), corresponding to the maximal solution $\mathcal{R}_1^*, \ldots, \mathcal{R}_p^*$, has been referred to as Σ_i^* in the Introduction.

Last but not least, we consider the system

$$\dot{z}_5^i = \tilde{f}_{z_5^i}(0, \ldots, 0, z_5^i) + \tilde{g}_{z_5^i,p+1}(0, \ldots, 0, z_5^i)u_{p+1}, \tag{10}$$

which is obtained by considering the restrictions $\tilde{f}|\mathcal{L}_{x_0}^{\mathcal{Q}}$ and $\tilde{g}_{p+1}|\mathcal{L}_{x_0}^{\mathcal{Q}}$ (note that $\tilde{g}_{p+1} \in \mathcal{Q}$). The system (10), corresponding to the maximal solution $\mathcal{R}_1^*, \ldots, \mathcal{R}_p^*$, has been referred to as Σ_{p+1}^* in the Introduction.

It can be easily shown that the dynamics (8)–(10) are *uniquely* associated to (1) and to a given regular solution, in the sense that their definition is independent of static feedback transformations, regular at x_0, and local changes of coordinates.

3. Noninteraction and stability via invertible feedback laws. Let us consider the system (1) (which is assumed to be already in the form (6)) with $\mathcal{R}_i = \mathcal{R}_i^*$. Moreover, let Δ_{MIX}^* be the distribution (7), defined correspondingly to the maximal solution $\mathcal{R}_1^*, \dots, \mathcal{R}_p^*$. As previously noted, the distribution $\Delta_{MIX}^* + \mathcal{Q}^*$ is *uniquely* associated to $\mathcal{R}_1^*, \dots, \mathcal{R}_p^*$ up to local changes of coordinates and static feedback transformations, regular at x_0.

(H5) The distributions $\Delta_{MIX}^* + \mathcal{Q}^*$, $\mathcal{R}_i^* + \Delta_{MIX}^*$ and $\mathcal{R}^* \cap (\mathcal{R}_i^* + \Delta_{MIX}^*)$ have constant dimension.

In this section we prove that for the class of systems here considered the asymptotic stability of Σ_0^* and the asymptotic stabilizability via dynamic feedback law of each system Σ_i^* are necessary to achieve noninteraction and stability via *invertible* feedback laws. To this end, let $(\mathcal{V}_i^e)^*$ be the maximal distribution that is contained in $\cap_{j \neq i} \mathcal{K}_j^e$ and such that $[\tilde{f}^e, (\mathcal{V}_i^e)^*] \subset (\mathcal{V}_i^e)^* + \mathcal{G}^e$ and $[\tilde{g}_j^e, (\mathcal{V}_i^e)^*] \subset (\mathcal{V}_i^e)^* + \mathcal{G}^e$, with $j = 1, \dots, p+1$. This distribution always exists, since, given two distributions $\mathcal{V}_1$ and $\mathcal{V}_2$ that satisfy the above properties, $\mathcal{V}_1 + \mathcal{V}_2$ also has these properties. Thus, the maximal distribution, which satisfies the above properties, is given by the sum of all the distributions with such properties. As a preliminary result, we prove the following lemma.

LEMMA 1. *Under assumptions* (H1)–(H4), *we have*

$$(\mathcal{V}_i^e)^* = \mathcal{V}_i^* + \mathcal{G}^w,$$

where $\mathcal{V}_i^*$ is thought of as a distribution of $\mathbb{R}^{n+n^w}$.

Proof. Let $(\mathbb{R}^s)^*$ be the dual of $\mathbb{R}^s$. For any smooth codistribution $\Omega \subset (\mathbb{R}^s)^*$, denote by $(\Omega)^\perp$ the distribution that assigns at each point $p \in \mathbb{R}^s$ the subspace $\{v \in \mathbb{R}^s : \langle w, v \rangle = 0, w \in \Omega(p)\}$. Let us consider the following sequence of codistributions

$$\begin{aligned} \tilde{\Omega}_{i0}^e &= \sum_{j \neq i} \operatorname{span}\{dh_j^e\}, \\ \tilde{\Omega}_{ik}^e &= L_{\tilde{f}^e}(\tilde{\Omega}_{i,k-1}^e \cap (\mathcal{G}^e)^\perp) + \sum_{j=1}^{p+1} L_{\tilde{g}_j^e}(\tilde{\Omega}_{i,k-1}^e \cap (\mathcal{G}^e)^\perp) + \tilde{\Omega}_{i,k-1}^e. \end{aligned} \tag{11}$$

Using essentially the same arguments as [1, Lem. 6.3.2], it is not hard to show that, if there exists $k_i^* \geq 0$ such that $\tilde{\Omega}_{i,k_i^*}^e = \tilde{\Omega}_{i,k_i^*+1}^e$ and $\tilde{\Omega}_{i,k_i}^e$ and $\tilde{\Omega}_{i,k_i}^e \cap (\mathcal{G}^e)^\perp$ have constant dimension for all k, we have $(\tilde{\Omega}_{ik_i^*}^e)^\perp = (\mathcal{V}_i^e)^*$. Moreover, if $\tilde{\Omega}_{ik}^e = \tilde{\Omega}_{i,k+1}^e$ for some $k \geq 0$, then $\tilde{\Omega}_{ik}^e = \tilde{\Omega}_{ih}^e$ for all $h > k$.

To prove the lemma, we will show that there exists $k_i^* \geq 0$ such that $\tilde{\Omega}_{i,k_i^*}^e = \tilde{\Omega}_{i,k_i^*+1}^e = \mathcal{V}_i^* + \mathcal{G}^w$, where $\mathcal{V}_i^*$ is thought of as a distribution of $\mathbb{R}^{n+n^w}$, and that $(\tilde{\Omega}_{ik}^e)^\perp$ and $\tilde{\Omega}_{ik}^e \cap (\mathcal{G}^e)^\perp$ have constant dimension for all k. For Ω_{i0}^e has constant dimension in a neighbourhood of x_0^e. The codistribution $\tilde{\Omega}_{i0}^e \cap (\mathcal{G}^e)^\perp$ is spanned at each x^e by the rows of $\tilde{\Omega}_{i0}^e(x^e)$ that annihilate the vectors of $\mathcal{G}^e(x^e)$. Since $\tilde{\Sigma}^e$ has vector relative degree at x_0^e, $\tilde{\Omega}_{i0}^e \cap (\mathcal{G}^e)^\perp$ is spanned by the covector fields $dh_{j_l}^e$, $j_l \in \{1, \dots, i-1, i+1, \dots, p\}$, such that $\langle dh_{j_l}^e, G^e \rangle (x^e) = 0$. Since $L_{\tilde{f}^e} dh_{j_l}^e = dL_{\tilde{f}^e} h_{j_l}^e$ and $L_{\tilde{g}_j^e} dh_{j_l}^e = dL_{\tilde{g}_j^e} h_{j_l}^e$, it follows that $L_{\tilde{f}^e}(\tilde{\Omega}_{i0}^e \cap (\mathcal{G}^e)^\perp) = L_{f^e}(\tilde{\Omega}_{i0}^e \cap (\mathcal{G}^e)^\perp)$ and $L_{\tilde{g}_i^e}(\tilde{\Omega}_{i0}^e \cap (\mathcal{G}^e)^\perp) = 0$ for $j = 1, \dots, p+1$. Suppose now that for some $k > 0$ the codistribution $\tilde{\Omega}_{ik}^e$ has constant dimension in a neighbourhood of x_0^e and that $\tilde{\Omega}_{ik}^e \cap (\mathcal{G}^e)^\perp$ is spanned by the covector fields $dL_{f^e}^{k_{j_l}} h_{j_l}^e$, $j_l \in \{1, \dots, i-1, i+1, \dots, p\}$, such that $\langle dL_{f^e}^{k_{j_l}} h_{j_l}^e, G^e \rangle (x^e) = 0$ for all k_l such that $0 \leq k_{j_l} \leq \max\{k, r_{j_l} - 2\}$ ($r_1, \dots, r_p$ is the vector relative degree of (1)). Similarly, we have $L_{\tilde{f}^e}(\tilde{\Omega}_{ik}^e \cap (\mathcal{G}^e)^\perp) = L_{f^e}(\tilde{\Omega}_{ik}^e \cap (\mathcal{G}^e)^\perp)$ and $L_{\tilde{g}_j^e}(\tilde{\Omega}_{ik}^e \cap (\mathcal{G}^e)^\perp) = 0$ for $j = 1, \dots, p+1$. Since Σ^e has vector relative degree at x_0^e, the codistribution $\tilde{\Omega}_{i,k+1}^e$ has always constant dimension in a neighbourhood of x_0^e (see [1, Lem. 5.12] for a proof). Clearly $\tilde{\Omega}_{i,k+1}^e \cap (\mathcal{G}^e)^\perp$ is spanned by the covector fields $dL_{f^e}^{k_{j_l}} h_{j_l}^e$, $j_l \in \{1, \dots, i-1, i+1, \dots, p\}$, such that $\langle dL_{f^e}^{k_l} h_{j_l}^e, G^e \rangle (x^e) = 0$ for all k_{j_l} such that $0 \leq k_{j_l} \leq \max\{k+1, r_{j_l} - 2\}$. The algorithm (11) stops for $k = k_i^* = \max_i\{r_i\}$ -1, when $\tilde{\Omega}_{ik_i^*}^e \cap (\mathcal{G}^e)^\perp = \tilde{\Omega}_{i,k_i^*+1}^e \cap (\mathcal{G}^e)^\perp$ and $k_{j_l} = r_{j_l} - 1$, $l = 1, \dots, p$. This, together with the fact that $(\mathcal{V}_i^*)^\perp = \text{span}\{dh_j, \dots, dL_{\tilde{f}}^{r_j - 1} h_j : j \neq i\}$ [1, Cor. 6.3.14], proves our thesis. □

A consequence of the above lemma is the following proposition. Let $\alpha(x, w)$ and $\beta(x, w)$ be partitioned, according to the (block) partitions of the input vectors u and v, respectively, as

$$\alpha(x,w) = (\alpha_1^T(x,w) \cdots \alpha_{p+1}^T(x,w))^T, \qquad \beta(x,w) = (\beta_1(x,w) \cdots \beta_{p+1}(x,w)),$$

$$\beta_j(x,w) = (\beta_{1j}^T \cdots \beta_{p+1,j}^T)^T, \qquad j = 1, \dots, p+1.$$

PROPOSITION 2. *Under assumptions* (H1) – (H4) *we have*

$$B_{ij}(x^e) = 0, \qquad j \neq i,\ i \neq p+1, \tag{12}$$

$$L_{\tilde{g}_j^e} \tilde{D}^e \alpha_i(x^e) = 0, \qquad j \neq i,\ i \neq p+1, \tag{13}$$

$$L_{\tilde{g}_j^e} \tilde{D}^e \beta_{ii}(x^e) = 0, \qquad j \neq i,\ i \neq p+1, \tag{14}$$

where D^e is an arbitrary composition of the Lie derivatives $L_{\tilde{f}^e}$ and $L_{\tilde{g}_j^e}$, $j = 1, \dots, p+1$.

Proof. Combine Lemma 1 above with [13, Lem. 1]. □

Now let us consider an invertible feedback law that solves LNSD for (1). By definition, the system

$$\begin{aligned} u &= \alpha(x,w) + \beta(x,w)v, \\ \dot{x} &= f(x) + G(x)\alpha(x,w) + G(x)\beta(x,w)v, \\ \dot{w} &= \gamma(x,w) + \delta(x,w)v, \end{aligned} \tag{15}$$

with (1) given by $\dot{x} = f(x) + G(x)v$, is invertible in the sense of Singh. From [14, Thm. 1] it follows that there exists a dynamic feedback law such that the system, resulting from (15), is noninteractive. In particular, if we assume for the moment that $\dim(\text{span}\{\tilde{g}_i\}(x)) > 1$, from (12) it follows that for each system

$$\begin{aligned} u_i &= \alpha_i(x,w) + \beta_{ii}(x,w)v_i, \\ \dot{x} &= f(x) + G(x)\alpha(x,w) + G(x)\beta(x,w)v, \\ \dot{w} &= \gamma(x,w) + \delta(x,w)v, \end{aligned} \tag{16}$$

there exists an invertible dynamic feedback law such that the system, resulting from plugging this feedback law into (16), is noninteractive. Let us consider the following auxiliary system:

$$\begin{aligned} \dot{\zeta}^i &= \alpha_i(x,w) + \beta_{ii}(x,w)v_i, \\ \dot{x} &= f(x) + G(x)\alpha(x,w) + G(x)\beta(x,w)v, \\ \dot{w} &= \gamma(x,w) + \delta(x,w)v, \\ y^i &= h^i(\zeta^i, x, w) = \zeta^i, \end{aligned} \tag{17}$$

where y^i is a vector of fictitious outputs and let $((\zeta^i)^T, x^T, w^T)^T = x^{ie}$ and $x_0^{ie} = (0,0,0)$. Moreover, if $\dim((\text{span}\{\tilde{g}_i\})(x)) = m_i$ and $\dim((\text{span}\{\tilde{g}_i^e\})(x^e)) = m_i^e$, let

$$\tilde{f}^{ie}(x^{ie}) = \begin{pmatrix}\alpha_i(x,w)\\ \tilde{f}^e(x,w)\end{pmatrix},\ \tilde{g}_j^{ie}(x^{ie}) = \begin{cases}\begin{pmatrix}\beta_{ii}(x,w)\\ \tilde{g}_i^e(x,w)\end{pmatrix} & \text{if} j = i,\\ \begin{pmatrix}0_{m_i \times m_j^e}\\ \tilde{g}_j^e(x,w)\end{pmatrix} & \text{if} j \neq i,\end{cases}$$

$$\tilde{G}^{ie}(x^{ie}) = (\tilde{g}_1^{ie}(x^{ie}) \dots \tilde{g}_{p+1}^{ie}(x^{ie})), \qquad \tilde{\mathcal{G}}^{ie} = \text{span}\{\tilde{G}^{ie}\}.$$

With the above notation, we introduce the following sequence of distributions:

$$\begin{aligned}\mathcal{S}_0^{ie} &= \tilde{\mathcal{G}}^{ie},\\ \mathcal{S}_k^{ie} &= [\tilde{f}^{ie}, \ker d\zeta^i \cap \mathcal{S}_{k-1}^{ie}] + \sum_{j=1}^{p+1} [\tilde{g}_j^{ie}, \ker d\zeta^i \cap \mathcal{S}_{k-1}^{ie}] + \mathcal{S}_{k-1}^{ie}.\end{aligned} \tag{18}$$

Note that if the distributions $\mathcal{S}_k^{ie}$ and $\mathcal{S}_k^{ie} \cap \ker d\zeta^i$ have constant dimension for all k, then there exists $k_i^* \geq 0$ such that $\mathcal{S}_{k_i^*}^{ie} = \mathcal{S}_k^{ie}$ for all $k \geq k_i^*$. Moreover, $\mathcal{S}^{ie*} = \mathcal{S}_{k_i^*}^{ie}$ is the *minimal* distribution that contains $\tilde{\mathcal{G}}^{ie}$ and satisfies $[\tilde{f}^{ie}, \mathcal{S}^{ie*} \cap \ker d\zeta^i] \subset \mathcal{S}^{ie*}$ and $[\tilde{g}_j^{ie}, \mathcal{S}^{ie*} \cap \ker d\zeta^i] \subset \mathcal{S}^{ie*}$, $j = 1, \dots, p+1$ (see also [15]). Each distribution $\mathcal{S}_k^{ie}$ is invariant under static feedback transformations, which are regular at x_0^{ie}.

We now proceed to illustrate how the algorithm (18) can be implemented in practice.

To compute $\mathcal{S}_1^{ie}$, we first must compute $\mathcal{G}^{ie} \cap \ker d\zeta^i$. Let $A_0^i(x^e) = \beta_{ii}(x^e)$. Moreover, assume that $A_0^i(x^e)$ has constant rank s_0^i in a neighbourhood of x_0^e and $\tilde{G}^{ie}(x^{ie})$ has constant rank in a neighbourhood of x_0^{ie} (or, equivalently, that $\mathcal{S}_0^{ie} \cap \ker d\zeta^i$ and $\mathcal{S}_0^{ie}$ have constant dimension in a neighbourhood of x_0^{ie}). Moreover, let l_0^i be the number of columns of $A_0^i(x^e)$. It is always possible to find smooth matrices $A_{0h}^i(x^e)$ and $B_{0h}^i(x^e)$, $h = 1, 2$, such that (after possibly permuting the columns of $A_0^i(x^e)$) we have

$$A_0^i(x^e) = (A_{01}^i(x^e) \quad A_{02}^i(x^e)), \quad A_0^i B_{01}^i(x^e) = A_{01}^i(x^e), \quad A_0^i B_{02}^i(x^e) = 0_{m_i \times (l_0^i - s_0^i)}$$

and $(B_{01}^i(x^e) \quad B_{02}^i(x^e))$ is an invertible matrix for all x^e in a neighbourhood of x_0^e. Moreover, using (12)–(14), it is not hard to show that the matrices $B_{0h}^i(x^e)$, $h = 1, 2$, can be chosen in such a way that

$$L_{\tilde{g}_j^e} \tilde{D}^e B_{0h}^i(x^e) = 0, \qquad j \neq i, \tag{19}$$

where $\tilde{D}^e$ is an arbitrary composition of the Lie derivatives $L_{\tilde{f}^e}$ and $L_{\tilde{g}_j^e}$, $j = 1, \dots, p+1$. Let

$$\begin{aligned}\bar{T}_{01}^i &= (\tilde{g}_1^{ie} \dots \tilde{g}_{i-1}^{ie}\, \tilde{g}_{i+1}^{ie} \dots \tilde{g}_{p+1}^{ie}),\\ \bar{T}_{02}^i &= \tilde{g}_i^{ie} B_{02}^i,\\ T_{02}^i &= (\bar{T}_{01}^i\, \bar{T}_{02}^i),\\ T_{01}^i &= \tilde{g}_i^{ie} B_{01}^i,\\ T_1^i &= (T_{01}^i\, T_{02}^i\, [\tilde{g}_1^{ie}, T_{02}^i] \dots [\tilde{g}_{p+1}^{ie}, T_{02}^i]\, [\tilde{f}^{ie}, T_{02}^i]),\end{aligned}$$

and

$$\begin{aligned}\bar{R}_{02}^i &= \tilde{g}_i^e B_{02}^i,\\ R_{01}^i &= \tilde{g}_i^e B_{01}^i.\end{aligned}$$

The distribution $\tilde{\mathcal{G}}^{ie} \cap \ker\, d\zeta^i$ is spanned by the vector fields T^i_{02} and the distribution $\mathcal{S}^{ie}_1$ is spanned by the vector fields T^i_1.

At the step $(k+1)$th, $k \geq 1$, to compute $\mathcal{S}^{ie}_{k+1}$ we first must compute $\mathcal{S}^{ie}_k \cap \ker\, d\zeta^i$. Let

$$A^i_k(x^e) = (A^i_{k-1,1}(x^e) \quad L_{\bar{R}^i_{k-1,2}}\beta_{ii}(x^e) \quad L_{\bar{R}^i_{k-1,2}}\alpha_i(x^e)).$$

Moreover, assume that $A^i_k(x^e)$ has constant rank s^i_k in a neighbourhood of x^e_0 and $T^i_k(x^{ie})$ has constant rank in a neighbourhood of x^{ie}_0 (or, equivalently, that $\mathcal{S}^{ie}_k \cap \ker\, d\zeta^i$ and $\mathcal{S}^{ie}_k$ have constant dimension in a neighbourhood of x^{ie}_0). In this case, it is possible to find matrices $A^i_{kh}(x^e)$ and $B^i_{kh}(x^e)$, $h = 1, 2$, such that (after possibly permuting the columns of $A^i_k(x^e)$) we have

$$A^i_k(x^e) = (A^i_{k1}(x^e) \quad A^i_{k2}(x^e)), \quad A^i_k B^i_{k1}(x^e) = A^i_{k1}(x^e), \quad A^i_k B^i_{k2}(x^e) = 0_{m_i \times (l^i_k - s^i_k)}$$

and $(B^i_{k1}(x^e) \quad B^i_{k2}(x^e))$ is an invertible matrix for all x^e in a neighbourhood of x^e_0. Moreover, using (12)–(14), it is not hard to show that the matrices $B^i_{kh}(x^e)$, $h = 1, 2$, can be chosen in such a way that

$$L_{\tilde{g}^e_j}\tilde{D}^e B^i_{kh}(x^e) = 0, \qquad j \neq i, \tag{20}$$

where $\tilde{D}^e$ is an arbitrary composition of the Lie derivatives $L_{\tilde{f}^e}$ and $L_{\tilde{g}^e_j}$, $j = 1, \ldots, p+1$. Let

$$\begin{aligned}
\bar{T}^i_{k1} &= (T^i_{k-1,2} \quad [\tilde{g}^{ie}_1, T^i_{k-1,2}] \cdots [\tilde{g}^{ie}_{i-1}, T^i_{k-1,2}] \quad [\tilde{g}^{ie}_i, \bar{T}^i_{k-1,1}] \quad [\tilde{g}^{ie}_{i+1}, T^i_{k-1,2}] \\
&\qquad \cdots [\tilde{g}^{ie}_{p+1}, T^i_{k-1,2}] \quad [\tilde{f}^{ie}, \bar{T}^i_{k-1,1}]), \\
\bar{T}^i_{k2} &= (T^i_{k-1,1} \quad [\tilde{g}^{ie}_i, \bar{T}^i_{k-1,2}] \quad [\tilde{f}^{ie}, \bar{T}^i_{k-1,2}])\, B^i_{k2}, \\
T^i_{k2} &= (\bar{T}^i_{k1} \quad \bar{T}^i_{k2}), \\
T^i_{k1} &= (T^i_{k-1,1} \quad [\tilde{g}^{ie}_i, \bar{T}^i_{k-1,2}] \quad [\tilde{f}^{ie}, \bar{T}^i_{k-1,2}])\, B^i_{k1}, \\
T^i_{k+1} &= (T^i_{k1} \quad T^i_{k2} \quad [\tilde{g}^{ie}_1, T^i_{k2}] \cdots [\tilde{g}^{ie}_{p+1}, T^i_{k2}] \quad [\tilde{f}^{ie}, T^i_{k2}]),
\end{aligned}$$

and

$$\begin{aligned}
\bar{R}^i_{k2} &= (R^i_{k-1,1} \quad [\tilde{g}^e_i, \bar{R}^i_{k-1,2}] \quad [\tilde{f}^e, \bar{R}^i_{k-1,2}])\, B^i_{k2}, \\
R^i_{k1} &= (R^i_{k-1,1} \quad [\tilde{g}^e_i, \bar{R}^i_{k-1,2}] \quad [\tilde{f}^e, \bar{R}^i_{k-1,2}])\, B^i_{k1}.
\end{aligned}$$

The distribution $\mathcal{S}^{ie}_k \cap \ker\, d\zeta^i$ is spanned by the vector fields T^i_{k2}, since, as a direct consequence of (12)–(14), (19), and (20), the only vector fields, which have the first m_i components possibly not identically equal to zero, are of the form T^i_{h1}, $[\tilde{g}^{ie}_i, \bar{T}^i_{h2}]$ or $[\tilde{f}^{ie}, \bar{T}^i_{h2}]$, with $h \leq k-1$, and the first m_i rows of $(T^i_{k-1,1} \quad [\tilde{g}^{ie}_i, \bar{T}^i_{k-1,2}] \quad [\tilde{f}^{ie}, \bar{T}^i_{k-1,2}])\,(x^{ie})$ are identically equal to $A^i_k(x^e)$. On the other hand, the distribution $\mathcal{S}^{ie}_{k+1}$ is spanned by the vector fields T^i_{k+1}. This completes the description of a possible implementation of the algorithm (18). We note that the proposed algorithm (18) gives a more general and intuitive version of the one given in [13].

In what follows, we assume that the distributions $\mathcal{S}^{ie}_k$ and $\mathcal{S}^{ie}_k \cap \ker\, d\zeta^i$, $k \geq 0$, have constant dimension in a neighbourhood of x^{ie}_0. In this case there exists indeed $k^*_i \geq 0$ such that $\mathcal{S}^{ie}_{k^*_i} = \mathcal{S}^{ie}_{k^*_i+1} = \mathcal{S}^{ie*}$ and we will say that $\mathcal{S}^{ie}$ is *regularly computable* at x^{ie}_0. The fact that (17) can be rendered noninteractive via dynamic feedback law has some interesting consequences in terms of $\mathcal{S}^{ie*}$. This is shown in the next lemma.

LEMMA 3. *Assume that $\mathcal{S}^{ie*}$ is regularly computable at x_0^{ie}. Then, $\mathcal{S}^{ie*}$* + ker $d\zeta^i = \mathbb{R}^{m_i+n+n^w}$ *for all x^{ie} in a neighbourhood of x_0^{ie}.*

Proof. Suppose that

$$\dim\left((\mathcal{S}^{ie*} + \ker\, d\zeta^i)(x_0^{ie})\right) < m_i + n + n^w. \tag{21}$$

We will show that this gives a contradiction. Assume first that $\dim(\text{span}\{\tilde{g}_i\}(x)) = 1$. In this case, reasoning as in [7], we can easily prove that each system (16) has relative degree at x_0^e (with respect to u_i). Without loss of generality, we can assume that (16) is in *normal form* (see [1] for a definition). From here it is straightforward to check that $\mathcal{S}^{ie*}$ +ker $d\zeta^i = \mathbb{R}^{m_i+n+n^w}$.

On the other hand, assume that $\dim(\text{span}\{\tilde{g}_i\}(x)) > 1$. Following [14], from our assumptions it follows that there exists a feedback law

$$\begin{aligned} v_i &= \alpha^i(x^{ie}, \eta^i) + \beta^i(x^{ie}, \eta^i)v', \\ \dot{\eta}^i &= \gamma^i(x^{ie}, \eta^i) + \delta^i(x^{ie}, \eta^i)v', \end{aligned} \tag{22}$$

such that the input-output behaviour of the system, obtained from plugging (22) into (17) and denoted in what follows by $\bar{\Sigma}^e$, is described by

$$(\zeta_k^i)^{(l_k)} = v'_{j_k}, \qquad k = 1, \ldots, m_i, \tag{23}$$

where the subscript (l_k) denotes the number of derivatives with respect to time, ζ_k^i is the kth component of the output vector ζ^i, v'_{j_k} is the j_kth component of the input vector v', $l_k \geq 0$ and $j_k \in \{1, \ldots, m\}$ with $j_r \neq j_s$ for r, $s = 1, \ldots, m_i$ and $r \neq s$. Let

$$\bar{f}^{ie} = \begin{pmatrix} f^{ie}(x^{ie}) + \tilde{G}^{ie}(x^{ie})\alpha^i(x^{ie}, \eta^i) \\ \gamma^i(x^{ie}, \eta^i) \end{pmatrix},$$

$$\bar{G}^{ie}(x^{ie}, \eta^i) = \begin{pmatrix} \tilde{G}^{ie}(x^{ie})\beta^i(x^{ie}, \eta^i) \\ \delta^i(x^{ie}, \eta^i) \end{pmatrix}, \qquad \bar{\mathcal{G}}^{ie} = \text{span}\{\bar{G}^{ie}\},$$

and $\bar{g}_j^{ie}(x^{ie}, \eta^i)$ be the jth column (block) of $\bar{G}^{ie}(x^{ie}, \eta^i)$. From (23), it follows that, after possibly changing coordinates, $\bar{\Sigma}^e$ is in the form

$$\begin{aligned} \dot{\theta}_{k1}^i &= \theta_{k2}^i, \\ \cdots &= \cdots \\ \dot{\theta}_{kl_k}^i &= v'_{j_k}, \qquad k = 1, \ldots, m_i, \\ \dot{\varphi}^i &= \psi^i(\theta_{k1}^i, \ldots, \theta_{kl_1}^i, \ldots, \theta_{m_i1}^i, \ldots, \theta_{m_il_{m_i}}^i, \varphi^i) \\ &\quad + \lambda^i(\theta_{k1}^i, \ldots, \theta_{kl_1}^i, \ldots, \theta_{m_i1}^i, \ldots, \theta_{m_il_{m_i}}^i, \varphi_i)v', \end{aligned}$$

for some (possibly vector-valued) smooth functions ψ^i and λ^i and with output vector h^i equal to $(\theta_{11}^i, \ldots, \theta_{m_i1}^i)^T$. Now, denote by $\bar{\mathcal{S}}^{ie*}$ the minimal distribution that contains $\bar{\mathcal{G}}^{ie}$ and is such that $[\bar{f}^{ie}, \bar{\mathcal{S}}^{ie*} \cap \ker\, dh^i] \subset \bar{\mathcal{S}}^{ie*}$ and $[\bar{g}_j^{ie}, \bar{\mathcal{S}}^{ie*} \cap \ker\, dh^i] \subset \bar{\mathcal{S}}^{ie*}$ for $j = 1, \ldots, p+1$, and assume for simplicity that $\bar{\mathcal{S}}^{ie*}$ is regularly computable at the origin of the state space (the proof of the lemma can be repeated also after removing this assumption). By straightforward computations, it can be seen that the distribution $\bar{\mathcal{S}}^{ie*}$ is such that

$$\langle dh^i, \bar{\mathcal{S}}^{ie*}\rangle(x_0^{ie}) = \mathbb{R}^{m_i} \tag{24}$$

($\langle dh^i, \bar{\mathcal{S}}^{ie*}\rangle(x_0^{ie})$ denotes the subspace of $\mathbb{R}^{m_i}$ spanned by the vectors $\langle dh^i, \tau^i\rangle(x_0^{ie})$, with $\tau^i \in \bar{\mathcal{S}}^{ie*}$). On the other hand, it is not hard to show by induction that, if π is the natural projection

$\pi((x^{ie}, \eta^i)) = x^{ie}$ and $\pi_{*(x^{ie},\eta^i)}$ its differential at (x^{ie}, η^i), we have $\pi_{(x^{ie},\eta^i)*}\bar{\mathcal{S}}^{ie*}(x^{ie}, \eta^i) \subset \mathcal{S}^{ie*}(\pi(x_i^e, \eta^i))$. This, together with (24), would contradict (23). □

We are ready now to state the main result of this section. Let

$$\mathcal{R}_i^e = \langle \tilde{f}^e, \tilde{g}_1^e, \dots, \tilde{g}_{p+1}^e | \text{span}\{\tilde{g}_i^e, \tilde{g}_{p+1}^e\}\rangle, \qquad i = 1, \dots, p,$$
$$\mathcal{Q}^e = \langle \tilde{f}^e, \tilde{g}_1^e, \dots, \tilde{g}_{p+1}^e | \text{span}\{\tilde{g}_{p+1}^e\}\rangle.$$

Moreover, let Δ_{MIX}^e be the distribution defined in the same way as in (7), but with $\tilde{f}$ and $\tilde{g}_j$, $j = 1, \dots, p+1$, replaced by $\tilde{f}^e$ and $\tilde{g}_j^e$, respectively.

THEOREM 4. *Assume* (H1)–(H5) *hold. Moreover, assume that the distributions* $\mathcal{S}^{ie*}$, $i = 1, \dots, p$, *are regularly computable at* x_0^{ie}, *the distributions* $\mathcal{R}_i^e$, $i = 1, \dots, p$, *and* $\mathcal{Q}^e$ *are regularly computable at* x_0^e, *and that the distributions* $\Delta_{MIX}^e + \mathcal{R}_i^e$, $i = 1, \dots, p$, *and* $\Delta_{MIX}^e + \mathcal{Q}^e$ *have constant dimension in a neighbourhood of* x_0^e. *If* LNSD *is solvable via invertible feedback laws, then*

(a) *the dynamics* Σ_0^* *is locally asymptotically stable at the origin,*

(b) *each system* Σ_i^*, $i = 1, \dots, p$, *is locally asymptotically stabilizable via dynamic feedback laws.*

Conversely, if (a) *and* (b) *are satisfied and, in addition,* Σ_{p+1}^* *is locally asymptotically stabilizable via dynamic feedback laws, then* LNSD *is solvable via invertible feedback laws.*

Remark 5. Conditions (a) and (b) require us to check the local asymptotic stability and, respectively, the local asymptotic stabilizability of a suitable nonlinear system of the form $\dot{x} = f(x) + G(x)u$. For stability the reader is referred to the standard text [16]. On the other hand, many interesting results are available to check asymptotic stabilizability via *static* state feedback. A first kind of test is given in terms of the Jacobian $\frac{\partial f}{\partial x}(x)$ of $f(x)$, computed at $x = 0$. If the pair $\left(\frac{\partial f}{\partial x}(0), G(0)\right)$ is stabilizable (in the sense that the uncontrollable modes are asymptotically stable) then the nonlinear system can be locally asymptotically stabilized through a *linear* state feedback law ([18]). If some eigenvalues of $\frac{\partial f}{\partial x}(0)$ are not controllable and have zero real part (*critical case*), the theory of center manifold can be used in some simple cases to design stabilizing controllers ([17]).

An important class of nonlinear system, which can always be locally asymptotically stabilized through *static* state feedback, is given by the systems that have *relative degree* and are *minimum phase* ([18]–[20]).

On the other hand, it has recently been shown in [21] that a nonlinear system of the above form can be asymptotically stabilized via smooth static state feedback if and only if there exist a smooth function $\varphi(x)$ such that the system is feedback equivalent to a *strictly passive* (with respect to φ), or, equivalently, to a system that has *relative degree* $\{1, \dots, 1\}$ and is *minimum phase*.

Interesting results on state feedback stabilization have been also given in terms of *control Lyapunov functions* ([22]–[29]). Necessary and sufficient conditions for state feedback stabilization are available for two- and three-dimensional analytic systems ([30], [31]). Moreover, stabilization via *dynamic* state feedback is considered in [29].

Proof of Theorem 4. The sufficiency part can be proved as in [10].

The necessity can be proved as follows. Let π be the natural projection $\pi(x^e) = x$ and π_{*x^e} be its differential at x^e. Moreover, let $\mathcal{Q}^*$ be thought of as a distribution of $\mathbb{R}^{n+n^w}$

Now let B_{kj}^i, $\bar{R}_{k2}^i$, R_{k1}^i, and s_k^i be as above and, if $q = \dim(\mathcal{Q}^*(x^e))$, denote by $(A \bmod(\mathcal{Q}^* + \mathcal{G}^w))(x^e)$ the first $n - q$ components (respectively, rows) of a given vector (respectively, matrix) $A(x^e)$. From (12)–(14), (19), and (20) it follows that

$$([\tilde{f}^e, \bar{R}_{k2}^i] \bmod(\mathcal{Q}^* + \mathcal{G}^w))(x^e) = \tilde{g}_i(x)(L_{\bar{R}_{k2}^i}\alpha_i)(x^e),$$

$$([\tilde{g}_i^e, \bar{R}_{k2}^i] \operatorname{mod}(\mathcal{Q}^* + \mathcal{G}^w))(x^e) = \tilde{g}_i(x)(L_{\bar{R}_{k2}^i}\beta_{ii})(x^e).$$

Lemma 3 implies that the first m_i rows of $T_{k_i^*}^i(x^{ie})$ are linearly independent for all x^{ie} in a neighbourhood of x_0^{ie}. Since the first m_i rows of $(T_{k-1,1}^i \quad [\tilde{g}_i^{ie}, \bar{T}_{k-1,2}^i] \quad [\tilde{f}^{ie}, \bar{T}_{k-1,2}^i])\,(x^{ie})$ are identically equal to $A_k^i(x^e)$ for all k, the matrix $A_{k_i^*,1}^i(x^e)$ has full row rank for all x^e in a neighbourhood of x_0^e and $(R_{k_i^*,1}^i \operatorname{mod}(\mathcal{Q}^* + \mathcal{G}^w))(x^e) = \tilde{g}_i(x)A_{k_i^*,1}^i(x^e)$. Since $\mathcal{R}_i^e$ is invariant under $\tilde{f}^e, \tilde{g}_1^e, \ldots, \tilde{g}_{p+1}^e$ and from (19) and (20), it follows that $[\tilde{g}_i^e, \bar{R}_{k2}^i] \in \mathcal{R}_i^e$ and $[\tilde{f}^e, \bar{R}_{k2}^i] \in \mathcal{R}_i^e$. We conclude that $\tilde{g}_i(\pi(x^e)) \in \pi_{*x^e}(\mathcal{R}_i^e + \mathcal{Q}^*)(x^e)$.

From here, we can proceed as in [13, Thm. 1]. For completeness, we give a short sketch of the proof. Reasoning as in [7] and using the above facts, we prove that $\pi_{*x^e}(\Delta_{MIX}^e + \mathcal{Q}^e + \mathcal{Q}^*)(x^e) \supset (\Delta_{MIX}^* + \mathcal{Q}^*)(\pi(x^e))$. Since from (12)–(14) it easily follows that $\pi_{*x^e}(\Delta_{MIX}^e + \mathcal{Q}^e + \mathcal{Q}^*)(x^e) \subset (\Delta_{MIX}^* + \mathcal{Q}^*)(\pi(x^e))$, we conclude that $\pi_{*x^e}(\Delta_{MIX}^e + \mathcal{Q}^e + \mathcal{Q}^*)(x^e) = (\Delta_{MIX}^* + \mathcal{Q}^*)(\pi(x^e))$. Since $\dot{x}^e = \tilde{f}^e(x^e)$ is locally asymptotically stable in x_0^e, it is such also on the invariant submanifold $\mathcal{L}_{x_0^e}^{\Delta_{MIX}^e + \mathcal{Q}^e}$. Moreover, $\pi_{*x^e}(\tilde{f}^e|\mathcal{L}_{x_0^e}^{\Delta_{MIX}^e + \mathcal{Q}^e})(x^e) \in (\Delta_{MIX}^* + \mathcal{Q}^*)(\pi(x^e))$ and from (13) it follows that $\alpha(x^e) = 0$ for $x^e \in \mathcal{L}_{x_0^e}^{\Delta_{MIX}^e + \mathcal{Q}^e}$. This, together with $\pi_{*x^e}(\Delta_{MIX}^e + \mathcal{Q}^e + \mathcal{Q}^*)(x^e) = (\Delta_{MIX}^* + \mathcal{Q}^*)(\pi(x^e))$, implies that Σ_0^* is an invariant subdynamics of $\dot{x}^e = \tilde{f}^e(x^e)$ for $x^e \in \mathcal{L}_{x_0^e}^{\Delta_{MIX}^e + \mathcal{Q}^e}$, which proves (a). Similarly, using the fact that $\alpha_j(x^e) = 0$ for $x^e \in \mathcal{L}_{x_0^e}^{\mathcal{R}_i^e + \Delta_{MIX}^e}$ and $j \neq \{i, p+1\}$ (as a consequence of (13)), we prove that $\pi_{*x^e}(\mathcal{R}_i^e + \Delta_{MIX}^e + \mathcal{Q}^*)(x^e) = (\mathcal{R}_i^* + \Delta_{MIX}^*)(\pi(x^e))$ and, as a consequence, (b). □

Theorem 4 generalizes [7] in the case $m = p$, since any *regular noninteraction* (in the sense of [7]) feedback law, which solves LND, is also *invertible*. Indeed, using the results contained in [7], it can be shown that the system (3), corresponding to a given regular noninteraction feedback law, has vector relative degree at x_0^e (with respect to u). Moreover, Lemmas 1 and 3 and Proposition 2 show that in the case of nonlinear systems (1) any invertible feedback law, which solves LND, is also *regular* in the sense of [13].

It is possible to construct some counterexamples, which show that the dynamics Σ_{p+1}^* is not necessarily locally asymptotically stabilizable at the origin if LNSD is solvable (see [13]). On the other hand, if $\mathcal{Q}^* = 0$, similarly to [13], we obtain the following necessary and sufficient condition.

COROLLARY 6. *Assume* (H1)–(H5) *hold. Moreover, suppose that* $\mathcal{Q}^* = 0$. *If*

(a) *the dynamics* Σ_0^* *is locally asymptotically stable at the origin,*

(b) *each system* Σ_i^*, $i = 1, \ldots, p$, *is locally asymptotically stabilizable at the origin via dynamic feedback laws,*

then LNSD *is solvable via invertible dynamic feedback laws. Conversely, assume that* $\mathcal{S}^{ie*}$ *is regularly computable at* x_0^{ie}, *the distributions* $\mathcal{R}_i^e$, $i = 1, \ldots, p$, *and* $\mathcal{Q}^e$ *are regularly computable at* x_0^e *and the distributions* $\mathcal{R}_i^e + \Delta_{MIX}^e$, $i = 1, \ldots, p$, *and* $\Delta_{MIX}^e + \mathcal{Q}^e$ *have constant dimension in a neighbourhood of* x_0^e. *Then, if* LNSD *is solvable,* (a) *and* (b) *are satisfied.*

4. Some existence conditions. In this section we prove some conditions that must be satisfied for the existence of a dynamic feedback law that solves LNSD for the class of nonlinear systems (1) satisfying assumptions (H1)–(H4). We *do not* constrain the class of feedback laws and this is quite a new result in the literature. The following theorem gives an interesting existence condition, which, however, cannot be checked a priori. It simply assesses, for each $i = 1, \ldots, p$, the existence of a smooth submanifold, transversal to $h_i^{-1}(0)$, that can be rendered (locally) an asymptotically stable manifold for Σ_i^*. This manifold depends, in general, on the decoupling feedback law, except when, for example, $\pi_{*(x,w)}(\mathcal{R}_i^e + \Delta_{MIX}^e)(x,w) = (\mathcal{R}_i^* + \Delta_{MIX}^*)(\pi(x,w))$ (this is the case when considering invertible feedback laws).

THEOREM 7. *Assume* (H1)–(H4) *hold. If* LNSD *is solvable then for each system* Σ_i^*, *there exist a dynamic feedback law* $u_i = \alpha^i(z_1^i, z_3^i, w^i)$, $\dot{w}^i = \delta^i(z_1^i, z_3^i, w^i)$, $w^i \in \mathbb{R}^{n^i}$, *and, correspondingly, a smooth invariant submanifold* $\mathcal{M}_i = \{(z_1^i, z_3^i, w^i) \in \mathbb{R}^{n+n^i} \colon \varphi^i(z_1^i, z_3^i, w^i) = 0\}$, *passing through the origin of the state space and transversal to the smooth submanifold* $\{(z_1^i, z_3^i, w^i) \in \mathbb{R}^{n+n^i} : h_i(z_1^i) = 0\}$, *such that the system, resulting from* Σ_i^* *with* $u_i = \alpha^i(z_1^i, z_3^i, w^i)$, $\dot{w}^i = \delta^i(z_1^i, z_3^i, w^i)$, *is locally asymptotically stable at the origin on* $\mathcal{M}^i$.

Proof. Using (12)–(14), it is easy to prove that

$$\pi_{*x^e}(\mathcal{R}_i^e + \Delta_{MIX}^e)(x^e) \subset (\mathcal{R}_i^* + \Delta_{MIX}^*)(\pi(x^e)), \tag{25}$$

where π_{*x^e} is defined as above. Since LNSD is solvable, it follows that $\dot{x}^e = \tilde{f}^e(x^e)$ is locally asymptotically stable in x_0^e. In particular, it is such for initial conditions lying on $\mathcal{L}_{x_0^e}^{\mathcal{R}_i^e + \Delta_{MIX}^e}$. From (25) it follows that $\pi_{*x^e}(\tilde{f}^e|\mathcal{L}_{x_0^e}^{\mathcal{R}_i^e + \Delta_{MIX}^e})(x^e) \subset (\mathcal{R}_i^* + \Delta_{MIX}^*)(\pi(x^e))$. This, together with the fact that $\tilde{f}^e(x_0^e) = 0$, implies that, in z^i-coordinates, the component of $\tilde{f}^e|\mathcal{L}_{x_0^e}^{\mathcal{R}_i^e + \Delta_{MIX}^e}$ along $\operatorname{span}\{\partial/\partial x_j, \partial/\partial z_2^i : j \neq \{i, p+1, p+2\}\}$ is identically zero. The same holds for the restrictions $\tilde{g}_i^e|\mathcal{L}_{x_0^e}^{\mathcal{R}_i^e + \Delta_{MIX}^e}$ and $\tilde{g}_{p+1}^e|\mathcal{L}_{x_0^e}^{\mathcal{R}_i^e + \Delta_{MIX}^e}$, since $\tilde{g}_i^e, \tilde{g}_{p+1}^e \in \mathcal{R}_i^e$. Moreover, from (12)–(14) it follows that $\alpha_j(x^e) = 0$ for all $x^e \in \mathcal{L}_{x_0^e}^{\mathcal{R}_i^e + \Delta_{MIX}^e}$ and $j \neq \{i, p+1\}$. The dynamics, described on $\mathcal{L}_{x_0^e}^{\mathcal{R}_i^e + \Delta_{MIX}^e}$ by the vector fields $\tilde{f}^e|\mathcal{L}_{x_0^e}^{\mathcal{R}_i^e + \Delta_{MIX}^e}$, $\tilde{g}_i^e|\mathcal{L}_{x_0^e}^{\mathcal{R}_i^e + \Delta_{MIX}^e}$ and $\tilde{g}_{p+1}^e|\mathcal{L}_{x_0^e}^{\mathcal{R}_i^e + \Delta_{MIX}^e}$, is clearly an invariant subdynamics of

$$\begin{aligned} \begin{pmatrix} \dot{z}_1^i \\ \dot{z}_3^i \\ \dot{z}_4^i \\ \dot{z}_5^i \\ \dot{w} \end{pmatrix} &= \begin{pmatrix} \tilde{f}_{z_1^i}(0, \ldots, 0, z_1^i) \\ \tilde{f}_{z_3^i}(0, \ldots, 0, z_1^i, 0, z_3^i) \\ \tilde{f}_{z_4^i}(0, \ldots, 0, z_1^i, 0, z_3^i, z_4^i) \\ \tilde{f}_{z_5^i}(0, \ldots, 0, z_1^i, 0, z_3^i, z_4^i, z_5^i) \\ \delta(0, \ldots, 0, z_1^i, 0, z_3^i, z_4^i, z_5^i, w) \end{pmatrix} \\ &\quad + \sum_{j \in \{i, p+1\}} \begin{pmatrix} \tilde{g}_{z_1^i, j}(0, \ldots, 0, z_1^i) \\ \tilde{g}_{z_3^i, j}(0, \ldots, 0, z_1^i, 0, z_3^i) \\ \tilde{g}_{z_4^i, j}(0, \ldots, 0, z_1^i, 0, z_3^i, z_4^i) \\ \tilde{g}_{z_5^i, j}(0, \ldots, 0, z_1^i, 0, z_3^i, z_4^i, z_5^i) \\ 0 \end{pmatrix} \alpha_j(0, \ldots, 0, z_1^i, 0, z_3^i, z_4^i, z_5^i, w). \end{aligned} \tag{26}$$

If we let $\mathcal{M}^i = \mathcal{L}_{x_0^e}^{\mathcal{R}_i^e + \Delta_{MIX}^e}$, since (26) is locally asymptotically stable at the origin on $\mathcal{M}^i$, it follows that there exist a dynamic feedback law $u_i = \alpha^i(z_1^i, z_3^i, w^i)$, $\dot{w}^i = \delta^i(z_1^i, z_3^i, w^i)$, $w^i \in \mathbb{R}^{n^i}$, and, correspondingly, a smooth invariant submanifold $\mathcal{M}^i = \{(z_1^i, z_3^i, w^i) \in \mathbb{R}^{n+n^i} : \varphi^i(z_1^i, z_3^i, w^i) = 0\}$, passing through the origin of the state space, such that the system, resulting from Σ_i^* with $u_i = \alpha^i(z_1^i, z_3^i, w^i)$, $\dot{w}^i = \delta^i(z_1^i, z_3^i, w^i)$, is locally asymptotically stable at the origin on $\mathcal{M}^i$. Moreover, it is not hard to show that in a neighbourhood of x_0^e

$$\mathcal{R}_i^e + \ker dh_i^e = \mathbb{R}^{n+n^w}. \tag{27}$$

Indeed, by definition of LNSD, the system $\tilde{\Sigma}^e$ has vector relative degree at x_0^e and it is noninteractive. From [1, Lem. 3.3.1] and the definition of vector relative degree, it follows that for each $i \in \{1, \ldots, p\}$ there exists r_i^e such that $L_{\tilde{g}_i^e} L_{\tilde{f}^e}^k h_i^e(x^e) = 0$ for all

$k < r_i^e - 1$ and $L_{\tilde{g}_i^e} L_{\tilde{f}^e}^{r_i^e - 1} h_i^e(x_0^e) \neq 0$. Equivalently, if we define $ad_{\tilde{f}^e}^k \tilde{g}_i^e$ as $ad_{\tilde{f}^e}^0 \tilde{g}_i^e = \tilde{g}_i^e$ and $ad_{\tilde{f}^e}^k \tilde{g}_i^e = [\tilde{f}^e, ad_{\tilde{f}^e}^{k-1} \tilde{g}_i^e]$ for all $k \geq 1$, we have $L_{ad_{\tilde{f}^e}^k \tilde{g}_i^e} h_i^e(x^e) = 0$ for all $k < r_i^e - 1$ and $L_{ad_{\tilde{f}^e}^{r_i^e - 1} \tilde{g}_i^e} h_i^e(x_0^e) \neq 0$. Since by construction $ad_{\tilde{f}^e}^{r_i^e - 1} \tilde{g}_i^e \in \mathcal{R}_i^e$, we obtain (27). This proves that $\mathcal{M}^i$ is transversal to the smooth submanifold $\{(z_1^i, z_3^i, w^i) \in \mathbb{R}^{n+n^i} : h_i(z_1^i) = 0\}$. □

Remark 8. If we consider invertible feedback laws, we obtain condition (b) of Theorem 4, since in this case it can be shown that $\pi_{*x^e}(\mathcal{R}_i^e + \Delta_{MIX}^e + \mathcal{Q}^*)(x^e) = (\mathcal{R}_i^* + \Delta_{MIX}^*)(\pi(x^e))$.

In §3 we remarked that the dynamics Σ_{p+1}^* is not necessarily locally asymptotically stabilizable at the origin if LNSD is solvable. This fact accounts for an apparent gap between necessary and sufficient conditions (Theorem 4). We might ask if something more can be said by considering the dynamics (8)–(10), associated to *some* regular solution, which is not necessarily the *maximal* one. A positive fact is that if the dynamics (8) is locally asymptotically stable at the origin, each system (9) is locally asymptotically stabilizable at the origin via dynamic feedback law and, in addition, (10) is locally asymptotically stabilizable at the origin via dynamic feedback law, then LNSD is solvable. This can be proved exactly as in the case we consider the maximal regular solution $\mathcal{R}_1^*, \ldots, \mathcal{R}_p^*$. Are the above conditions also necessary to solve LNSD? Toward partially answering this question, we first give the following definition.

DEFINITION 9. *An invertible feedback law* (2) *is said to be a* noninteraction *feedback law if there exists a block partition* $J_1, \ldots, J_{p+1}$ *of* $\{1, \ldots, m\}$, *with* $J_i \cap J_h = \{\phi\}$ *for* $h \neq i$, *such that* (2) *is of the form*

$$\begin{aligned} u_{J_i} &= \alpha_{J_i}(x, w) + \beta_{J_i, i}(x, w) v_i, \qquad i = 1, \ldots, p, \\ u_{J_{p+1}} &= \alpha_{J_{p+1}}(x, w) + \sum_{j=1}^{p+1} \beta_{J_{p+1}, j}(x, w) v_j, \\ \dot{w} &= \delta(x, w) + \gamma(x, w) v, \end{aligned}$$

and

$$\begin{aligned} L_{\tilde{g}_j^e} \tilde{D}^e \alpha_{J_i}(x^e) &= 0, \qquad j \neq i,\ i \neq p+1, \\ L_{\tilde{g}_j^e} \tilde{D}^e \beta_{J_i, i}(x^e) &= 0, \qquad j \neq i,\ i \neq p+1, \end{aligned}$$

where $\tilde{D}^e$ is an arbitrary composition of the Lie derivatives $L_{\tilde{f}^e}$ and $L_{\tilde{g}_j^e}$, $j = 1, \ldots, p+1$. □

The definition of noninteraction feedback law was first given in [7] in the case $m = p$.

We have seen that any invertible feedback law, which solves LND for (1), can be expressed as the cascade of a static feedback law, *regular* at x_0, together with an *invertible noninteraction* feedback law. In the case of linear systems, the proof of this fact is extremely simple and goes as follows. Let $u = Fx^e + \sum_{j=1}^{p+1} H_j v_j$, $\dot{w} = Kx^e + \sum_{j=1}^{p+1} D_j v_j$, be a feedback law that achieves noninteraction. By reasoning as in [16], it can be shown that $\pi\mathcal{R}_1^e, \ldots, \pi\mathcal{R}_p^e$ and $\pi\mathcal{Q}^e$ are controllability subspaces and that the set $\pi\mathcal{R}_1^e, \ldots, \pi\mathcal{R}_p^e$ is a regular solution. After possibly a regular static feedback transformation, we can assume that there exists a block partition $J_1, \ldots, J_{p+1}$ of $\{1, \ldots, m\}$, with $J_i \cap J_h = \{\phi\}$ for $h \neq i$, such that $\pi\mathcal{R}_i^e = \langle \tilde{A} | \text{span}\{\tilde{B}_{J_i}, \tilde{B}_{J_{p+1}}\} \rangle$ and $\pi\mathcal{R}_i^e \cap \mathcal{G} = \text{span}\{\tilde{B}_{J_i}, \tilde{B}_{J_{p+1}}\}$ for $i = 1, \ldots, p$ (here we replaced $\tilde{f}(x)$ by $\tilde{A}x$ and $\tilde{g}_{J_i}(x^e)$ by $\tilde{B}_{J_i}$). Now, according to the partition $J_1, \ldots, J_{p+1}$, let us partition H_i as $(H_{J_1,i}^T \ldots H_{J_{p+1},i}^T)^T$ and F as $(F_{J_1}^T \ldots F_{J_{p+1}}^T)^T$. For any $X_i^e \in \mathcal{R}_i^e$, we have $\pi(\tilde{A}^e X_i^e) = \tilde{A}\pi(X_i^e) + \sum_{l=1}^{p+1} \tilde{B}_{J_l} F_{J_l} \pi(X_i^e)$ (we replaced $\tilde{f}^e(x^e)$ by $\tilde{A}^e x^e$). Since $\tilde{A}\pi(X_i^e) \in \pi\mathcal{R}_i^e$ and $\tilde{B}_{J_l} \in \pi\mathcal{R}_l^e$, we conclude that $F_{J_l}\pi(X_i^e) = 0$ for $l \neq \{i, p+1\}$. Moreover, since $\tilde{B}H_i \in \pi\mathcal{R}_i^e$

and by construction $\text{span}\{\tilde{B}_{J_h}\} \cap \pi\mathcal{R}_i^e = 0$ for $h \neq \{i, p+1\}$ (here $\tilde{G}(x)$ has been replaced by $\tilde{B}$), it follows that $H_{J_i,l} = 0$ for $l \neq i$ and $i \neq p+1$. This proves our thesis.

The subspace $\pi\mathcal{Q}^e$ is a controllability subspace contained in $\bigcap_{i=1}^p \sum_{j\neq i} \pi\mathcal{R}_j^e$, but it is not the *maximal* one with such property. Indeed, if we take the subspaces $\mathcal{R}_1^e, \ldots, \mathcal{R}_p^e$ to be linearly independent and such that $\pi\mathcal{R}_i^e = \mathcal{R}_i^*$ (this is always possible by the construction given in [32]), we have $\pi\mathcal{Q}^e = 0$ but the maximal controllability subspace contained in $\bigcap_{i=1}^p \sum_{j\neq i} \pi\mathcal{R}_j^e$ is $\mathcal{Q}^*$. This fact suggests the following approach to the case of nonlinear systems. Consider the class of feedback laws that can be expressed as the cascade of a static feedback law, regular at x_0, together with an invertible noninteraction feedback law, with the following additional property: for each system

$$
\begin{aligned}
u_{J_i} &= \alpha_{J_i}(x,w) + \beta_{J_i,i}(x,w)v_i,\\
u_{J_{p+1}} &= \alpha_{J_{p+1}}(x,w) + \sum_{l\in\{i,p+1\}} \beta_{J_{p+1},l}(x,w)v_l,\\
\dot{w} &= \delta(x,w) + \sum_{l\in\{i,p+1\}} \gamma_l(x,w)v_l,
\end{aligned}
\tag{28}
$$

and

$$
\begin{aligned}
u_{J_{p+1}} &= \alpha_{J_{p+1}}(x,w) + \beta_{J_{p+1},p+1}(x,w)v_{p+1},\\
\dot{w} &= \delta(x,w) + \gamma_{p+1}(x,w)v_{p+1},
\end{aligned}
\tag{29}
$$

where $\gamma_i(x,w)$ is the ith column block of $\gamma(x,w)$ corresponding to the partition $\tilde{g}_1^e, \ldots, \tilde{g}_{p+1}^e$, LND is solvable. This property corresponds, in the case of linear systems, to the fact that each system (28) and (29) has a transfer function matrix with rank equal to the number of its rows ([33]). Using the results and proofs of §3, it is not hard to show that, if LNSD is solvable via invertible feedback laws of the above class, we have

$$
\begin{aligned}
\pi_{*x^e}\mathcal{R}_i^e(x^e) &= \langle \tilde{f}, \tilde{g}_{J_1}, \ldots, \tilde{g}_{J_{p+1}} | \text{span}\{\tilde{g}_{J_i}, \tilde{g}_{J_{p+1}}\}\rangle(\pi(x^e)), \qquad i = 1, \ldots, p,\\
\pi_{*x^e}\mathcal{Q}^e(x^e) &= \langle \tilde{f}, \tilde{g}_{J_1}, \ldots, \tilde{g}_{J_{p+1}} | \text{span}\{\tilde{g}_{J_{p+1}}\}\rangle(\pi(x^e)).
\end{aligned}
$$

Moreover, $\pi_{*x^e}\mathcal{R}_1^e, \ldots, \pi_{*x^e}\mathcal{R}_p^e$ is a regular solution. Now, let us consider the dynamics (8)–(10), as defined in §2 with $\mathcal{R}_i = \pi_{*x^e}\mathcal{R}_i^e$ and with $\mathcal{Q}$ replaced by $\pi_{*x^e}\mathcal{Q}^e$, which, in general, is *not* the maximal controllability distribution contained in $\bigcap_{i=1}^p \sum_{j\neq i} \pi_{*x^e}\mathcal{R}_j^e$ (for this reason, the dynamics (8)–(10) in this case may depend on the feedback law, used to obtain (6) from (1)). Using the same arguments of Theorem 4, it can be easily shown that, under suitable regularity assumptions, the dynamics (8) is locally asymptotically stable at the origin and each system (9) and (10) is locally asymptotically stabilizable at the origin via dynamic feedback law. However, the problem of finding out if these conditions are also necessary to solve LNSD (or a stronger version of it) via *invertible* feedback laws is still an open question.

Conclusions. For the class of nonlinear systems (1), for which *noninteraction* can be achieved by means of static state feedback laws, regular at x_0, it results that the asymptotic stability of Σ_0^* and the asymptotic stabilizability via dynamic state feedback of the systems Σ_i^*, $i = 1, \ldots, p$, are necessary to achieve noninteraction and stability via invertible feedback laws. These conditions, plus the asymptotic stabilizability of Σ_{p+1}^* via dynamic feedback law, are also sufficient to solve LNSD via invertible feedback laws. Unfortunately, there is an apparent gap between necessary and sufficient conditions, given by the distribution $\mathcal{Q}^*$. In the case $m = p$, this is not an issue, since $\mathcal{Q}^* = 0$. Moreover, in the case of linear systems, Σ_{p+1}^* is always asymptotically stabilizable by definition.

Some interesting existence conditions have also been given. One result, among others, is that if noninteraction and stability can be achieved via dynamic feedback laws, which are possibly noninvertible, for each $i \in \{1, \ldots, p\}$ there must exist a dynamic feedback law that locally asymptotically stabilizes Σ_i^* on a suitable invariant submanifold $\mathcal{M}^i$ of the state space of the system, resulting from Σ_i^* after applying this feedback law. If we consider invertible feedback laws, we obtain a necessary condition of Theorem 4.

Acknowledgments. The author thanks Prof. A. Isidori for his constant encouragement.

REFERENCES

[1] A. Isidori, *Nonlinear Control Systems*, 2nd ed., Springer Verlag, New York, Berlin, 1989.

[2] S. N. Singh, *A modified algorithm for invertibility in nonlinear systems*, IEEE Trans. Automat. Control, 26 (1981), pp. 595–598.

[3] M. D. Di Benedetto, J. W. Grizzle, and C. H. Moog, *Rank invariants for nonlinear systems*, SIAM J. Control Optim., 27 (1989), pp. 658–672.

[4] W. M. Wonham and S. Morse, *Decoupling and pole assignment by dynamic compensation*, SIAM J. Control, 8 (1970), pp. 331-337.

[5] S. Battilotti, *A sufficient condition for noninteracting control with stability via dynamic state-feedback*, IEEE Trans. Automat. Control, 36 (1991), pp. 1033–1045.

[6] S. Battilotti and W. P. Dayawansa, *Necessary and sufficient conditions for noninteracting control with stability for a class of nonlinear systems*, Systems Control Lett., 17 (1991), pp. 327–338.

[7] K. Wagner, *Nonlinear noninteraction with stability by dynamic state-feedback*, SIAM J. Control Optim., 29 (1991), pp. 609–621.

[8] W. Zhan, T. J. Tarn, and A. Isidori, *A canonical dynamic extension for noninteraction with stability for affine nonlinear square systems*, Systems Control Lett., 17 (1991), pp. 177–184.

[9] S. Morse and W. M. Wonham, *Status of noninteracting control*, IEEE Trans. Automat. Control, 16 (1971), pp. 568–581.

[10] S. Battilotti, *A sufficient condition for noninteracting control with stability via dynamic state-feedback: block-partitioned outputs*, Internat. Control, 55 (1992), pp. 1141–1160.

[11] H. Nijmeijer and J. M. Schumacher, *Zeroes at infinity for affine nonlinear control systems*, IEEE Trans. Automat. Control, 30 (1985), pp. 566–573.

[12] J. W. Grizzle and A. Isidori, *Block noninteracting control with stability via static state-feedback*, Math. Control Systems Signals, 2 (1989), pp. 315–341.

[13] S. Battilotti, *Necessary conditions for block noninteracting control with stability via dynamic feedback laws*, Systems Control Lett., 15 (1992), pp. 481–491

[14] S. N. Singh, *Decoupling of invertible nonlinear systems with state-feedback and precompensation*, IEEE Trans. Automat Control, 26 (1981), pp. 331–345.

[15] H. Nijmeijer and W. Respondek, *Decoupling via dynamic compensation for nonlinear control systems*, Proc. 25th Conference on Decision and Control, Athens, 1986, pp. 212–217.

[16] W. Hahn, *Stability of motion*, Springer-Verlag, New York, 1967.

[17] W. M. Boothby and R. Marino, *Feedback stabilization of planar systems*, Systems Control Lett., 12 (1989), pp. 87–92.

[18] C. Byrnes and A. Isidori, *Local stabilization of minimum phase nonlinear systems*, Syst. and Contr. Lett., 11 (1988), pp. 9–17.

[19] ———, *New results and examples in nonlinear feedback stabilization*, Systems Control Lett., 12 (1989), pp. 437–442.

[20] ———, *Asymptotic stabilization of minimum phase systems*, IEEE Trans. Automat. Control, 36 (1991), pp. 1122–1137.

[21] C. Byrnes, A. Isidori, and J. Willems, *Passivity, feedback equivalence and the global stabilization of minimum phase nonlinear systems*, IEEE Trans. Automat. Control, 36 (1991), pp. 1228–1240.

[22] J. Tsinias, *Asymptotic feedback stabilization: A sufficient condition for the existence of control Lyapunov functions*, Systems Control Lett. 15 (1990), pp. 441–448.

[23] ———, *Existence of control Lyapunov functions and applications to state feedback stabilizability of nonlinear systems*, SIAM J. Control Optim., 29 (1991), pp. 457–473.

[24] ———, *On the existence of control Lyapunov functions: Generalizations of Vidyasagar's theorem on nonlinear stabilizability*, SIAM J. Control Optim., 30 (1992), pp. 879–893.

[25] E. D. Sontag, *A universal construction of Artstein's theorem on nonlinear stabilization*, Systems Control Lett., 13 (1989), pp. 117–123.

[26] L. Praly, B. D'Andrea-Novel, and J. M. Coron, *Lyapunov design of stabilizing controllers*, Proc. of the 28th Conference on Decision and Control, Tampa, Florida, 1989.

[27] J. M. Coron and L. Praly, *Adding an integrator for the stabilization problem*, Systems Control Lett., 17 (1991), pp. 89–104.

[28] J. Tsinias, *A theorem on global stabilization of nonlinear systems by linear feedback*, Systems Control Lett., 17 (1991), pp. 357–362.

[29] ———, *An extension of Artstein's theorem on stabilization by using ordinary feedback integrators*, Systems Control Lett., 20 (1993), pp. 141–148.

[30] W. P. Dayawansa and C. F. Martin, *Asymptotic stabilization of two dimension real analytic systems*, Systems Control Lett., 12 (1989), pp. 205–211.

[31] W. P. Dayawansa, C. F. Martin, and G. Knowles, *Asymptotic stabilization of a class smooth two dimensional systems*, SIAM J. Control Optim.

[32] W. M. Wonham, *Linear Multivariable Control: A Geometric Approach*, 3rd ed., Springer-Verlag, New York, Berlin, 1979.

[33] L. M. Silverman and H. J. Payne, *Input-output structure of linear systems with application to the decoupling problem*, SIAM J. Control Optim., 9 (1971), pp. 219–233.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 33, No. 1, pp. 126–148, January 1995

© 1995 Society for Industrial and Applied Mathematics

006

NONHOLONOMIC CONTROL SYSTEMS ON RIEMANNIAN MANIFOLDS*

A. M. BLOCH† AND P. E. CROUCH‡

Abstract. This paper gives a general formulation of the theory of nonholonomic control systems on a Riemannian manifold modeled by second-order differential equations and using the unique Riemannian connection defined by the metric. The main concern is to introduce a reduction scheme, replacing some of the second-order equations by first-order equations. The authors show how constants of motion together with the nonholonomic constraints may be combined to yield such a reduction. The theory is applied to a particular class of nonholonomic control systems that may be thought of as modeling a generalized rolling ball. This class reduces to the classical example of a ball rolling without slipping on a horizontal plane.

Key words. Riemannian manifold, nonholonomic, control, optimization

AMS subject classifications. 70F25, 70G05, 93A30, 93B05

1. Introduction.

In this paper we consider the formulation of controlled classical nonholonomic systems on Riemannian manifolds. As described in Vershik [1984], Vershik and Fadeev [1981], Vershik and Gershkovich [1988], and Arnold [1988], one can divide the theory of nonholonomic systems into two classes: variational systems derivable from a Lagrangian, and classical (mechanical) systems derivable from D'Alembert's principle. The latter is normally used for describing mechanical systems, and our goal here is to introduce some aspects of control into the theory at the general level of systems evolving on Riemannian manifolds. This paper expands and extends some of the results announced in Bloch and Crouch [1992].

The classical nonholonomic systems have been considered by many others. We mention some of the most important here: Cartan [1952], using the general theory of equivalence; Weber [1986], using the Hamiltonian setting; Hermann [1982], using the Lagrangian setting and Ehresmann's theory of connections on jet spaces; and finally, Bates and Sniatycki [1993], again using a Hamiltonian setting but introducing a theory of reduction under symmetry.

Although little work has been done on the role of control in such systems until very recently, for example Yang [1992] and Bloch et al. [1989], [1992b], a lot of work has been done on the kinematic problem, where the nonintegrable velocity constraints define nonlinear (kinematic) control problems through the direct control of some of the velocities. Examples of this work include Brockett and Dai [1993], Krishnaprasad and Yang [1991], Lafferiere and Sussmann [1991], Montgomery [1990], and Murray and Sastry [1990]; see Bloch et al. [1992b] for further references.

Our work sets out to show that reduction of the dynamic equations of classical nonholonomic systems with external forces or controls can be expressed in a general Lagrangian setting, and that if symmetries are also present they can be incorporated into the same setting. The reduction is effected by introducing a bundle structure, so that the reduced system is defined by first-order differential equations in the fiber space and second-order differential equations in the base, defined through the Riemannian connection induced by a natural Riemannian metric or kinetic energy. This approach is most closely aligned with the important work of Koiller [1992], who considers the reduction process on a principal bundle, the structure group of which is a symmetry group for the problem. These are termed nonabelian

* Received by the editors August 3, 1992; accepted for publication (in revised form) September 28, 1993.

† Department of Mathematics, The Ohio State University, Columbus, Ohio 43210, and University of Michigan, Ann Arbor, Michigan 48109. This author's research was partially supported by National Science Foundation grant PYI DMS-9157556, and Air Force Office of Scientific Research grant F49620-93-1-0037.

‡ Center for Systems Science and Engineering, Arizona State University, Tempe, Arizona 85287. This author's research was partially supported by National Science Foundation grant DMS-91011964.

Caplygin systems. Our framework incorporates this setting as a special case, even though it is an important case.

We illustrate our general prescription by introducing an extended example. We generalize the case of a ball rolling on a flat plane with external forces parallel to the plane to push the ball, as studied by Brockett and Dai [1993]. The generalization incorporates the well-known model of a generalized rigid body, using a compact semi-simple Lie group, together with velocity constraints, linking the generalized rotational motion to a generalized translational motion. We call this system the generalized rolling ball. In the case that the ball is not symmetric (i.e., has an inertia tensor not equal to a multiple of the identity), the nonholonomic velocity constraints are not symmetries, and our reduction procedure generalizes that of Koiller [1988].

We also consider some controllability and optimality properties of the generalized rolling ball in the symmetric case. The controllability result generalizes the local argument of Bloch et al. [1992b], using a general controllability result on principal bundles. We also formalize a minimum force control problem as the higher-order analogue of the minimum energy control problem. In the holonomic case the minimum force control problem for second-order Newtonian systems was studied in Noakes et al. [1989] and Crouch and Leite [1991a], [1991b]. These works treat higher-order variational problems on Riemannian manifolds. We outline the modifications necessary to treat the nonholonomic case.

The outline of the paper is as follows. In §2 we give a general formulation of holonomic control systems on Riemannian manifolds, and in §3 we give the formulation of nonholonomic systems. In §4 we introduce the concept of symmetries and describe the general reduction procedure based on the nonholonomic constraints and the constants of motion derived from the symmetries. In §5, we have four subsections dealing with the control of the generalized rolling ball, in which we describe some preliminaries on Riemannian structures on Lie groups, the generalized rigid body, the generalized rolling ball in both body fixed and inertial axes, and finally controllability and optimality questions.

2. Holonomic control systems and optimal control. In this section we give a brief formulation of mechanical systems, without any velocity constraints, under the influence of external forces. We broadly follow the formulation given by Hermann [1982]. We consider the case of systems with "nonintegrable" velocity constraints in the following section.

Mechanical systems, in which the velocity constraints are integrable, are referred to here as holonomic mechanical systems. These integrable velocity constraints yield constraints on the configuration variables only, and thereby determine a manifold in which the configuration variables are constrained to evolve (see Arnold [1978]). This provides the motivation for considering holonomic control systems in the generality discussed below.

We let M denote a smooth (infinitely differentiable), n-dimensional manifold with a Riemannian metric denoted $\mathcal{K}(\cdot,\cdot)$. TM will denote the tangent bundle to M. The norm of a vector $X_p \in T_pM$ will be denoted by

$$||X_p||_{\mathcal{K}} = \sqrt{\mathcal{K}(X_p, X_p)}.$$

M denotes the configuration space of a mechanical system, whereas TM denotes the phase (or state) space. The notion of an inertia tensor will be modeled by a bundle mapping

$$J : TM \to TM,$$

such that J is the identity on M. Thus for each $p \in M$ we have a linear mapping

$$J_p : T_pM \to T_pM.$$

We assume that for each p, J_p is an isomorphism satisfying for each X_p, $Y_p \in T_pM$

(i) $\mathcal{K}(J_p X_p, Y_p) = \mathcal{K}(X_p, J_p Y_p)$,
(ii) $\mathcal{K}(J_p X_p, X_p) \geq 0$ ($= 0$ if and only if $X_p = 0$).
From J we may define another Riemannian metric on M by setting

$$\langle X, Y\rangle = \mathcal{K}(JX, Y)$$

for all vector fields X, Y on M. We refer to $\mathcal{K}$ as the ambient metric and $\langle\cdot,\cdot\rangle$ as the mechanical metric. The norm of a vector $X_p \in T_pM$ with respect to the mechanical metric will be denoted by

$$\|X_p\| = \sqrt{(\langle X_p, X_p\rangle)}.$$

The mechanical metric determines a unique Riemannian connection on M, denoted ∇, and thereby determines a covariant derivative, denoted $D/\partial t$. A mechanical system is determined by its kinetic energy $T : TM \longrightarrow R$ given by the expression

$$T_q = \frac{1}{2}\left\langle \frac{dq}{dt}, \frac{dq}{dt}\right\rangle$$

and the potential energy $U : M \longrightarrow \mathbf{R}$ given by an arbitrary smooth function on M. Denote the cotangent bundle to M by T^*M. An external force is modeled as a covector field F on M, in general time varying. Thus, $F_p \in T_p^*M$ for each $p \in M$. We define the momentum P as a covector field along the trajectories of the mechanical system on M by setting

$$P_q = \left\langle \frac{dq}{dt}, \cdot\right\rangle = \mathcal{K}\left(J_q\frac{dq}{dt}, \cdot\right).$$

We define an holonomic mechanical system on M to be that given by the basic Newtonian system of equations in T^*M

$$\frac{DP}{\partial t} = F - dU. \tag{1}$$

For each smooth vector field W on M, we have along the motion of (1)

$$\begin{aligned}\frac{D}{\partial t}(P(W)) &= \frac{D}{\partial t}\left\langle\frac{dq}{dt}, W\right\rangle = \left\langle\frac{D^2q}{\partial t^2}, W\right\rangle + \left\langle\frac{dq}{dt}, \frac{DW}{\partial t}\right\rangle\\ &= \frac{DP}{\partial t}(W) + P\left(\frac{DW}{\partial t}\right) = \frac{DP}{\partial t}(W) + \left\langle\frac{dq}{dt}, \frac{DW}{\partial t}\right\rangle.\end{aligned}$$

We deduce that

$$\frac{DP}{\partial t} = \left\langle\frac{D^2q}{\partial t^2}, \cdot\right\rangle = \mathcal{K}\left(J_q\frac{D^2q}{\partial t^2}, \cdot\right). \tag{2}$$

From the definition of kinetic energy and potential energy, equation (1) yields

$$F\left(\frac{dq}{dt}\right) = \frac{d}{dt}U(q(t)) + \frac{d}{dt}T_{q(t)}. \tag{3}$$

We may rewrite the holonomic mechanical system (1) as a system of equations in TM as follows. Let $\hat{X}_i$, $1 \leq i \leq N$ be $N \leq n$ independent vector fields on M, and let $u_i(\cdot)$,

$1 \leq i \leq N$, be input or control functions (real valued functions of time). We then model the force field F by setting

$$F_{q(t)} = \sum_{i=1}^{N} u_i(t)\langle \hat{X}_i(q(t)), \cdot \rangle. \tag{4}$$

Let $J^* : TM \to T^*M$ denote the bundle isomorphism determined on fibers by

$$J_p^* X_p = \langle X_p, \cdot \rangle = \mathcal{K}(J_p X_p, \cdot).$$

From equations (2) and (4) we may rewrite equation (1) in the form

$$\frac{D^2 q}{\partial t^2} = \sum_{i=1}^{N} u_i \hat{X}_i(q) - J_q^{*-1}\, dU_q. \tag{5}$$

Equation (5) now represents a general holonomic mechanical system with inputs. From now on we shall ignore the potential term in this equation, but it may be added without any extra difficulty. When $F \equiv 0$ (and $U \equiv 0$) equation (5) reduces to the geodesic equations on the Riemannian manifold $(M, \langle \cdot, \cdot \rangle)$

$$\frac{D^2 q}{\partial t^2} = 0.$$

This flow is known to be an extremal of the variational problem (Milnor [1963]),

$$\min_{\substack{q \\ q(0)=q_0,\, q(T)=q_T}} \int_0^T \left\| \frac{dq}{dt} \right\|^2 dt. \tag{6}$$

We now introduce a natural optimal control problem for the system (5). First we define a norm on fibers of $T^*\, M$ in the usual way:

$$\|F_q\| = \sup_{|W_q\|_{K} \neq 0} \frac{|F(W_q)|}{\|W_q\|_K}. \tag{7}$$

Note that we use the ambient metric in this definition. We introduce the minimum force control problem as

$$\min_q \int_0^T \frac{1}{2} \|F_{q(t)}\|^2\, dt \tag{8}$$

subject to the dynamics $\frac{DP}{\partial t} = F$ and the boundary conditions

$$q(0) = q_0, \quad \frac{dq}{dt}(0) = \dot{q}_0, \quad q(T) = q_T, \quad \frac{dq}{dt}(T) = \dot{q}_T, \tag{9}$$

or equivalently we may specify $P_{q(0)}$ and $P_{q(T)}$. From (2) and (7) we obtain

$$\|F_q\| = \sup_{\|W_q\|_K \neq 0} \frac{\mathcal{K}\left(J_q \frac{D^2 q}{\partial t^2}, W_q\right)}{\sqrt{(\mathcal{K}(W_q, W_q)}} = \left\| J_q \frac{D^2 q}{\partial t^2} \right\|_{\mathcal{K}}.$$

Thus the cost functional (8) may be reformulated as

$$\min_q \int_0^T \frac{1}{2} \left\langle \frac{D^2 q}{\partial t^2}, J_q \frac{D^2 q}{\partial t^2} \right\rangle dt. \tag{10}$$

It is now natural to consider the formulation (5) of the holonomic control system. It is convenient to assume a little more structure for the force field F defined in (4). We modify the definition as follows:

$$F_{q(t)} = \sum_{i=1}^{N} u_i(t)\langle J_{q(t)}^{-1} X_i(q(t)), \cdot\rangle, \tag{11}$$

where X_i, $1 \le i \le n$ is an orthonormal base of vector fields with respect to the ambient metric

$$\mathcal{K}(X_i, X_j) = \delta_{ij}, \qquad 1 \le i, j \le n. \tag{12}$$

With the force field (11) the system (5) may be rewritten as

$$\frac{D^2 q}{\partial t^2} = \sum_{i=1}^{N} J_q^{-1} X_i(q) u_i(t). \tag{13}$$

The orthonormality assumption (12) now implies that

$$\left\langle \frac{D^2 q}{\partial t^2}, J_q \frac{D^2 q}{\partial t^2} \right\rangle = \sum_{i=1}^{N} u_i^2(t).$$

It follows that for system (13) the minimum force control problem is defined by the cost functional

$$\min_u \int_0^T \frac{1}{2} \sum_{i=1}^{N} u_i^2(t)\, dt \tag{14}$$

subject to the boundary conditions (9).

In the case $N = n$ this optimal control problem corresponds to the higher-order variational problem posed by the functional (10) with boundary conditions (9). Similar higher-order variational problems have been treated in various contexts, most notably as the minimum curvature problem (Griffiths [1983], Jurdjevic [1991]), but recently they have occurred in the context of interpolation problems in Gabriel and Kajiya [1988], Noakes et al. [1989], and Crouch and Silva-Leite [1991a], [1991b]. In the latter three works a simpler functional is considered, namely

$$\min_q \int_0^T \frac{1}{2} \left\| \frac{D^2 q}{\partial t^2} \right\|^2 dt. \tag{15}$$

The "normal" extremals of such functionals satisfy an equation of the form

$$\frac{D^4 q}{\partial t^2} + R\left(\frac{D^2 q}{\partial t^2}, \frac{Dq}{\partial t}\right) \frac{Dq}{\partial t} \equiv 0, \tag{16}$$

where R is the curvature tensor associated with the connection ∇. It follows that for $N = n$, the minimum force control problem introduced above is a natural higher-order version of the classical variational problem (6), which is often interpreted as a minimum energy problem.

In the control literature another class of optimal control problem has received much attention. It may be characterized by the cost functionals (14) subject to systems of the form

$$\frac{dq}{dt} = \sum_{i=1}^{N} v_i \bar{X}_i(q), \quad (N < n), \quad q \in M, \tag{17}$$

where $\bar{X}_i$ are independent vector fields on M and v_i are control functions. Although these problems are governed by first-order rather than second-order equations (13), the singular nature of this optimal control problem, because of the fact that $N < n$, is of the same character as that of the minimum force control problem when $N < n$. Brockett [1982] considered the control problem posed by (17) and (14) (see also Baillieul [1978]), and subsequently the analysis has been taken up in the mathematics and control theory literature as sub-Riemannian geometry. In the next section we reinterpret the class of systems (17) as kinematic nonholonomic control systems (Remark 3).

3. Nonholonomic control systems. We now consider the formulation of controlled nonholonomic control systems. Nonholonomic systems may be divided into two classes (see, e.g., Vershik and Gershkovich [1988])—variational nonholonomic systems (dubbed vakonomic systems by Arnold [1988]) and classical (or mechanical) nonholonomic systems. In either case the basic ingredients are "nonintegrable" constraints on phase space, defined by m, $m < n$, one-forms on M, $\omega_1, \dots, \omega_m$, such that their span, over the smooth functions on M, contains no nontrivial exact forms and in particular none of the forms $\omega_1, \dots, \omega_m$ is exact. These forms then define a smooth distribution H on M. For each $p \in M$, H_p is the subspace of T_pM defined by

$$H_p = \{X_p \in T_pM;\ \omega_k(X_p) = 0,\ 1 \le k \le m\}.$$

We also stipulate that the distribution H is nonsingular, so that the dimension of the subspace H_p does not vary with p, although the significance of this is not fully understood.

Variational nonholonomic systems are obtained in our context as solutions of variational problems of the form

$$\min_q \int_0^T \frac{1}{2} \left\| \frac{dq}{dt} \right\|^2 dt$$

subject to

$$\omega_k\left(\frac{dq}{dt}\right) \equiv 0, \qquad 1 \le k \le m$$

and boundary conditions $q(0) = q_0$, $q(T) = q_T$.

This problem is solved in the usual way by introducing Lagrange multipliers μ_k and solving the new variational problem defined by

$$\min_q \int_0^T \left[\frac{1}{2} \left\| \frac{dq}{dt} \right\|^2 + \sum_{k=1}^m \mu_k \omega_k\left(\frac{dq}{dt}\right)\right] dt,$$

$$\omega_k\left(\frac{dq}{dt}\right) \equiv 0, \qquad 1 \le k \le m.$$

If we consider the kinematic system (17), in which the vector fields $\hat{X}_i$ are orthonormal with respect to the mechanical metric

$$\langle \hat{X}_i, \hat{X}_j \rangle = \delta_{ij}, \qquad 1 \le i, \quad j \le m,$$

and we define $n - m$ independent dual one-forms ω_k satisfying

$$\omega_k(\hat{X}_i) = 0, \quad 1 \le k \le n - m, \quad 1 \le i \le m,$$

it follows that the variational problem described above coincides with the optimal control problem posed by the cost functional (14), and subject to the kinematics (17). In this context the reader should be aware of work by Kirshnaprasad and Yang [1991] and many others including Brockett [1982], Baillieul [1978], Murray and Sastry [1990], and Lafferiere and Sussmann [1991]. See also Remark 3 below.

Classical nonholonomic systems are not obtained from a variational principle in the usual sense (see Remark 2 below), but from D'Alembert's principle. The equations may be written in the form

$$\frac{D^2q}{\partial t^2} = \sum_{i=1}^{m} \lambda_i W_i, \tag{18}$$

$$\omega_k\left(\frac{dq}{dt}\right) \equiv 0, \qquad 1 \leq k \leq m. \tag{19}$$

The vector fields W_i on M are determined through the identity

$$\omega_k(X) = \langle W_k, X\rangle, \qquad 1 \leq k \leq M$$

for any vector field X on M. The multipliers λ_i are also uniquely determined from the equations (18) and (19) as shown in Remark 2 below.

After our formulation of a holonomic controlled mechanical system (5), we define a nonholonomic controlled mechanical system to have the form

$$\frac{D^2q}{\partial t^2} = \sum_{i=1}^{m} \lambda_i W_i + \sum_{i=1}^{N} u_i \hat{X}_i, \tag{20}$$

$$\omega_k\left(\frac{dq}{dt}\right) = \left\langle W_k, \frac{dq}{dt}\right\rangle \equiv 0, \qquad 1 \leq k \leq m. \tag{21}$$

At this point we do not impose any relation between m and N.

We now summarize some remarks about the formulation (20).

Remark 1. After our general formulation of a holonomic system with external inputs as equations in T^*M, we may similarly define a nonholonomic system with external inputs as follows:

$$\begin{aligned} &\frac{DP}{\partial t} = F, \\ &\omega_k\left(\frac{dq}{dt}\right) \equiv 0, \qquad 1 \leq k \leq m. \end{aligned} \tag{22}$$

Differentiating the constraints we obtain

$$0 = \frac{d}{dt}\omega_k\left(\frac{dq}{dt}\right) = \frac{d}{dt}\left\langle W_k, \frac{dq}{dt}\right\rangle = \left\langle \frac{DW_k}{\partial t}, \frac{dq}{dt}\right\rangle + \left\langle W_k, \frac{D^2q}{\partial t^2}\right\rangle. \tag{23}$$

From equations (2) and (22) we obtain

$$F_q(W_k) = \left\langle \frac{D^2q}{\partial t^2}, W_k\right\rangle, \qquad 1 \leq k \leq m,$$

and so from (23) we deduce that

$$F_q(W_k) + P_q\left(\frac{DW_k}{\partial t}\right) \equiv 0, \qquad 1 \le k \le m. \tag{24}$$

Thus the force field F in a nonholonomic system (22) is not arbitrary but satisfies the constraints (24). It follows that regardless of N, or the vector fields $\hat{X}_i$ in the nonholonomic control system (20), after solving for the multipliers λ_i, the resulting system has at most $n - m$ independent control directions (as specified by vector fields that multiply the control functions u_i).

Remark 2. We note that although the system (20) and (21) may not arise from a variational principle in the usual Lagrangian sense, it is a solution of an instantaneous variational problem (Vershik [1984])

$$\min_{\frac{D^2q}{\partial t^2}} \left\| \frac{D^2q}{\partial t^2} - \sum_{i=1}^{N} u_i\hat{X}_i \right\| \quad \text{subject to } \left\langle W_k, \frac{dq}{dt}\right\rangle \equiv 0, \qquad 1 \le k \le m.$$

This is made clear by differentiating the constraints, as in (23), yielding m affine constraints in $D^2q/\partial t^2$. The multipliers λ_k in (20) are the solutions of the following system of equations

$$\sum_{j=1}^{m} \langle W_k, W_j\rangle\lambda_j = -\left\langle \frac{DW_k}{\partial t}, \frac{dq}{dt}\right\rangle - \sum_{i=1}^{N} u_i\langle W_k, \hat{X}_i\rangle, \qquad 1 \le k \le m. \tag{25}$$

Remark 3. We define the kinematic nonholonomic control system, corresponding to the dynamic nonholonomic control system (20), (21), to be a system of the form

$$\frac{dq}{dt} = \sum_{i=1}^{N} v_i\bar{X}_i(q), \tag{26}$$

where v_i are also control functions and $\bar{X}_i$ are vector fields spanning the distribution H. In general $\bar{X}_i$ differ from $\hat{X}_i$ appearing in the dynamic equations (20), the particular choice governed by physical consideration of the control mechanism. However, the general principle governing the choice of $\bar{X}_i$ and v_i is that v_i control all of the independent components of the velocity $\frac{dq}{dt}$. For controllability analysis of the system (26), the assumption is made that the distribution H is completely nonholonomic (see Vershik and Gershkovich [1988]), that is, the involutive closure H_L of H satisfies

$$H_L(p) = T_pM, \qquad p \in M.$$

This is also an important observation in Brockett [1982]. It is shown in Bloch et al. [1992b] that under certain circumstances, controllability of the associated dynamic nonholonomic control system (20), (21) follows from the assumption above.

Remark 4. Computing the derivative of the kinetic energy along solutions of the nonholonomic control system (20), (21) yields

$$\frac{d}{dt}\frac{1}{2}\left\langle \frac{dq}{dt}, \frac{dq}{dt}\right\rangle = \left\langle \frac{dq}{dt}, \sum_{i=1}^{m} \lambda_iW_i\right\rangle + \left\langle \frac{dq}{dt}, \sum_{i=1}^{N} u_i\hat{X}_i\right\rangle.$$

Thus

$$\frac{d}{dt}T_{q(t)} = F_{q(t)}\left(\frac{dq}{dt}\right),$$

where F is defined by (4). Because the same equation holds for the holonomic system (equation (3) with $U \equiv 0$), it is clear that the constraints do no work, as is well known (see, e.g., Neimark and Fufaev [1972]).

Example 1. We consider a penny rolling on the x–y plane with rotation angle θ and heading angle ϕ, as in Bloch and McClamroch [1989] and Bloch et al. [1992b]. We have two controls, one that rolls the penny about its center of mass and another that turns it about its vertical axis. The constraints are given by

$$\dot{x} = \dot{\theta} \cos \phi,$$
$$\dot{y} = \dot{\theta} \sin \phi.$$

Letting $q = (x, y, \theta, \phi)^T$, the dynamic equations of motion (20) may be written as

$$\ddot{q} = u_1 \hat{X}_1 + u_2 \hat{X}_2 + \lambda_1 W_1 + \lambda_2 W_2,$$

where $\hat{X}_1 = (0,0,1,0)^T$, $\hat{X}_2 = (0,0,0,1)^T$, $W_1 = (1,0,-\cos \phi, 0)^T$, and $W_2 = (0,1,-\sin \phi, 0)^T$. It easily follows that $\lambda_1 = \frac{d}{dt}(\dot{\theta} \cos \phi)$, $\lambda_2 = \frac{d}{dt}(\dot{\theta} \sin \phi)$, and so the resulting system is given by

$$\ddot{x} = \frac{d}{dt}(\dot{\theta} \cos \phi), \quad \ddot{y} = \frac{d}{dt}(\dot{\theta} \sin \phi), \quad 2\ddot{\theta} = u_1, \quad \ddot{\phi} = u_2.$$

On the other hand, the kinematic equations are

$$\dot{q} = v_1 \bar{X}_1 + v_2 \bar{X}_2,$$

where $\bar{X}_1 = (\cos \phi, \sin \phi, 1, 0)^T$ and $\bar{X}_2 = (0,0,0,1)^T$ lie in the distribution H.

Example 2. We now consider the "Heisenberg" system (so called because its vector fields generate the Heisenberg algebra)—see Brockett [1982] and the work of Vershik et al. (see e.g., Vershik and Gershkovich [1988]). Here we have a system on $\mathbf{R}^3$ in the variables (x, y, z) and subject to the constraint

$$\dot{z} = y\dot{x} - x\dot{y}.$$

We have controls in the x and y directions and wish to control the system in $\mathbf{R}^3$. Letting $q = (x, y, z)^T$, the natural dynamic nonholonomic control system (20) may be written as

$$\ddot{q} = u_1 X_1 + u_2 X_2 + \lambda W,$$

where $X_1 = (1,0,0)^T$, $X_2 = (0,1,0)^T$, and $W = (-y, x, 1)^T$. It is easy to see that in this case

$$\lambda = (yu_1 - xu_2)/(1 + y^2 + x^2),$$

so that the dynamics become

$$\phi\ddot{x} = (1 + x^2)u_1 + xyu_2,$$
$$\phi\ddot{y} = (1 + y^2)u_2 + xyu_1,$$
$$\phi\ddot{z} = yu_1 - xu_2,$$

where $\phi(x, y) = 1 + y^2 + x^2$.

The corresponding kinematic system, as discussed in Brockett [1982] is given by

$$\dot{q} = v_1 \bar{X}_1 + v_2 \bar{X}_2,$$

where $\bar{X}_1 = (1, 0, y)^T, \bar{X}_2 = (0, 1, -x)^T$.

4. Symmetries and reduction. Symmetries in mechanics give rise to constants of the motion (see for example Abraham and Marsden [1978]). In the Riemannian context isometries are generated by Killing vector fields. Recall that Z is a Killing vector field with respect to the mechanical metric if

$$\langle \nabla_Y Z, Y\rangle = 0$$

for all vector fields Y. Further, a sufficient condition for $\langle Z(q), \frac{dq}{dt}\rangle$ to be a constant of motion for a geodesic flow is that Z is a Killing vector field. For controlled nonholonomic systems we have the following restatement of Theorem 6, p. 82 in Arnold [1988].

LEMMA 1. *Sufficient conditions for* $\langle Z, \frac{dq}{dt}\rangle$ *to be a constant of motion for the controlled nonholonomic system* (20), (21) *are*

(i) $Z \in H$,

(ii) $Z \in \operatorname{Span}\{\hat{X}_1, \dots, \hat{X}_N\}^{\perp}$ (*with respect to* $\langle\cdot,\cdot\rangle$),

(iii) Z *is a Killing vector field.*

Proof. As in equation (23) we have

$$\frac{d}{dt}\left\langle Z, \frac{dq}{dt}\right\rangle = \left\langle \frac{DZ}{\partial t}, \frac{dq}{dt}\right\rangle + \left\langle Z, \frac{D^2 q}{\partial t^2}\right\rangle.$$

The first term is zero by (iii), and the second is zero by the expression for $D^2q/\partial t^2$ in (20) and (i) and (ii). □

Note that when $M = \mathbf{R}^n$ and the metric $\langle\cdot,\cdot\rangle$ is independent of the coordinate function x_i, then $\partial/\partial x_i$ is a Killing vector field.

Note also that if the uncontrolled nonholonomic system (18) is determined by constraints (19) in which the vector fields W_i are indeed Killing vector fields, then equation (23) gives $\langle W_k, D^2q/\partial t^2\rangle \equiv 0$. It then follows from (18) that $\lambda_k = 0$, $1 \le k \le m$, so that the flow of such a nonholonomic system is a restriction of the geodesic flow. This is illustrated in Example 2.

It is often convenient to introduce a bundle structure in M, $\pi : M \to B$, with fiber F, dim. $B = r$, and dim. $F = n - r$. This structure reflects the natural geometric structure of the system induced by the constraints and must be compatible with the constraints in the sense that

$$\pi_* H_p = T_{\pi(p)} B \quad \text{for all}\, p \in M.$$

Clearly this forces dim. $H = n - m \ge$ dim. $B = r$. Our aim is to reduce the controlled nonholonomic system (20), (21) so that the evolution on the fiber is given by a first-order equation. To this end we introduce two further assumptions. Either Assumption 1 or Assumption 2 follows.

Assumption 1. dim. $H =$ dim. B, that is, $n = m + r$. In this case $\hat{H} = H$ clearly defines a horizontal distribution on the bundle.

Assumption 2. dim. $H-$ dim. $B = n - m - r = s > 0$, and there exist s linearly independent vector fields $Z_1, \dots, Z_s$ that satisfy conditions (i)–(iii) of Lemma 1. In particular,

$$\left\langle Z_i, \frac{dq}{dt}\right\rangle = c_i = \text{const.} \tag{27}$$

are constants of the motion for (20), (21). We define a distribution $\hat{H}_0$ on M by setting $X \in \hat{H}_0$ if

$$\begin{aligned} \langle W_k, X\rangle &= 0 \qquad 1 \le k \le m, \\ \langle Z_k, X\rangle &= 0 \qquad 1 \le k \le s, \end{aligned}$$

and assume that $\hat{H}_0$ satisfies

$$T_pF \cap \hat{H}_0 = \{0\} \quad \text{for all } p \in M. \tag{28}$$

Finally we define the r-dimensional affine distribution $\hat{H}$ on M by setting $X \in \hat{H}$ if

$$\begin{aligned} \langle W_k, X\rangle &= 0 \qquad 1 \le k \le m, \\ \langle Z_k, X\rangle &= c_k \qquad 1 \le k \le s. \end{aligned} \tag{29}$$

Note that Assumption 2 ensures that $\hat{H}_0$ is a constant r-dimensional distribution on M, and (28) ensures that $\hat{H}_0$ is a horizontal distribution on the bundle. If either Assumption 1 or 2 holds, we have as a direct sum of affine spaces

$$T_pM = \hat{H}_p \oplus T_pF \quad \text{for all } p \in M.$$

It follows that any vector field Y on M can be decomposed uniquely into components

$$Y_p = Y_p^H + Y_p^F, \quad p \in M,\, Y_p^H \in \hat{H}_p, \quad Y_p^F \in T_pF.$$

With this structure we may decompose the velocity

$$\frac{dq}{dt} = \dot{q}^H + \dot{q}^F, \quad \dot{q}^H \in \hat{H}_q, \quad \dot{q}^F \in T_qF.$$

In general $\dot{q}^H$ and $\dot{q}^F$ are not derivatives of functions q^H and q^F on M, although in many applications one can indeed identify such functions. From equations (21) and (29) we obtain

$$\begin{aligned} \langle W_k, \dot{q}^F\rangle &= -\langle W_k, \dot{q}^H\rangle, \qquad 1 \le k \le m, \\ \langle Z_k, \dot{q}^F\rangle &= -\langle Z_k, \dot{q}^H\rangle + c_k, \qquad 1 \le k \le s. \end{aligned}$$

Condition (28) allows us to solve these equations uniquely for $\dot{q}^F$

$$\dot{q}^F = f_F(q, \dot{q}^H). \tag{30}$$

Note that f_F is affine in $\dot{q}^H$. From the original controlled nonholonomic system (20) we may deduce equations of the form

$$\frac{D\dot{q}^H}{\partial t} = f_H(q, \dot{q}, u). \tag{31}$$

Equations (30) and (31) provide a reduction of the $2n$ first-order equations (20), with constraints (21), to $n + r$ first-order equations, without constraints. Locally we can write $\dot{q}^H = \frac{d}{dt}q^B$ for some trajectory $q^B(t) \in B$. In some cases we may be able to rewrite equations (30) and (31) globally in terms of a trajectory $q(t) = (q^B(t), q^F(t))$, $q^B \in B$, $q^F \in F$.

The class of Caplygin control systems introduced in Bloch et al. [1992b] corresponds to Assumption 1, in the case where M is a product $M = B \times F$, $B = \mathbf{R}^r$, $F = \mathbf{R}^{n-r}$, $\langle \cdot, \cdot \rangle$ is the Euclidean metric, and the whole dynamics (30), (31) is invariant with respect to q^F. A global prescription for these systems is given in the form

$$\begin{aligned} \dot{q}^F &= f_F(q^B, \dot{q}^B), \\ \ddot{q}^B &= f_B(q^B, \dot{q}^B, u), \end{aligned}$$

where f_F is affine in $\dot{q}^B$.

Another class of systems in which a global reduction is possible is the class of controlled nonabelian Caplygin systems. The uncontrolled systems of this type were discussed in Koiller [1988]. We define a controlled nonabelian Caplygin system as a system (30), (31), in which M is a principal G bundle $M(G, B)$ for a Lie group G, G acts by isometries on the mechanical metric, H (in either Assumption 1 or 2) is invariant under G in the sense that

$$H_{g \cdot p} = g_* H_p, \quad g \in G, \quad p \in M,$$

and finally the vector fields $\hat{X}_i$ are G invariant

$$g_* \hat{X}_i(p) = \hat{X}_i(g \cdot p), \quad g \in G, \quad p \in M, \quad 1 \leq i \leq N.$$

The reduced system of equations can be written (locally) (by analogy with the work of Koiller) in the form

$$\begin{aligned} \dot{g} &= f_F(q_B, \dot{q}_B, g), & g &\in G, \\ \frac{D\dot{q}_B}{\partial t} &= f_B(q_B, \dot{q}_B, u), & q_B &\in B \cong M/G, \end{aligned} \tag{32}$$

where f_F is a G invariant vector field on G.

These reduced systems fall directly within the class of control systems on principal bundles $M(G, B)$ studied by San Martin and Crouch [1984]. If D is a set of vector fields on M, we say that D is projectable if for each $X \in D$ there exists X' on B such that $\pi_* X = X' \circ \pi$. We have the following result.

THEOREM 1 (San Martin and Crouch [1984]). *Let $M(B, G)$ be a connected principal fiber bundle with G a compact Lie group and D a G-invariant, projectable family of vector fields on M defining a control system Σ_D, which is accessible. Denote by Σ'_D the system on B defined by $D' = \pi(D)$. Then Σ_D is controllable if and only if Σ'_D is controllable.*

We have the following corollary.

COROLLARY 1. *Consider a nonabelian Caplygin control system with compact structure group. Assume that the reduced system* (32) *is accessible and the system on the base is controllable. Then the reduced system on M is controllable.*

5. An example: The generalized rolling ball.

5.1. Background on Lie groups.

In this section we briefly review some of the material relating to the dynamical equations of the generalized rigid body. The material in the next two sections may be found in books by Arnold [1978], Hermann [1977], Boothby [1975], etc. Let G be an L-dimensional compact semi-simple Lie group with identity element e, and let $\mathcal{G}$ denote its Lie algebra. There exists a positive definite inner product on $\mathcal{G}$ that we denote by $\langle\langle \cdot, \cdot \rangle\rangle$, defined as a multiple of the Killing form. If X is an element of $\mathcal{G}$, then we can define left and right invariant vector fields on M by setting

$$X_g^l = L_{g_*} X, \qquad X_g^r = R_{g_*} X,$$

where L_g and R_g are the left and right translations on G by $g \in G$.

Let Ad denote the adjoint mapping on $\mathcal{G}$; if $\phi_h : G \to G$ is defined by

$$\phi_h(g) = h^{-1} g h = R_h L_{h^{-1}}(g),$$

then

$$\text{Ad } h = \phi_{h_*}|_{g=e}.$$

Note that with this definition Ad is an anti-homomorphism Ad $gh = $ Ad h Ad g. Now we have

$$X^l_g = L_{g*}X = R_{g*}(R_{g^{-1}_*}L_{g*})X = R_{g*}(\text{Ad } g^{-1}(X)).$$

Thus,

$$X^l_g = (\text{Ad } g^{-1}X)^r_g. \tag{33}$$

For compact semi-simple Lie groups, Ad g is an isometry of $\mathcal{G}$, for every $g \in G$, with respect to the Killing form. Thus we have

$$\langle\langle \text{Ad } gX, \text{ Ad } gY\rangle\rangle = \langle\langle X, Y\rangle\rangle, \quad X, Y \in \mathcal{G}, \quad g \in G. \tag{34}$$

We may now define a bi-invariant Riemannian metric on G by setting

$$\mathcal{K}(X^l_g, Y^l_g) = \langle\langle X, Y\rangle\rangle = \mathcal{K}(X^r_g, Y^r_g).$$

The bi-invariance follows directly from (33) and (34).

If $h(\cdot)$ is a curve in G satisfying the equation

$$\dot{h} = Z^r_h, \quad h(0) = e, \quad h \in G, \tag{35}$$

then

$$\frac{d}{dt} \text{ Ad } h(t)Y|_{t=0} = [Z, Y] \overset{\Delta}{=} \text{ad } Z(Y).$$

Note that the definition of Lie bracket on $\mathcal{G}$ coincides here, under the identification of $\mathcal{G}$ with T_eG, with the standard definition of the Lie bracket

$$[W, V]_g(f) = W_g(V(f)) - V_g(W(f))$$

for all functions f and G and vector fields W and V on G.

Applying (34) to Ad $h(t)$ and differentiating, yields

$$\langle\langle \text{ad } Z(X), Y\rangle\rangle + \langle\langle X, \text{ ad } Z(Y)\rangle\rangle = 0, \qquad X, Y, Z \in \mathcal{G}.$$

Let J be a positive definite linear mapping $J : \mathcal{G} \to \mathcal{G}$ satisfying

(i) $\langle\langle J(X), Y\rangle\rangle = \langle\langle X, J(Y)\rangle\rangle$,

(ii) $\langle\langle J(X), X\rangle\rangle \geq 0$ ($= 0$ if and only if $X = 0$).

We now define a right invariant metric on G by setting

$$\langle X^r_g, Y^r_g\rangle = \langle\langle X, J(Y)\rangle\rangle, \qquad X, Y \in \mathcal{G}.$$

We may extend J to a linear isomorphism $J_g : T_gG \to T_gG$ by setting

$$J_gY^r_g = (JY)^r_g, \quad g \in G, \quad Y \in \mathcal{G}.$$

It follows that

$$\langle V_g, W_g\rangle = \mathcal{K}(J_gV_g, W_g)$$

for all vector fields V and W on M. Corresponding to the right invariant metric $\langle\cdot,\cdot\rangle$ there exists a unique Riemannian connection ∇. Explicit formulas for ∇ are given in Arnold [1978] or Nomizu [1954]. Specifically, ∇ defines a bilinear form on $\mathcal{G}$

$$(X, Y) \mapsto \nabla_X Y = \frac{1}{2}\{[X, Y] + J^{-1}[X, JY] + J^{-1}[Y, JX]\}. \tag{36}$$

∇ is now extended to right invariant vector fields on G by setting

$$(\nabla_{X^r}Y^r)(g) = (\nabla_X Y)^r_g, \quad g \in G, \quad X, Y \in \mathcal{G}.$$

If J is a multiple of the identity, $\langle \cdot, \cdot \rangle$ is the same multiple of K, and ∇ reduces to

$$\nabla_{X^r}Y^r = \frac{1}{2}[X^r, Y^r]. \tag{37}$$

For the purposes of our later analysis we prove the following result.

LEMMA 2.

$$(\nabla_{X^r}Y^l)(g) = (\nabla'_X \operatorname{Ad} g^{-1}Y)^r_g, \quad g \in G, \quad X, Y \in \mathcal{G},$$

where ∇' is the bilinear form on $\mathcal{G}$ given by

$$\nabla'_X Y = \nabla_X Y - [X, Y].$$

Proof. Let $X_1, \ldots, X_L$ be a basis of $\mathcal{G}$, and write $\operatorname{Ad} g^{-1}Y = \sum_{k=1}^L f_k(g)X_k$, where f_k are functions on G. Thus

$$\begin{aligned} (\nabla_{X^r}Y^l)(g) &= \sum_{k=1}^L f_k(g)(\nabla_{X^r}X^r_k)(g) + \sum_{k=1}^L X^r(f_k)(g)X^r_k(g), \\ X^r(f_k)(g) &= \frac{d}{dt}f_k(R_g(h(t)))\Big|_{t=0}, \end{aligned} \tag{38}$$

where $h(\cdot)$ is a solution of (35). But

$$\frac{d}{dt}\operatorname{Ad}\,(R_g(h(t)))^{-1}Y\Big|_{t=0} = \frac{d}{dt}\operatorname{Ad} h^{-1}(t)\operatorname{Ad} g^{-1}Y\Big|_{t=0} = -[X, \operatorname{Ad} g^{-1}Y].$$

Thus from (38) we obtain

$$(\nabla_{X^r}Y)^l)(g) = (\nabla_X \operatorname{Ad} g^{-1}Y - [X, \operatorname{Ad} g^{-1}Y])^r_g. \qquad \square$$

5.2. Background on the generalized rigid body. In this section we briefly review material on the motion of a generalized rigid body. This is modeled by geodesic equations on a compact semi-simple Lie group G, with the right invariant metric defined by a positive definite mapping J on $\mathcal{G}$ as described in the previous section. We have two representations of the velocity defined by a right invariant frame and left invariant frame. If $X_1, \ldots, X_L$ is a basis for $\mathcal{G}$, we set

$$\frac{dg}{dt} = \sum_{i=1}^L w_i(t)X^l_i(g), \tag{39}$$

$$\frac{dg}{dt} = \sum_{i=1}^L v_i(t)X^r_i(g). \tag{40}$$

The kinetic energy expressed in terms of the representation (40) is given by

$$T(g) = \frac{1}{2}\left\langle \frac{dg}{dt}, \frac{dg}{dt} \right\rangle = \frac{1}{2}\left\langle\left\langle \sum_{i=1}^L v_iX_i, \sum_{i=1}^L v_iJX_i \right\rangle\right\rangle.$$

Thus the coordinates v_i refer to the velocity with respect to a frame moving in the body, whereas the coordinates w_i refer to the velocity with respect to a frame fixed in inertial space. (See also Arnold [1978, p. 323], but there the roles of left and right are reversed.) We have

$$\sum_{i=1}^{L} v_i X_i^r(g) = \sum_{i=1}^{L} w_i X_i^l(g).$$

Hence, if we now assume that the basis X_i of $\mathcal{G}$ is orthonormal

$$\langle\langle X_i, X_j \rangle\rangle = \delta_{ij},$$

we obtain

$$w_k = \sum_{i=1}^{L} \langle\langle X_i, \mathrm{Ad}\; g^{-1} X_k \rangle\rangle v_i.$$

We now obtain two representations for the acceleration, based on the left and right invariant settings. First we obtain the representation based on (40). Differentiating, we obtain

$$\frac{D^2 g}{\partial t^2} = \sum_{i=1}^{L} \dot{v}_i(t) X_i^r(g) + \sum_{j,i=1}^{L} v_i(t) v_j(t) (\nabla_{X_j^r} X_i^r)(g).$$

Setting $V_t = \sum_{i=1}^{L} v_i(t) X_i$ and $\partial V_t / \partial t = \sum_{i=1}^{L} \dot{v}_i(t) X_i$ as vectors in $\mathcal{G}$, using the expression (36) gives the usual expression (see, for example, Hermann [1977])

$$\frac{D^2 g}{\partial t^2} = \left(\frac{\partial V_t}{\partial t} + J^{-1}[V_t, J V_t] \right)_g^r . \tag{41}$$

Now turning to the representation (39) we have the following result in which we write

$$W_t = \sum_{i=1}^{L} w_i(t) X_i, \quad \frac{\partial W_t}{\partial t} = \sum_{i=1}^{L} \dot{w}_i(t) X_i, \quad \mathbf{J}(g) = \mathrm{Ad}\; g J \,\mathrm{Ad}\; g^{-1}. \tag{42}$$

LEMMA 3.

$$\frac{D^2 g}{\partial t^2} = \left(\frac{\partial W_t}{\partial t} + \mathbf{J}(g)^{-1} [W_t, \mathbf{J}(g) W_t] \right)_g^l . \tag{43}$$

Proof. Differentiating the representation (39), we obtain

$$\frac{D^2 g}{\partial t^2} = \sum_{i=1}^{L} \dot{w}_i(t) X_i^l(g) + \sum_{ij=1}^{L} w_i(t) w_j(t) (\nabla_{X_j^l} X_i^l)(g).$$

But $\nabla_{X_j^l} X_i^l = \nabla_{(\mathrm{Ad}\, g^{-1} X_j)^r} X_i^l$, so using (36) and Lemma (2)

$$\nabla_{X_j^l} X_i^l = \frac{1}{2} \{ -[\mathrm{Ad}\; g^{-1} X_j, \mathrm{Ad}\; g^{-1} X_i] + J^{-1}[\mathrm{Ad}\; g^{-1} X_j, J \,\mathrm{Ad}\; g^{-1} X_i]$$
$$+ J^{-1}[\mathrm{Ad}\; g^{-1} X_i, J \,\mathrm{Ad}\; g^{-1} X_j] \}_g^r .$$

Using the skew symmetry of the Lie bracket we get

$$\begin{aligned}\frac{D^2 g}{\partial t^2} &= \left(\frac{\partial W_t}{\partial t}\right)^l_g + (J^{-1}[\mathrm{Ad}\, g^{-1}W_t, J\,\mathrm{Ad}\, g^{-1}W_t])^r_g \\ &= \left(\frac{\partial W_t}{\partial t} + \mathrm{Ad}\, gJ^{-1}\,\mathrm{Ad}\, g^{-1}[W_t,\, \mathrm{Ad}\, gJ\,\mathrm{Ad}\, g^{-1}W_t]\right)^l_g .\end{aligned}$$

Now using the definition of $\mathbf{J}(g)$ we obtain the stated result in equation (43). □

5.3. The generalized rolling ball. In this section we describe a generalization of a ball rolling on a flat table, in which we model the ball as a generalized rigid body described in the preceding section. We take the configuration space to be given as $M = G \times \mathbf{R}^N$, where G is an L-dimensional compact semi-simple Lie group as in the previous sections with $L > N$. We put two Riemannian structures on M; the ambient structure is defined by setting

$$\mathcal{K}^M_{(g,x)}((X_g, V)(Y_g, W)) = \mathcal{K}_g(X_g, Y_g) + \langle V, W\rangle^E,$$

where $\langle\cdot,\cdot\rangle^E$ is the Euclidean structure on $\mathbf{R}^N$, and (X_g, V), (Y_g, W) are vectors in $T_gG \times T_x\mathbf{R}^N$. The mechanical metric is defined by setting

$$\langle (X_g, V),\, (Y_g, W)\rangle^M_{(g,x)} = \langle X_g, Y_g\rangle_g + \langle V, W\rangle^E,$$

where $\langle\cdot,\cdot\rangle$ is the right invariant mechanical metric on G, defined in terms of a positive definite mapping J of $\mathcal{G}$. The mechanical metric on M determines a Riemannian connection on M, but the product structure defined by the metric enables us to rely on the connections defined by the metrics $\langle\cdot,\cdot\rangle$ and $\langle\cdot,\cdot\rangle^E$ on G and $\mathbf{R}^N$, respectively.

We wish to define a nonholonomic control system on M according to the prescription given by the equations (20) and (21). For this example we suppose that $m = N$, that is, the number of independent control forces is equal to the number of kinematic constraints. The nonholonomic control system is completely described by defining the input forces

$$F = \sum_{i=1}^N u_i\, dx_i = \sum_{i=1}^N u_i \left\langle \frac{\partial}{\partial x_i}, \cdot \right\rangle^E , \tag{44}$$

where x_i, $1 \le i \le N$ are coordinates in $\mathbf{R}^N$, and the constraints ω_k, $1 \le k \le N$

$$\omega_k\left(\left(\frac{dg}{dt}, \frac{dx}{dt}\right)\right) = \left\langle \frac{dg}{dt}, J_g^{-1}X_k^l(g)\right\rangle - \left\langle \frac{dx}{dt}, \frac{\partial}{\partial x_k}\right\rangle^E = 0 \tag{45}$$

where $\left(\frac{dg}{dt}, \frac{dx}{dt}\right)$ is the velocity of the trajectory (g, x) in M, and $X_1, \ldots, X_L$ is an orthonormal basis for $\mathcal{G}$, with respect to $\langle\langle\cdot,\cdot\rangle\rangle$. Note that in the physical (three-dimensional) case $L = 3$, $N = 2$ (see Bloch and Crouch [1992b]) the constraints are

$$\left\langle \frac{dg}{dt}, J_g^{-1}X_1^l(g)\right\rangle - \left\langle \frac{dx}{dt}, \frac{\partial}{\partial x_2}\right\rangle^E = \left\langle \frac{dg}{dt}, J_g^{-1}X_2^l(g)\right\rangle + \left\langle \frac{dx}{dt}, \frac{\partial}{\partial x_1}\right\rangle^E = 0.$$

This follows from setting the velocity of the point of contact of the ball equal to zero and using the fact that a three-dimensional rotation has a unique axis of rotation. Because this is not true in higher dimensions, the constraint (45) appears to be the natural generalization. The system (20) and (21) now becomes

$$\frac{d^2}{dt^2}x_k = -\lambda_k + u_k, \qquad 1 \le k \le N, \tag{46}$$
$$\frac{D^2 g}{\partial t^2} = \sum_{k=1}^{N} \lambda_k J_g^{-1} X_k^l(g),$$

$$\dot{x}_k = \left\langle \frac{dg}{dt}, J_g^{-1} X_k^l(g) \right\rangle, \qquad 1 \le k \le N. \tag{47}$$

THEOREM 2. *The following are constants of motion for the controlled nonholonomic system* (46), (47).

$$\left\langle X_k^l(g), \frac{dg}{dt} \right\rangle, \qquad N < k \le L. \tag{48}$$

Proof. We show that the conditions of Lemma 1 are satisfied.

(i) We must show that X_k^l, $N < k \le L$ belong to the distribution H on M defined by the constraints (45). This is equivalent to the identity

$$\langle X_k^l, J_g^{-1} X_j^l(g) \rangle = 0, \quad 1 \le j \le N, \quad N < k \le L.$$

But this follows from the definition of the mechanical metric on G and the orthonormality of the vector fields X_j.

(ii) We must show that the vector fields X_k^l, $N < k \le L$ are orthogonal to the control vector fields $\partial/\partial x_k$. This follows trivially from the definition of the metric on M.

(iii) We must show that X_k^l, $N < k \le L$ are Killing vector fields with respect to the mechanical metric on M, which reduces to the same problem for the mechanical metric on G. Thus it is sufficient to show that

$$\langle W^r, \nabla_{W^r} X_k^l \rangle \equiv 0, \quad \text{for all vectors } W \in \mathcal{G}.$$

From Lemma 2 we obtain

$$\begin{aligned}\langle W^r, \nabla_{W^r} X_k^l \rangle_g &= \mathcal{K}_g(W_g^r, J_g(\nabla'_W \operatorname{Ad} g^{-1} X_k)_g^r) \\ &= \mathcal{K}_g(W_g^r, (J\nabla'_W \operatorname{Ad} g^{-1} X_k)_g^r) \\ &= \langle\langle W, J\nabla'_W \operatorname{Ad} g^{-1} X_k \rangle\rangle \\ &= \frac{1}{2}\langle\langle W, -J[W, \operatorname{Ad} g^{-1} X_k] + [W, J \operatorname{Ad} g^{-1} X_k] + [\operatorname{Ad} g^{-1} X_k, JW] \rangle\rangle \\ &= 0. \quad \square\end{aligned}$$

It follows directly from this result that along the trajectories of (46), (47) we have for suitable constants c_k,

$$\left\langle \frac{dg}{dt}, X_k^l(g) \right\rangle \equiv c_k, \qquad N < k \le L. \tag{49}$$

We may calculate the multipliers λ_k in (46) as was done in Remark 2. Differentiating (47) and substituting for the second derivatives given in (46), we obtain

$$-\lambda_k + u_k = \sum_{j=1}^{N} \langle J_g^{-1} X_k^l, J_g^{-1} X_j^l \rangle \lambda_j + \left\langle \frac{dg}{dt}, \frac{D}{\partial t} J_g^{-1} X_k^l \right\rangle$$

or

$$u_k - \left\langle \frac{dg}{dt}, \frac{D}{\partial t} J_g^{-1} X_k^l \right\rangle = \lambda_k + \sum_{j=1}^{N} \langle\langle \mathrm{Ad}\; g^{-1} X_k, J^{-1}\, \mathrm{Ad}\; g^{-1} X_j \rangle\rangle \lambda_j. \tag{50}$$

Define a matrix Λ with components given by

$$\Lambda_{kj} = \delta_{kj} + \langle\langle \mathrm{Ad}\; g^{-1} X_k, J^{-1}\, \mathrm{Ad}\; g^{-1} X_j \rangle\rangle, \quad 1 \le k, \quad j \le N. \tag{51}$$

The assumed positive definiteness of J ensures that Λ is always invertible, so that (50) defines the multipliers λ_k uniquely in equations (46).

We now describe the two reductions of equations (46), (47), described in §4. We first consider the reduction based on Assumption 1. In this case we take the base $B = G$ and the fiber $F = \mathbf{R}^N$. The N-independent constraints (47) ensure that dim. $H = (N + L) - N = L =$ dim. B, as required. The reduction procedure described in §4 simply rewrites the second-order equations on the fiber by the constraint equations (47). We may employ equation (50) to eliminate the multipliers λ_k; however, we first employ a simple feedback control, defined by

$$u_k = \left\langle \frac{dg}{dt}, \frac{D}{\partial t} J_g^{-1} X_k^l \right\rangle + \sum_{j=1}^{N} \Lambda_{kj} \bar{u}_j, \qquad 1 \le k \le N, \tag{52}$$

which defines $\bar{u}_j$, $1 \le j \le N$ uniquely. It follows that under this reduction and feedback the system (46), (47) becomes

$$\begin{aligned} \dot{x}_k &= \left\langle \frac{dg}{dt}, J_g^{-1} X_k^l(g) \right\rangle, \qquad 1 \le k \le N \\ \frac{D^2 g}{\partial t^2} &= \sum_{k=1}^{N} J_g^{-1} X_k^l(g) \bar{u}_k. \end{aligned} \tag{53}$$

We note that along solutions of (53), we have

$$\frac{d}{dt}\frac{1}{2}\left\langle \frac{dg}{dt}, \frac{dg}{dt} \right\rangle = \sum_{k=1}^{N} \dot{x}_k \bar{u}_k$$

compared with a similar computation for system (46), (47), given in Remark 4,

$$\frac{d}{dt}\left[\frac{1}{2}\left\langle \frac{dg}{dt}, \frac{dg}{dt} \right\rangle + \frac{1}{2}\langle \dot{x}, \dot{x} \rangle^E\right] = \sum_{k=1}^{N} \dot{x}_k u_k.$$

Thus the work done by the force control $\bar{u}$ in (53) simply changes the generalized rotational energy and not the generalized translational energy.

Finally we may rewrite the reduced system (53) in terms of an "inertial frame" (39), as described in the previous section. Using Lemma 3 to express the acceleration in terms of the velocity W_t in equation (42), we obtain

$$\begin{aligned} \dot{x}_k &= w_k, \qquad 1 \le k \le N, \\ \frac{\partial W_t}{\partial t} + \mathbf{J}(g)^{-1}[W_t, \mathbf{J}(g) W_t] &= \sum_{k=1}^{N} \mathbf{J}(g)^{-1} X_k \bar{u}_k, \end{aligned} \tag{54}$$

$$\frac{dg}{dt} = (W_t)_g^l = \sum_{k=1}^{L} w_k(t) X_k^l(g).$$

We now describe the reduction based on Assumption 2 in §4. In this case we take the base $B = \mathbf{R}^N$ and the fiber $F = G$. We employ the L constraints (47) and (49) to determine the velocity $\frac{dg}{dt}$. Note that we employ all N velocity constraints and $L - N$ constants of motion. To calculate $\frac{dg}{dt}$ explicitly, it is useful to employ the expression for it in terms of an "inertial frame" as given in equations (39). (47) and (49) can then be written as

$$\begin{aligned} \dot{x}_k &= w_k, \qquad 1 \le k \le N \\ c_k &= \sum_{i=1}^{L} \langle\langle X_k, \mathbf{J}(g) X_k \rangle\rangle w_i, \qquad N < k \le L. \end{aligned}$$

From these equations we may define functions f_j, $N < j \le L$,

$$f_j : \mathbf{R}^N \times G \to \mathbf{R}$$

by the following system of equations, using the positive definiteness of J.

$$c_k = \sum_{i=1}^{N} \langle\langle \mathbf{J}(g) X_k, X_i \rangle\rangle \dot{x}_i + \sum_{i=N+1}^{L} \langle\langle \mathbf{J}(g) X_k, X_i \rangle\rangle f_i(\dot{x}_i, g), \qquad N + 1 \le k \le L. \tag{55}$$

It follows that we may rewrite the expression (39) for the velocity $\frac{dg}{dt}$ by substituting $w_k = \dot{x}_k$, $1 \le k \le N$, $w_k = f_k(\dot{x}, g)$, $N + 1 \le k \le L$. The remaining equations of the reduced system are those for $\ddot{x}_k$ in system (46). However, we employ feedback again, expressed in vector form as

$$(\Lambda - I)u + e = \Lambda \hat{u}, \tag{56}$$

where e is the N vector with components

$$\left\langle \frac{dg}{dt}, \frac{D}{\partial t} J_g^{-1} X_k^l \right\rangle, \qquad 1 \le k \le N.$$

The control $\hat{u}$ is defined uniquely by the invertibility of the matrix Λ defined by (51). The resulting reduced equations have the form

$$\begin{aligned} \ddot{x}_k &= \hat{u}_k, \qquad 1 \le k \le N, \\ \frac{dg}{dt} &= \sum_{k=1}^{N} \dot{x}_k X_k^l(g) + \sum_{k=N+1}^{L} f_k(\dot{x}, g) X_k^l(g). \end{aligned} \tag{57}$$

Note that for this system we have

$$\frac{d}{dt} \frac{1}{2} \langle \dot{x}, \dot{x} \rangle^E = \sum_{k=1}^{N} \dot{x}_k \hat{u}_k.$$

Thus in this formulation the force control $\hat{u}$ in (57) simply changes the generalized translational energy and not the generalized rotational energy.

5.4. Controllability and optimal control. In this section we make some comments about the problems associated with the controllability and optimal control of the reduced models (54) and (57). We first comment on the controllability aspects. Clearly the system (54) as written cannot be controllable, because of the constants of motion (49), which can be reexpressed in the form

$$c_k = \langle\langle W_t, \mathbf{J}(g) X_k \rangle\rangle.$$

Reduction of the equations (54) by these constraints would be complicated to analyze, and the reduction has already been performed in the system (57). A necessary condition for controllability is accessibility (or weak controllability) (see Isidori [1989] or Nijmeijer and Van der Schaft [1990]). However, as stated, the Lie algebra associated with the system (57) is also complicated to analyze. We therefore content ourselves to the case where $J = I$, in which case it is easily deduced that both systems (54) and (57) reduce to the system

$$\begin{aligned} \dot{x}_k &= v_k, \\ \dot{v}_k &= \hat{u}_k, \\ \frac{dg}{dt} &= \sum_{k=1}^{N} v_k X_k^l(g) + \sum_{k=N+1}^{L} c_k X_k^l(g), \qquad 1 \le k \le N. \end{aligned} \tag{58}$$

To analyze the accessibility and controllability of this system, we introduce some subspaces of the Lie algebra $\mathcal{G}$. Let P be the subspace spanned by the vectors $X_1, \ldots, X_N$, let $\mathcal{G}_P$ be the subalgebra of $\mathcal{G}$ generated by P and let I_P denote the ideal of $\mathcal{G}_P$ generated by the subspace, $[P, P]$, of $\mathcal{G}_P$.

THEOREM 3. *Assume that $\mathcal{G}$ is a simple Lie algebra. Then the reduced system* (58) *is controllable and accessible if and only if $\mathcal{G}_P = \mathcal{G}$ and $[P, P] \neq 0$.*

Proof. We first analyze accessibility of the system. It is sufficient to analyze the Lie algebra of the system (58) as represented by the Lie algebra L on the vector space $\mathbf{R}^{2N} \times \mathcal{G}$ with generators

$$g_k = \frac{\partial}{\partial v_k}, \quad 1 \le k \le N, \quad f = \sum_{k=1}^{N} v_k \left(\frac{\partial}{\partial x_k} + X_k \right) + \sum_{j=N+1}^{L} c_j X_j.$$

Because the elements of L depend only on $v \in \mathbf{R}^N$, accessibility of (58) is equivalent to the fact that $L_v = \mathbf{R}^{2N} \times \mathcal{G}$, where L_v is the subspace of $\mathbf{R}^{2N} \times \mathcal{G}$ spanned by elements of L evaluated at v. We construct a subalgebra $\hat{L} \subset L$ with the property that none of its elements depend on v and such that $\hat{L} = L_0$. Accessibility is therefore equivalent to $\hat{L} = \mathbf{R}^{2N} \times \mathcal{G}$. (Because $\hat{L}$ does not depend on v, we may identify it with a subspace of $\mathbf{R}^{2N} \times \mathcal{G}$.) Explicitly we set $\hat{L}$ to be the subalgebra of L generated by the vectors $\{g_k, [g_k, f];\ 1 \le k \le N\}$. All Lie brackets of generators g_k and f not in $\hat{L}$ vanish at $v = 0$. It is easily verified that

$$\hat{L} = \text{span}\left\{ \frac{\partial}{\partial v_k};\ \frac{\partial}{\partial x_k} + X_k,\ 1 \le k \le N;\ I_P \right\}.$$

Indeed, span $\{[[f, g_j], [f, g_i]];\ 1 \le j, i \le N\} = [P, P]$. Now $\hat{L} = \mathbf{R}^{2N} \times \mathcal{G}$ if and only if $I_P = \mathcal{G}$, and so accessibility is equivalent to $I_P = \mathcal{G}$. Consider the situation in which $I_P \subseteq \mathcal{G}_P = \mathcal{G}$. Thus I_P is an ideal of a simple Lie algebra $\mathcal{G}$, so that $I_P = 0$ or $I_P = \mathcal{G}$. Because $I_P = 0$ if and only if $[P, P] = 0$, accessibility is equivalent to $\mathcal{G}_P = \mathcal{G}$ and $[P, P] \neq 0$.

We now turn to controllability. We appeal to the result in Theorem 1. Specifically, we treat system (58) as a system on a principal bundle $G \times \mathbf{R}^{2N}$, in which the system is G invariant and projectable. Because we have already assumed that G is compact and the projected system on the base $B = \mathbf{R}^{2N}$ is a controllable linear system, controllability of the total system is equivalent to accessibility of the system. □

We now turn to the question of minimum force control for the reduced systems (53) and (57). First recall the minimum force control problem for the holonomic system (13) with cost functional (14). The reduced nonholonomic system (53) with cost functional

$$\min_u \int_0^T \frac{1}{2} \sum_{i=1}^{N} \bar{u}_k(t)^2 \, dt$$

may be considered as a generalization of the holonomic problem in which we consider the nonholonomic constraints as defining motion on fibers of a bundle over the original holonomic base dynamics. The optimal control minimizes force in the base as measured by the analogue of equation (10)

$$\int_0^T \frac{1}{2} \left\langle \frac{D^2 g}{\partial t^2}, J_g \frac{D^2 g}{\partial t^2} \right\rangle dt.$$

In this setting the minimum force control problem for the nonholonomic system (53) may be viewed as a direct generalization of the optimal control problems posed by nonholonomic kinematic systems discussed at the end of §2 and Remark 3 of §3.

However, the constants of motion (49) for the system (53) make the optimal control problem well posed only on level sets. Thus we consider instead the formulation (57) of the reduced nonholonomic system. The remarks above may be applied equally to the minimal force optimal control problem for this system in which the optimal control minimizes force in the base $B = \mathbf{R}^{2N}$ as measured by the functional

$$\int_0^T \frac{1}{2} \langle \ddot{x}, \ddot{x} \rangle^E \, dt, \tag{59}$$

while subject to the motion in the fibers $F = G$ described in (57). The difficulty in establishing controllability results, except in the case where $J = I$ directs consideration at the system (58). We show how to define the associated minimum force control problem as a constrained variational problem.

Using the skew symmetry of the connection on G in the case $J = I$ (see equation (37)), differentiating along trajectories of (58) gives

$$\frac{D^2 g}{\partial t^2} = \sum_{k=1}^{N} \ddot{x}_k X_k^l(g),$$

so that the functional (59) may be rewritten as

$$\int_0^T \frac{1}{2} \mathcal{K} \left(\frac{D^2 g}{\partial t^2}, \frac{D^2 g}{\partial t^2} \right) dt. \tag{60}$$

The constraints on the evolution of g may be expressed as

$$\frac{dg}{dt} = \sum_{k=1}^{N} \dot{x}_k X_k^l(g) + \sum_{k=N+1}^{L} c_k X_k^l(g).$$

Setting $Z_t(g) = \sum_{k=1}^{L} \delta_k(t) X_k^l(g)$, we obtain

$$\mathcal{K}\left(\frac{dg}{dt}, Z_t(g)\right) - \sum_{k=1}^{N} \delta_k(t)\dot{x}_k - \sum_{k=N+1}^{L} \delta_k(t)c_k \equiv 0. \tag{61}$$

Thus the minimum force control problem for the system (58) has been reduced to the variational problem defined by the functional (60) subject to the constraints (61). Clearly, δ_k, $1 \le k \le L$ are Lagrange multipliers, and in fact δ_k, $1 \le k \le N$ are constants. In this setting the problem now becomes a generalization of the one described in §2 for a certain class of holonomic control problems, where the functional (15) coincides with (60). Although a full resolution of the optimal control problem must consider exceptional (or rigid) trajectories (see Bliss [1946]), the nonexceptional trajectories may be analyzed as in Crouch and Silva-Leite [1991b]. Indeed it is shown there that the necessary conditions for nonexceptional trajectories of the problem (60) subject to (61) are given by

$$\frac{D^3 W_t}{\partial t^3} + R\left(\frac{DW_t}{\partial t}, W_t\right) W_t - \frac{DZ_t}{\partial t} + \frac{1}{2}[Z_t, W_t] \equiv 0,$$

where $W_t = \sum_{k=1}^{L} w_k(t) X_k^l(g)$, generalizing the expression (16) in the unconstrained case. Crouch and Silva-Leite [1991b] show how to reduce these equations to the algebra $\mathcal{G}$, using a well-known expression for the curvature tensor R.

A specific application of the results presented in §5 is given by the ball rolling on a plane, as analyzed in Bloch and Crouch [1992a]. This corresponds to the case where $G = SO(3), L = 3$, and $N = 2$, and the conditions of Theorem 3 are automatically satisfied.

REFERENCES

R. Abraham and J. E. Marsden, *Foundations of Mechanics*, 2nd ed., Addison-Wesley, Reading, MA, 1978.

V. I. Arnold, *Mathematical Methods of Classical Mechanics*, Springer-Verlag, Berlin, New York, 1978.

V. I. Arnold, ed. *Dynamical Systems* III, Springer-Verlag, Berlin, New York, 1988.

J. Baillieul, *Geometric methods for nonlinear optimal control problems*, J. Optim. Theory Appl., 25 (1978), pp. 519–548.

L. Bates and J. Sniatycki, *Nonholonomic reduction*, Rep. Math. Phys, 32 (1993), pp. 99–115.

G. A. Bliss, *Lectures on Calculus of Variations*, University of Chicago Press, Chicago, IL, 1946.

A. M. Bloch and N. H. McClamroch, *Control of mechanical systems with classical nonholonomic constraints*, Proc. of the 28th IEEE Conf. on Decision and Control, IEEE, New York, 1989, pp. 201–205.

A. M. Bloch and P. E. Crouch, *On the dynamics and control of nonholonomic systems on Riemannian manifolds*, Proc. of NOLCOS '92, Bordeaux, 1992a.

A. M. Bloch, M. Reyhanoglu, and N. H. McClamroch, *Control and stabilization of nonholonomic systems*, IEEE Trans. Automat. Control, 37 (1992b), pp. 1746–1757.

W. A. Boothby, *An Introduction to Differentiable Manifolds and Riemannian Geometry*, Academic Press, New York, 1975.

R. W. Brockett, *Control theory and singular Riemannian geometry*, in New Directions in Applied Mathematics, P. J. Hilton and G. S. Young, eds., Springer-Verlag, Berlin, New York, 1982.

R. W. Brockett and Liyi Dai, *Nonholonomic kinematics and the role of elliptic functions in constructive controllability*, in Nonholonomic Motion Planning, Z. Li and J. F. Canny, eds., Kluwer, Boston, 1993, pp. 1–22.

E. Cartan, *Sur La Représentation Géométrique Des Systèmes Matériels Non Holonomes*, Collected Works, Ouevres Complètes, Gauthier-Villars, Paris, 1952.

P. E. Crouch and F. Silva-Leite, *Geometry and the dynamic interpolation problem*, Proc. A.C.C., Boston, MA, 1991a, pp. 1131–1136.

———, *The dynamic interpolation problem: On Riemannian manifolds, Lie groups and symmetric spaces*, J. Dynamical and Control Systems, 1991b.

L. E. Faibusovich, *Explicitly solvable non-linear optimal control problems*, Internat. J. Control, 48 (1988), pp. 2507–2526.

S. Gabriel and J. Kajiya, *Spline interpolation in curved space*, in M.S.R.I. Conf., Berkeley, CA, 1988.

V. Gershkovich and A. Vershik, *Nonholonomic manifolds and nilpotent analysis*, J. Geom. Phys., 5 (1988), pp. 407–452.

P. A. Griffiths, *Exterior Differential Systems*, Birkhauser, Boston, MA, 1983.

R. Hermann, *Differential geometry and the calculus of variations*, in Interdisciplinary Mathematics, Vol XVII, Math Science Press, Brookline, MA, 1977.

———, *The differential geometric structure of general mechanical systems from the lagrangian point of view*, J. Math. Phys., 23 (1982), pp. 2077–2089.

A. Isidori, *Nonlinear Control Systems*, 2nd ed., Springer-Verlag, Berlin, New York, 1989.

V. Jurdjevic, *Non-Euclidean elastica*, Amer. J. Math., 116 (1994), to appear.

———, *The geometry of the plate-ball problem*, Arch. Rational Mech. Anal., 124 (1993), pp. 305–328.

J. Koiller, *Reduction of some classical nonholonomic systems with symmetry*, Arch. Rational Mech. Anal., 118 (1992), pp. 113–148.

K. Nomizu, *Invariant affine connections on homogeneous spaces*, Amer. J. Math., 76 (1954), pp. 33–65.

P. S. Krishnaprasad and R. Yang, *Geometric phases, anholonomy and optimal movement*, Proc. of the Conf. on Robotics and Automation 1991, Sacramento, CA, 1991, pp. 2185–2189.

G. Lafferiere and H. J. Sussman, *Motion planning for completely nonholonomic systems without drift*, Proc. of the Conf. on Robotics and Automation 1991, Sacramento, CA, 1991, pp. 1148–1153.

J. E. Marsden, R. Montgomery, and T. Ratiu, *Reduction, Symmetry and Phases in Mechanics*, Mem. Amer. Math. Soc., Vol. 436, 1990.

J. E. Marsden and J. Scheurle, *Lagrangian reduction and the double spherical pendulum*, Z. Angew. Math. Phys., 44 (1993), pp. 17–43.

W. D. McMillan, *Dynamics of Rigid Bodies*, Duncan MacMillan Rowles, United Kingdom, 1936.

J. Milnor, *Morse Theory*, Princeton University Press, Princeton, NJ, 1963.

R. Montgomery, *Isoholonomic problems and some applications*, Comm. Math. Phys., 128 (1990), pp. 565–592.

R. Murray and S. Sastry, *Steering nonholonomic systems using sinusoids*, Proc. of the 29th IEEE Conf. on Decision and Control, IEEE, New York, 1990, pp. 2097–2101.

J. I. Neimark and F. A. Fufaev, *Dynamics of Nonholonomic Systems*, Transl. Math. Monographs, Vol. 33, American Mathematical Society, Providence, RI, 1972.

H. Nijmeijer and A. J. van der Schaft, *Nonlinear Dynamical Control Systems*, Springer-Verlag, Berlin, New York, 1990.

L. Noakes, G. Heinzinger, and B. Paden, *Cubic splines on curved spaces*, IMA J. Math. Control Inform., 6 (1989), pp. 405–473.

L. San Martin and P. E. Crouch, *Controllability on principal fibre bundles with compact structure group*, Systems Control Lett., 5 (1984), pp. 35–40.

A. M. Vershik, *Classical and nonclassical dynamics with constraints*, in New in Global Analysis, Voronteh, Gos. Univ. (In Russian) (English trans. Lecture Notes in Mathematics, Vol. 1108, Springer-Verlag, Berlin, New York, 1984, pp. 278–301.)

A. M. Vershik and L. D. Faddeev, *Lagrangian mechanics in invariant form*, Selecta Math. Soviet, 1 (1981), pp. 339–350.

A. M. Vershik and V. Ya. Gershkovich, *Nonholonomic problems and the theory of distributions*, Acta Appl. Math., 12 (1988), pp. 181–209.

R. W. Weber, *Hamiltonian systems with constraints and their meaning in mechanics*, Arch. Rational Mech. Anal., 91 (1986), pp. 309–335.

R. Yang, *Nonholonomic Geometry Mechanics and Control*, Ph.D. thesis, University of Maryland, College Park, MD, 1992.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 33, No. 1, pp. 149–167, January 1995

©1995 Society for Industrial and Applied Mathematics
007

THE MINIMIZATION OF SEMICONTINUOUS FUNCTIONS: MOLLIFIER SUBGRADIENTS*

YURI M. ERMOLIEV†, VLADIMIR I. NORKIN‡, AND ROGER J-B. WETS§

Abstract. To minimize discontinuous functions that arise in the context of systems with jumps, for example, we propose a new approach based on approximation via averaged functions (obtained by convolution with mollifiers). The properties of averaged functions are studied, after it is shown that they can be used in an approximation scheme consistent with minimization. A new notion of subgradient is introduced based on approximations generated by mollifiers and is exploited in the design of minimization procedures.

Key words. impulse control, discrete events systems, averaged functions, subgradients, subdifferentiability, stochastic quasi-gradients, epi-convergence

AMS subject classifications. 49J52, 49J55, 49J45

1. Introduction.

It is not unusual to have to deal with optimization problems involving discontinuous functions, for example: optimization problems involving set-up costs or impulse controls (Bensoussan and Lions [5]), the control of discrete events systems (Gong and Ho [14], Rubinstein [36], Ermoliev and Gaivoronski [9]), and control problems with pre- and post-accident regimes whose systems' parameters do not evolve continuously. Even a convex optimization problem is sometimes replaced by one involving discontinuous penalties such as indicator or characteristic functions. Problems defined in terms of marginal functions, expressing the dependence of the optimal value of some subproblem (as in stochastic programming problems, for example) on certain parameters are often discontinuous. To deal with such applications, a number of efforts have been made to develop a subdifferential calculus for nonsmooth, and possibly discontinuous, functions. Among the many possibilities let us mention the notions due to Rockafellar [31], Aubin [3], Clarke [6], Ioffe [18], Frankowska [11], Michel and Penot [25], and Mordukhovich [26] in the context of variational analysis; those due to Warga [43] for subdifferentials obtained via certain approximating scheme; those due to Demyanov and Rubinov [7] for quasi-differentiable functions; and those due to Ermoliev [9] and Polyak [30] in the context of stochastic approximation techniques for optimization problems.

Another approach to the differentiation of "nonclassical" functions, which eventually became known as the *theory of distributions* (in Russia, as the *theory of generalized functions*), was developed in 1930's by Sobolev [38] and Schwartz [37]. This technique is in wide use in mathematical physics and related engineering problems. Although one can find in the literature occasional reference to a connection between these two approaches, the notion of differentiability in the sense of distributions is not used in variational analysis or in the design of solution procedures for optimization problems involving "nonclassical" functions. Probably one of the reasons for this is that in the theory of distributions, (standard) functions defined on $\mathbf{R}^n$ are redefined as functionals on a certain functional space. The same applies to their gradients.

*Received by the editors October 8, 1992; accepted for publication (in revised form) August 20, 1993. This work was supported in part by a grant from the U.S.-Israel Binational Science Foundation.

†International Institute for Applied System Analysis, A-2361 Laxenburg, Austria.

‡Glushkov Institute of Cybernetics, 252207 Kiev, Ukraine.

§Department of Mathematics, University of California, Davis, California 95616.

In the development of a subdifferential calculus for (discontinuous) functions, we appeal to some of the results of the theory of distributions, but our aim is to bring back the algebraic manipulations to operations that can be carried out in $\mathbf{R}^n$, in particular, by assigning certain distributions to points in $\mathbf{R}^n$. More specifically, we associate with a point $x \in \mathbf{R}^n$ a family of mollifiers (density functions) whose support tends toward x and converges to the dirac function $\delta(x - \cdot)$. Given such a family, say $\{\psi_\theta, \theta \in \mathbf{R}_+\}$, a "generalized" function associated with a function $f : \mathbf{R}^n \to \mathbf{R}$ is then defined as the clusters of all possible values generated by the pairings of f with ψ_θ. A set of generalized gradients, here called *mollifier subgradients*, is defined in a similar fashion.

From another angle, we can also link this approach to a technique involving "averaged" functions introduced by Steklov [39], [40] and Sobolev [38]. In the case of continuous functions, these averaged functions converge uniformly to f, and it is then related to an approach suggested by Warga [42]–[44] (see also Frankowska [12]).

For the gradients of averaged functions there are simple unbiased stochastic estimators based on finite differences (some will be mentioned in our development). This opens up the possibility of minimizing the original (discontinuous) function through the minimization of a sequence of smooth approximating averaged functions. Such an approach, initiated in §5, relies on the ideas inherent in stochastic quasi-gradient methods and dynamic nonstationary optimization as were used by Ermoliev and Nurminski [10], Gaivoronski [13], Katkovnik [19], and Nikolaeva [27] in convex nondifferentiable optimization; by Gupal [15], and Mayne and Polak [24] in the Lipschitz continuous case; and by Gupal and Norkin [17] in the discontinuous case.

Section 2 introduces a notion of convergence for discontinuous functions and prepares the way for a justification that averaged functions provide consistent approximations from a minimization viewpoint. Section 3 is devoted to the properties of averaged functions, and §4 introduces the notion of a mollifier subgradient based on the approximation of a discontinuous function by averaged functions. Finally, §5 outlines some potential optimization procedures.

2. eh-convergence. Let $f : \mathbf{R}^n J \to \overline{\mathbf{R}}$ be a proper ($f \not\equiv \infty$, $f > -\infty$) extended real-valued function with $\operatorname{dom} f = \{x \in \mathbf{R}^n | f(x) < \infty\}$ the (nonempty) set on which it is finite. Its *epigraphical* (*or lower semicontinuous*) *closure* $\mathrm{cl}_e f$ is given by

$$\mathrm{cl}_e f(x) := \liminf_{x' \to x} f(x') = \inf_{x^\nu \to x} \liminf_{\nu \to \infty} f(x^\nu)$$

and its *hypographical* (*or upper semicontinuous*) *closure* $\mathrm{cl}_h f$ is

$$\mathrm{cl}_h f(x) := \limsup_{x' \to x} f(x') = \sup_{x^\nu \to x} \limsup_{\nu \to \infty} f(x^\nu);$$

inf and sup are taken over all sequences x^ν converging to x. The function $\mathrm{cl}_e f$ is lower semicontinuous and $\mathrm{cl}_h f$ is upper semicontinuous.

For an arbitrary sequence of functions $\{f^\nu : \mathbf{R}^n \to \overline{\mathbf{R}}, \nu \in \mathbf{N}\}$, we denote by $\mathrm{e}\!-\!\mathrm{li}\, f^\nu$ its *lower epi-limit*, i.e.,

$$(\mathrm{e}\!-\!\mathrm{li}\, f^\nu)(x) := \inf_{x^\nu \to x} \liminf_{\nu \to \infty} f^\nu(x^\nu),$$

and by $\mathrm{h}\!-\!\mathrm{ls}\, f^\nu$ its *upper hypo-limit*, i.e.,

$$(\mathrm{h}\!-\!\mathrm{ls}\, f^\nu)(x) := \sup_{x^\nu \to x} \limsup_{\nu \to \infty} f^\nu(x^\nu);$$

here also inf and sup are calculated with respect to all sequences converging to x. It is easy to see that $\text{e-li}\, f^\nu$ is lower semicontinuous and that $\text{h-ls}\, f^\nu$ is upper semicontinuous (if necessary cf. [33] for more details); note that $\text{h-ls}\, f^\nu = -\,\text{e-li}(-f^\nu)$.

DEFINITION 2.1. *Given a sequence of functions* $\{ f^\nu : \mathbf{R}^n \to \overline{\mathbf{R}}, \nu \in \mathbf{N} \}$*, a function* $f : \mathbf{R}^n \to \overline{\mathbf{R}}$ *is an* epi-sublimit *of the sequence* $\{f^\nu\}$ *if* $\text{cl}_e f \le \text{e-li}\, f^\nu$*. It is a* hypo-suplimit *if* $\text{h-ls}\, f^\nu \le \text{cl}_h f$*. If* f *is both an epi-sublimit and a hypo-suplimit, we say that the sequence* f^ν eh-converges *to* f*.*

We can view eh-convergence as an *extended graph-convergence.* With $\text{gph}\, f^\nu$, the graph of the function f^ν, eh-convergence implies that

$$\text{Limsup}_{\nu\to\infty} \text{gph}\, f^\nu \subset \{ (x,\alpha) \in \mathbf{R}^n \times \overline{\mathbf{R}} \mid \text{cl}_e f(x) \le \alpha \le \text{cl}_h f(x) \}$$

where Limsup is the *outer* (*superior*) set-limit; for a sequence of sets C^ν, $\text{Limsup}_\nu C^\nu$ consists of the cluster points of all sequences $\{u^\nu\}$ with $u^\nu \in C^\nu$ for ν sufficiently large.

A notion of eh-convergence (for functions with values in a function space) also surfaced in the study of the stability properties of integral functionals with discontinuous integrands; cf. Artstein and Wets [1].

3. Averaged functions.

Averaged functions are defined relative of a specific family of mollifiers; our usage of the term mollifier differs somewhat from the standard one in that we do not require that mollifiers be necessarily infinitely differentiable.

DEFINITION 3.1. *Given a locally integrable function* $f : \mathbf{R}^n \to \mathbf{R}$ *and a family of bounded* mollifiers $\{ \psi_\theta : \mathbf{R}^n \to \mathbf{R}_+, \theta \in \mathbf{R}_+ \}$ *that satisfy*

$$\int_{\mathbf{R}^n} \psi_\theta(z)\, dz = 1, \quad \text{supp}\, \psi_\theta := \{ z \in \mathbf{R}^n \mid \psi_\theta(z) > 0 \} \subset \rho_\theta \mathbf{B} \quad \textit{with } \rho_\theta \downarrow 0 \quad \textit{as } \theta \downarrow 0,$$

the associated family $\{ f_\theta, \theta \in \mathbf{R}_+ \}$ *of* averaged functions *is given by*

$$f_\theta(x) := \int_{\mathbf{R}^n} f(x-z)\psi_\theta(z)\, dz = \int_{\mathbf{R}^n} f(z)\psi_\theta(x-z)\, dz.$$

For example, the family of mollifiers could be of the following type: let ψ *be a density function with* $\text{supp}\, \psi$ *bounded,* $\alpha_\theta \downarrow 0$ *as* $\theta \downarrow 0$*, and*

$$\psi_\theta(z) := \frac{\psi(z/\alpha_\theta)}{(\alpha_\theta)^n}.$$

A mollifier is thus a probability density function defined on $\mathbf{R}^n$ *but the family* $\{\psi_\theta\}$ *must possess some specific properties. We can also express* f_θ *as a convolution*

$$f_\theta = f \star \psi_\theta.$$

Sobolev [38] introduced "averaged functions" in his study of generalized functions (distributions) that could serve as solutions of certain equations in mathematical physics; he also required that the mollifiers ψ_θ be of class C^∞. In terms of the theory of distributions, $f_\theta(x)$ is the value of the distribution f at $\psi_\theta(x - \cdot)$, x playing the role of a parameter.

THEOREM 3.2. *Let* $\{ f_\theta, \theta \in \mathbf{R}_+ \}$ *be a family of averaged functions associated with a locally integrable function* $f : \mathbf{R}^n \to \mathbf{R}$*, and suppose that* $x^\theta \to x$ *as* $\theta \downarrow 0$*. Then*

$$\text{cl}_e f(x) \le \liminf_{\theta\downarrow 0} f_\theta(x^\theta) \le \limsup_{\theta\downarrow 0} f_\theta(x^\theta) \le cl_h f(x).$$

Consequently, the averaged functions f_θ eh-converge to f.

Proof. It will suffice to prove the first inequality, the second one is evident and the proof of the last one is similar to that of the first. eh-convergence is an immediate consequence of this string of inequalities.

By definition of lower semicontinuity, for all $x \in \mathbf{R}^n$ and $\varepsilon > 0$ there exists V, a neighborhood of 0, such that $f(x-z) \geq \mathrm{cl}_e f(x) - \varepsilon$ for all $z \in V$. For θ sufficiently small, $\operatorname{supp}\psi_\theta \subset V$ and then

$$f_\theta(x^\theta) = \int_{\mathbf{R}^n} f(x^\theta - z)\psi_\theta(z)\,dz = \int_V f(x^\theta - z)\psi_\theta(z)\,dz \geq \int_V \mathrm{cl}\, f(x^\theta - z)\psi_\theta(z)\,dz$$
$$\geq (\mathrm{cl}_e f(x^\theta) - \varepsilon)\int \psi_\theta(z)\,dz.$$

Hence, $\liminf_{\theta\downarrow 0} f_\theta(x^\theta) \geq \mathrm{cl}_e f(x) - \varepsilon$. The proof is completed by letting $\varepsilon \downarrow 0$. □

COROLLARY 3.3. *Let $f : \mathbf{R}^n \to \mathbf{R}$ be continuous, and let $\{f_\theta, \theta \in \mathbf{R}_+\}$ be an associated family of averaged functions. Then the averaged functions f_θ converge continuously to f, i.e., $f_\theta(x^\theta) \to f(x)$ for all $x^\theta \to x$. In fact, the averaged functions f_θ converge uniformly to f on every bounded subset of $\mathbf{R}^n$.*

Proof. The proof is evident. □

When the function f is not continuous, we cannot expect to have continuous convergence of the averaged functions to f. But that is also more than what is required. For our purposes, we only need to establish that the averaged functions converge to f in a sense that will guarantee the convergence of minimizers and infima. This is precisely what is accomplished by epi-convergence.

DEFINITION 3.4 (Aubin and Frankowska [4], Rockafellar and Wets [33]). *A sequence of functions $\{f^\nu : \mathbf{R}^n \to \overline{\mathbf{R}}, \nu \in \mathbf{N}\}$* epi-converges *to $f : \mathbf{R}^n \to \overline{\mathbf{R}}$* at *$x$ if*

(i) $\liminf_{\nu\to\infty} f^\nu(x^\nu) \geq f(x)$ *for all $x^\nu \to x$;*

(ii) $\lim_{\nu\to\infty} f^\nu(x^\nu) = f(x)$ *for some sequence $x^\nu \to x$.*

The sequence $\{f^\nu\}_{\nu\in\mathbf{N}}$ epi-converges *to f if this holds for all $x \in \mathbf{R}^n$, in which case we write $f = \mathrm{e}\text{-}\mathrm{lm}\, f^\nu$.*

Clearly, if f is the epi-limit of some sequence, then f is necessarily lower semicontinuous. Moreover, if the f^ν converge continuously, and a fortiori uniformly, to f, they also epi-converge to f.

For example, if $(x,y) \mapsto g(x,y) : \mathbf{R}^n \times \mathbf{R}^m \to \overline{\mathbf{R}}$ is (jointly) lower semicontinuous at $(\bar{x},\bar{y})$ and is continuous in y at $\bar{y}$, then for any sequence $y^\nu \to \bar{y}$, the corresponding sequence of functions $\{f^\nu = g(\cdot, y^\nu), \nu \in \mathbf{N}\}$ epi-converges to $f = g(\cdot,\bar{y})$ at $\bar{x}$.

THEOREM 3.5 (Attouch and Wets [2]). *If the sequence $\{f^\nu : \mathbf{R}^n \to \overline{\mathbf{R}}, \nu \in \mathbf{N}\}$ epi-converges to $f : \mathbf{R}^n \to \overline{\mathbf{R}}$ at all $x \in D \subset \mathbf{R}^n$, then*

$$\limsup_\nu(\inf_O f^\nu) \leq \inf_O f, \quad \forall O \subset \mathbf{R}^n \textit{ open},$$
$$\liminf_\nu(\inf_K f^\nu) \geq \inf_K f, \quad \forall K \subset \mathbf{R}^n \textit{ compact},$$

and

$$\forall x^\nu \to x : \quad [f^\nu(x^\nu) \leq \inf f^\nu + \varepsilon_\nu, \quad \varepsilon_\nu \downarrow 0,] \quad \Longrightarrow \quad x \in \operatorname{argmin} f.$$

Epi-convergence of the averaged functions f_θ to f will be guaranteed by the following property of f.

DEFINITION 3.6. *A function $f : \mathbf{R}^n \to \overline{\mathbf{R}}$ is strongly lower semicontinuous at x, if it is lower semicontinuous at x and there exists a sequence $x^\nu \to x$ with f*

continuous at x^ν (for all ν) such that $f(x^\nu) \to f(x)$. The function f is strongly lower semicontinuous if this holds at all x.

Roughly speaking, strong lower semicontinuity excludes the possibility of discontinuities at isolated points. If we think of $(x, f(x))$ as the state of a system, strong lower semicontinuity means that this state can always be reached by following a path along which the evolution of the system is continuous (with no jumps). If x is "time-dependent", then although we may expect sudden changes from one state to another, either before or after the jump, the evolution will be continuous, one doesn't expect instantaneous jumps followed by an immediate return to normal regime.

THEOREM 3.7. *For any strongly lower semicontinuous, locally integrable function $f : \mathbf{R}^n \to \mathbf{R}$, and any associated family $\{ f_\theta,\ \theta \in \mathbf{R}_+ \}$ of averaged functions, we have that $f = \mathrm{e-lm}\, f_\theta$, i.e., for any sequence $\theta^\nu \downarrow 0$, $f = \mathrm{e-lm}\, f_{\theta^\nu}$.*

Proof. Pick any x. For condition (i) of Definition 3.4, simply appeal to Theorem 3.2. For condition (ii), proceed as follows: The strong lower semicontinuity of f at x provides a sequence $x^\nu \to x$ such that $f(x^\nu) \to f(x)$ with f continuous at x^ν. From Corollary 3.3, it follows that for all ν, $f_\theta(x^\nu) \to f(x^\nu)$. Given any sequence $\theta^k \to 0$ as $k \to \infty$, we need to come up with a sequence x^k such that $f_{\theta^k}(x^k) \to f(x)$. But this follows from the following observation: The set $S := \{ f(x^\nu) \,|\, \nu \in \mathbf{N} \}$ is contained in $\mathrm{Liminf}_k\, S^k$ where $S^k := \{ f_{\theta^k}(x^\nu) \,|\, \nu \in \mathbf{N} \}$; $\mathrm{Liminf}_k\, S^k$ consists of all limits points of all sequences $\{\alpha_k\}_{k \in \mathbf{N}}$ with $\alpha_k \in S_k$. Since $\mathrm{Liminf}_k\, S^k$ is closed and $f(x) \in \mathrm{cl}\, S$, it follows that $f(x) \in \mathrm{Liminf}_k\, S^k$, and that means that there exists $\alpha^k \to f(x)$ with $\alpha^k \in S^k$. These points α^k are the $f_{\theta^k}(x^k)$ we were looking for. □

Theorem 3.7 tells us that if we have to minimize the function f, the averaged functions f_θ could be used in a consistent approximation scheme, i.e., that implies the convergence of the minimizers. However, before we follow this route, we would have to make sure that their properties make them amenable to minimization by existing—or possibly, modified—algorithmic procedures. The remainder of this section is devoted to the continuity and differentiability properties of averaged functions, in particular for the class of Steklov (averaged) functions.

DEFINITION 3.8. *Given a locally integrable function $f : \mathbf{R}^n \to \mathbf{R}$, the* Steklov (averaged) functions *are defined as follows: For $\alpha > 0$,*

$$f_\alpha(x) = \int_{\mathbf{R}^n} f(x - z)\psi_\alpha(z)\, dz,$$

where

$$\psi_\alpha(z) = \begin{cases} 1/\alpha^n & \textit{if } \max_{1,\dots,n} |z_i| \le \alpha/2; \\ 0 & \textit{otherwise.} \end{cases}$$

Equivalently,

$$f_\alpha(x) = \frac{1}{\alpha^n} \int_{x_1 - \alpha/2}^{x_1 + \alpha/2} dy_1 \dots \int_{x_n - \alpha/2}^{x_n + \alpha/2} dy_n\, f(y).$$

This class of averaged functions was introduced by Steklov [39] in 1907, and used by Kolmogorov and Fréchet for compactness tests in $\mathcal{L}^p$. In the context of smooth optimization, they were used by Katkovnik [19], Nikolaeva [27], Gupal [15] and [16], and Mayne and Polak [24].

The next proposition records the well-know fact that Steklov functions are locally Lipschitz continuous.

PROPOSITION 3.9. *For locally bounded and integrable functions $f : \mathbf{R}^n \to \mathbf{R}$, the associated Steklov functions f_α are locally Lipschitz, i.e., on each compact set $K \subset \mathbf{R}^n$,* *the function f_α is Lipschitz continuous on K with Lipschitz constant κ,*

$$\kappa = (2n/\alpha) \sup_{x\in K_\alpha} f(x), \quad \textit{where } K_\alpha := \{\, x+z \,|\, x \in K,\ \max_{i=1,\dots,n} |z_i| \le \alpha/2 \,\}.$$

Differentiability of average functions, however, cannot be guaranteed in general, unless the mollifiers ψ_θ are sufficiently smooth or if f itself has a sufficient level of continuity.

PROPOSITION 3.10 (Sobolev [38], Schwartz [37]). *Let $f : \mathbf{R}^n \to \mathbf{R}$ be locally integrable. Whenever the mollifiers ψ_θ are smooth (of class C^1), so are the associated averaged functions f_θ with gradient*

$$\nabla f_\theta(x) = \int_{\mathbf{R}^n} f(y) \nabla \psi_\theta(x-y)\, dy.$$

PROPOSITION 3.11 (Gupal [15]). *For $f : \mathbf{R}^n \to \mathbf{R}$ continuous, the Steklov (averaged) functions f_α are continuously differentiable, and their gradients are given by*

$$\begin{aligned}
&\nabla f_\alpha(x) \\
&= \sum_{i=1}^n e_i \frac{1}{\alpha^n - 1} \int_{x_1-\alpha/2}^{x_1+\alpha/2} dy_1 \cdots \int_{x_{i-1}-\alpha/2}^{x_{i-1}+\alpha/2} dy_{i-1} \int_{x_{i+1}-\alpha/2}^{x_{i+1}+\alpha/2} dy_{i+1} \cdots \int_{x_n-\alpha/2}^{x_n+\alpha/2} dy_n \\
&\quad \frac{1}{\alpha}[f(y_1,\dots,y_{i-1},x_i - \tfrac{1}{2}\alpha, y_{i+1},\dots y_n) - f(y_1,\dots,y_{i-1},x_i + \tfrac{1}{2}\alpha, y_{i+1},\dots y_n)],
\end{aligned}$$

where e_i is the ith unit coordinate vector.

This gradient can also be expressed as

$$\nabla f_\alpha(x) = \sum_{i=1}^n e_i \int_{-\frac{1}{2}}^{\frac{1}{2}} d\xi_1 \cdots \int_{-\frac{1}{2}}^{\frac{1}{2}} d\xi_{i-1} \int_{-\frac{1}{2}}^{\frac{1}{2}} d\xi_{i+1} \cdots \int_{-\frac{1}{2}}^{\frac{1}{2}} d\xi_n \, \lambda_\alpha(x,\xi),$$

where

$$\begin{aligned}
\lambda_\alpha(x,\xi) =& \frac{1}{\alpha}[f(x_1+\alpha\xi_1,\dots,x_{i-1}+\alpha\xi_{i-1}, x_i + \tfrac{1}{2}\alpha, x_{i+1}+\alpha\xi_{i+1},\dots,x_n+\alpha\xi_n) \\
&- f(x_1+\alpha\xi_1,\dots,x_{i-1}+\alpha\xi_{i-1}, x_i - \tfrac{1}{2}\alpha, x_{i+1}+\alpha\xi_{i+1},\dots,x_n+\alpha\xi_n)].
\end{aligned}$$

This means that $\nabla f_\theta(x)$ is the expectation of the random vector $\boldsymbol{\lambda}_\alpha(x,\boldsymbol{\xi})$, where $\boldsymbol{\xi} = (\boldsymbol{\xi}_1,\dots\boldsymbol{\xi}_n)$ is a random vector whose elements are independent and uniformly distributed on $[-\frac{1}{2},\frac{1}{2}]$. In other words, $\boldsymbol{\lambda}_\alpha(x,\boldsymbol{\xi})$ is an unbiased estimator of the gradient of f_α at x.

Remark 3.12. Although in the case of discontinuous functions f we cannot "reach" differentiability for Steklov functions, it is always possible to do so if the averaging process is repeated a second time. This follows immediately from Propositions 3.9 and 3.11. Given a locally integrable function $f : \mathbf{R}^n \to \mathbf{R}$, let

$$\begin{aligned}
f_{\alpha\beta}(x) :&= \int_{\mathbf{R}^n} f_\alpha(x-z)\psi_\beta(z)\, dz \\
&= \int_{\mathbf{R}^n} dy \int_{\mathbf{R}^n} dz\, f(x-y-z)\psi_\alpha(y)\psi_\beta(z)
\end{aligned}$$

with the densities ψ_α and ψ_β as in Definition 3.8. We can also express this as an expectation,

$$f_{\alpha\beta}(x) = E\{f(x - \alpha\boldsymbol{\xi} - \beta\boldsymbol{\eta})\}$$

with $\boldsymbol{\xi}$ and $\boldsymbol{\eta}$ random vectors whose elements are independent and uniformly distributed on $[-\frac{1}{2}, \frac{1}{2}]$. The gradient can be calculated from Proposition 3.11. We have

$$\begin{aligned}&\nabla f_{\alpha\beta}(x)\\ &\quad = \int_{-\frac{1}{2}}^{\frac{1}{2}} d\xi_1 \cdots \int_{-\frac{1}{2}}^{\frac{1}{2}} d\xi_n \left(\int_{-\frac{1}{2}}^{\frac{1}{2}} d\eta_1 \cdots \int_{-\frac{1}{2}}^{\frac{1}{2}} d\eta_{i-1} \int_{-\frac{1}{2}}^{\frac{1}{2}} d\eta_{i+1} \cdots \int_{-\frac{1}{2}}^{\frac{1}{2}} d\eta_n \, \lambda_{\alpha\beta}(x, \xi, \eta) \right),\end{aligned}$$

where, with $z_i^{\alpha\beta}(\xi, \eta) := x_i - \alpha\xi_i - \beta\eta_i$,

$$\begin{aligned}&\lambda_{\alpha\beta}(x,\xi,\eta)\\ &\quad := \sum_{i=1}^{n} e_i \big[f(z_1^{\alpha\beta}(\xi,\eta), \ldots, z_{i-1}^{\alpha\beta}(\xi,\eta), x_i + \alpha\xi_i + \tfrac{\beta}{2}, z_{i+1}^{\alpha\beta}(\xi,\eta), \ldots, z_n^{\alpha\beta}(\xi,\eta))\\ &\qquad\qquad - f(z_1^{\alpha\beta}(\xi,\eta), \ldots, z_{i-1}^{\alpha\beta}(\xi,\eta), x_i + \alpha\xi_i - \tfrac{\beta}{2}, z_{i+1}^{\alpha\beta}(\xi,\eta), \ldots, z_n^{\alpha\beta}(\xi,\eta)) \big] \beta^{-1}.\end{aligned}$$

Again, $\boldsymbol{\lambda}_{\alpha\beta}(x, \boldsymbol{\xi}, \boldsymbol{\eta})$ is an unbiased estimate of the gradient $\nabla f_{\alpha\beta}(x)$ with $\boldsymbol{\xi}, \boldsymbol{\eta}$ random vectors whose elements are independent and uniformly distributed on $[-\frac{1}{2}, \frac{1}{2}]$. □

Remark 3.13. Let us also record an important relationship between the estimates of the gradients of averaged functions and stochastic gradients. We consider the following averaged functions:

$$f_\theta(x) = \frac{1}{\theta^n} \int_{\mathrm{R}^n} f(z)\psi\left(\frac{x-z}{\theta}\right) dz = \int_{\mathrm{R}^n} f(x - \theta z)\psi(z)\, dz,$$

with f locally integrable, ψ is a density function with compact support and such that $\nabla\psi$ is Lipschitz continuous. Then the gradient of f_θ,

$$\nabla f_\theta = \frac{1}{\theta^{n+1}} \int_{\mathrm{R}^n} f(z)\nabla\psi\left(\frac{x-z}{\theta}\right) dz$$

is locally Lipschitz with constants proportional to $1/\theta^{n+1}$. The random vector (cf. Gupal [16])

$$\boldsymbol{\lambda}_{\theta,\triangle}(x, \boldsymbol{\xi}, \boldsymbol{\eta}) = \frac{1}{\triangle}[f(x - \theta\boldsymbol{\xi} + \triangle\boldsymbol{\eta}) - f(x - \theta\boldsymbol{\xi})]\boldsymbol{\eta}$$

is a stochastic quasi-gradient of f_θ at x (Ermoliev [9]), where $\boldsymbol{\xi}$ is distributed in accordance with the density function ψ, and $\boldsymbol{\eta}$ is a random vector whose elements are independent and uniformly distributed on $[-1, 1]$. To see this, note that

$$\begin{aligned}E^{\xi,\eta}\{\boldsymbol{\lambda}_{\theta,\triangle}(x, \boldsymbol{\xi}, \boldsymbol{\eta})\} &= E^{\eta}\frac{1}{\triangle}[f_\theta(x + \triangle\boldsymbol{\eta}) - f_\theta(x)]\boldsymbol{\eta}\\ &= \frac{2}{3}\nabla f_\theta(x) + \frac{\triangle}{2}O(x, \theta, \triangle),\end{aligned}$$

where $O(x, \theta, \triangle)$ is locally bounded.

Observe also that if $\boldsymbol{\xi}$ is distributed in accordance with the density function ψ_θ and $\boldsymbol{\eta}$ is a random vector whose elements are independent and uniformly distributed on $[-1,1]$, then

$$\boldsymbol{\lambda}_{\theta,\triangle}(x,\boldsymbol{\xi},\boldsymbol{\eta}) = \frac{1}{\triangle}[\, f(x-\boldsymbol{\xi}+\triangle\boldsymbol{\eta}) - f(x-\boldsymbol{\xi})\,]\boldsymbol{\eta}$$

is a quasi-gradient for the averaged function f_θ, i.e., it provides a, possibly biased, estimate of the gradient of f_θ as calculated in Proposition 3.10.

Remark 3.14. To complete this analysis of averaged functions, let us point out that the class of averaged functions that we have introduced is based on convolutions with mollifiers that have bounded support. We could however have worked with a more general class, including mollifiers with unbounded support, and still obtain a convergence result similar to that of Theorem 3.2; in fact, not just eh-convergence, but most of the results in this section. Let $\{\, \varphi_\theta : \mathbf{R}^n \to \mathbf{R}_+, \theta \in \mathbf{R}_+ \,\}$ be a class of integrable functions such that $\int \varphi_\theta(z)\,dz = 1$. Suppose that the function $f : \mathbf{R}^n \to \mathbf{R}$ and the $\{\varphi_\theta\}$ are such that $f_\theta = f \star \varphi_\theta$ is well defined (on $\mathbf{R}^n$) and that for all $\delta > 0$,

$$\lim\nolimits_{\theta\downarrow 0} \int_{|z|>\delta} |f(z)|\varphi_\theta(x-z)\,dz = 0, \quad \text{uniformly in } x, \quad \lim\nolimits_{\theta\downarrow 0} \int_{|z|\le\delta} \varphi_\theta(z)\,dz = 1.$$

To see that the functions f_θ still eh-converge to f, note for all $x \in \mathbf{R}^n$ and $\varepsilon > 0$ there exists V, a neighborhood of 0, such that $f(x-z) \ge \mathrm{cl}_e f(x) - \varepsilon$ for all $z \in V$ and that for $x^\theta \to x$ as $\theta\downarrow 0$, for all $\delta > 0$ and θ sufficiently small,

$$\begin{aligned} f_\theta(x^\theta) &= \int_{|z|\le\delta} f(z)\varphi_\theta(x^\theta - z)\,dz + \int_{|z|>\delta} f(z)\varphi_\theta(x^\theta - z)\,dz \\ &\ge (\mathrm{cl}_e f(x) - \varepsilon) \int_{|z|\le\delta} \varphi_\theta(z)\,dz - \frac{\varepsilon}{2} \end{aligned}$$

and hence $\liminf_{\theta\downarrow 0} f_\theta(x^\theta) \ge \mathrm{cl}_e f(x)$ (after letting $\varepsilon\downarrow 0$). For example, let φ be the Gaussian density function, i.e.,

$$\varphi(y) = \frac{1}{(2\pi)^{n/2}} e^{-|y|^2}.$$

Consider the following family of functions:

$$f_\theta(x) = \frac{1}{\theta^n} \int_{\mathbf{R}^n} f(y)\, \varphi\left(\frac{x-y}{\theta}\right) dy, \qquad \theta > 0.$$

Suppose that $|f(x)| \le \gamma_1 + \gamma_2 |x|^{\gamma_3}$ with γ_1, γ_2, γ_3 positive constants. Then the functions f_θ eh-converge to f as $\theta\downarrow 0$ and each functions f_θ is pf class C^∞. We have

$$\nabla f_\theta(x) = \frac{1}{\theta^{n+2}} \int_{\mathbf{R}^n} f(x-y) \varphi\left(\frac{y}{\theta}\right) dy = \frac{1}{\theta} \int_{\mathbf{R}^n} [\, f(x-\theta z) - f(x)\,] z \varphi(z)\,dz;$$

passing differentiation under the integral sign is justified by the theory of tempered distributions, cf. Schwartz [37]. Thus the random vector $\boldsymbol{\lambda}_\theta(x,\boldsymbol{\xi})$, defined by

$$\lambda_\theta(x,\xi) = \frac{1}{\theta}[\, f(x-\theta\xi) - f(x)\,]\xi$$

with $\boldsymbol{\xi}$ a Gaussian random variable (density φ), is an unbiased statistical estimator of $\nabla f_\theta(x)$.

4. Mollifier subgradients. We are going to exploit the fact that averaged functions determine an epi-convergent family of approximating functions, and that rather explicit expressions can be obtained for their gradients, to define a new notion of subgradient based on a family of mollifiers. In the next section, these subgradients are used to design minimization procedures aimed, in particular, at the minimization of discontinuous functions.

DEFINITION 4.1. *Let $f : \mathbf{R}^n \to \mathbf{R}$ be locally integrable and let $\{f^\nu := f_{\theta^\nu}, \nu \in \mathbf{N}\}$ be a sequence of averaged functions obtained from f by convolution with the sequence of mollifiers $\{\psi^\nu := \psi_{\theta^\nu} : \mathbf{R}^n \to \mathbf{R}_+, \nu \in \mathbf{N}\}$ where $\theta^\nu \downarrow 0$ as $\nu \to \infty$. Assume that the mollifiers are such that the averaged functions f^ν are smooth (of class C^1), as would be the case if the mollifiers ψ^ν are smooth. The* subgradient set of ψ-mollifier *of f at x is*

$$\partial_\psi f(x) := \operatorname{Limsup}_{\nu\to\infty}\{\nabla f^\nu(x^\nu) \mid x^\nu \to x\},$$

i.e., the cluster points of all possible sequences $\{\nabla f^\nu(x^\nu)\}$ such that $x^\nu \to x$. The full Ψ- mollifier subgradient set *is*

$$\partial_\Psi f(x) := \bigcup\nolimits_\psi \partial_\psi f(x),$$

where ψ ranges over all possible sequences of mollifiers that generate smooth averaged functions.

The set $\partial_\psi f(x)$ of ψ-mollifier subgradients is closed, and in general, depends on the choice of the sequence $\{\psi^\nu\}$ used in its construction. The full mollifier subgradient set $\partial_\Psi f(x)$ clearly does not depend on any particular choice of mollifiers. The sets $\partial_\psi f(x)$ and $\partial_\Psi f(x)$ are always nonempty if the function f is almost everywhere smooth and its gradient is locally bounded on the set where it exists (as in Corollary 3.3, but applied here to ∇f).

DEFINITION 4.2. *Let $f : \mathbf{R}^n \to \mathbf{R}$ be locally integrable and let $\{f^\nu := f_{\theta^\nu}, \nu \in \mathbf{N}\}$ be a sequence of averaged functions obtained from f by convolution with the sequence of mollifiers $\{\psi^\nu := \psi_{\theta^\nu} : \mathbf{R}^n \to \mathbf{R}_+, \nu \in \mathbf{N}\}$ where $\theta^\nu \downarrow 0$ as $\nu \to \infty$. Assume that the mollifiers are such that the averaged functions f^ν are smooth (of class C^1), as would be the case if the mollifiers ψ^ν are smooth (of class C^1). The ψ-*mollifier subderivative *of f at x in direction u is*

$$f'_\psi(x; u) := \mathrm{h{-}ls}\,(f^\nu)'(x; u) = \sup_{\{x^\nu \to x\}} \limsup_{\nu\to\infty} (f^\nu)'(x^\nu; u),$$

where $(f^\nu)'(x; u)$ is the derivative of f^ν at x in direction u; sup *is taking with respect to all sequences $x^\nu \to x$. The* full (Ψ-)mollifier subderivative *of f at x in direction u is*

$$f'_\Psi(x; u) := \sup_\psi f'_\psi(x; u),$$

where ψ ranges over all possible sequences of mollifiers generating smooth averaged functions.

Henceforth, when referring to f we always assume that it is locally integrable and that $\{f^\nu\}$ is a sequence of smooth averaged functions obtained from f by convolution with a sequence of mollifiers $\{\psi^\nu, \nu \in \mathbf{N}\}$.

PROPOSITION 4.3. *The ψ-mollifier subgradient mapping $x \mapsto \partial_\psi f(x)$ is outer semicontinuous (closed graph) and f'_ψ is upper semicontinuous. Also*

$$\begin{aligned} f'_\psi(x;u) &\geq \sup\{\,\langle g,u\rangle \mid g \in \partial_\psi f(x)\,\},\\ f'_\Psi(x;u) &\geq \sup\{\,\langle g,u\rangle \mid g \in \partial_\Psi f(x)\,\}. \end{aligned}$$

Proof. The proof follows directly from the definitions.

PROPOSITION 4.4. *The function $u \mapsto f'_\psi(x;u)$ is sublinear, i.e., $f'_\psi(x;\cdot)$ is convex and positively homogeneous. The set-valued mapping*

$$x \mapsto G_\psi(x) := \{\, g \in \mathbf{R}^n \mid \langle g,u\rangle \leq f'_\psi(x;u),\ \forall u \in \mathbf{R}^n \,\}$$

is closed-, convex-valued.

Proof. Since the functions f^ν are smooth, we have

$$(f^\nu)'(x^\nu;u_1+u_2) = (f^\nu)'(x^\nu;u_1) + (f^\nu)'(x^\nu;u_2).$$

Taking lim sup on both sides over all sequences $x^\nu \to x$ yields

$$f'_\psi(x;u_1+u_2) \leq f'_\psi(x;u_1) + f'_\psi(x;u_2).$$

Similarly, the positive homogeneity of $f'_\psi(x;\cdot)$ follows from the linearity of the derivatives of the functions $(f^\nu)'(x;\cdot)$. The assertions about the set-valued mapping G_ψ follow directly from the sublinearity of $f'_\psi(x;\cdot)$. □

PROPOSITION 4.5. *We always have*

$$\operatorname{con} \partial_\psi f(x) \subset G_\psi(x) := \{\, g \in \mathbf{R}^n \mid \langle g,u\rangle \leq f'_\psi(x;u),\ \forall u \in \mathbf{R}^n \,\},$$

where con *denotes the convex hull. If $G_\psi(x)$ is bounded, then* $\operatorname{con}\partial_\psi f(x) = G_\psi(x)$.

Proof. We begin with the inclusion. To any $g \in \partial_\psi f(x)$, there corresponds a subsequence $\{\nu_k\} \subset \{\nu\}$ and $x^k \to x$ such that $\nabla f^{\nu_k}(x^k) \to g$. Since $(f^{\nu_k})'(x^k;u) = \langle \nabla f^{\nu_k}(x^k), u\rangle$, it follows that

$$\langle g,u\rangle = \lim_{k\to\infty} \langle \nabla f^{\nu_k}(x^k), u\rangle = \lim_{k\to\infty} (f^{\nu_k})'(x^k;u) \leq f'_\psi(x;u).$$

Thus $\partial_\psi f(x) \subset G_\psi(x)$ and the convexity of $G_\psi(x)$ then yields $\operatorname{con}\partial_\psi f(x) \subset G_\psi(x)$.

Suppose now that $G_\psi(x)$ is bounded. If $h \in G_\psi(x) \setminus \operatorname{con}\partial_\psi f(x)$, i.e., $G_\psi(x) \not\subset \operatorname{con}\partial_\psi f(x)$, then by the separation theorem for convex sets, there exists $\bar u$ such that $\langle h,\bar u\rangle > \langle g,\bar u\rangle$ for all $g \in \operatorname{con}\partial_\psi f(x)$. But $f'_\psi(x;\bar u) \geq \langle h,\bar u\rangle$ and, passing to a subsequence whenever necessary, there exists $x^\nu \to x$ so that

$$\nabla f^\nu(x^\nu) \longrightarrow g \in \partial_\psi f(x)$$

and

$$(f^\nu)'(x^\nu;\bar u) = \langle \nabla f^\nu(x^\nu), \bar u\rangle \longrightarrow f'_\psi(x;\bar u).$$

Thus, we would have

$$f'_\psi(x;\bar u) = \langle g,\bar u\rangle \geq \langle h,\bar u\rangle > \langle g,\bar u\rangle,$$

clearly contradicting the existence of such an h. □

Remark 4.6. The approach laid out here could be used to define subdifferentials of higher order. For example, if the mollifiers ψ_{θ^ν} are of class C^2, then the resulting averaged function f^ν are also twice continuously differentiable. With $\nabla^2 f^\nu(x)$ the Hessian of f^ν at x, we could define the second-order ψ-mollifier subhessian of f at x as

$$\partial^2_\psi f(x) := \mathrm{Limsup}_{\nu\to\infty}\{\, \nabla^2 f^\nu(x^\nu) \mid x^\nu \to x \,\},$$

i.e., the cluster points of all possible sequences $\{\nabla^2 f^\nu(x^\nu)\}$ of matrices with $x^\nu \to x$. The function

$$f''_\psi(x;H) := \limsup_{x^\nu\to x} \langle \nabla^2 f^\nu(x^\nu), H\rangle = \limsup_{x^\nu \to x} \sum_{i,j=1}^{n} \frac{\partial}{\partial x_i \partial x_j} f^\nu(x^\nu) h_{ij}$$

could be called the second-order ψ-mollifier subderivative of f in direction H. The mapping $x \mapsto \partial^2_\psi f(x)$ is closed, the function $f''_\psi(x;\cdot)$ is upper semicontinuous, and we have

$$\operatorname{con} \partial^2_\psi f(x) = \{\, H \in \mathbf{R}^{n^2} \mid Hu \le f''_\psi(x;U),\ \forall U \in \mathbf{R}^{n^2} \,\}.$$

The next theorem justifies a minimization approach based on mollifier subgradients.

THEOREM 4.7. *Suppose that $f : \mathbf{R}^n \to \mathbf{R}$ is strongly lower semicontinuous and locally integrable. Then, for any sequence $\{\psi^\nu\}$ of smooth mollifiers, we have*

$$0 \in \partial_\psi f(x) \quad \textit{whenever } x \textit{ is a local minimizer of } f.$$

Proof. Let x be a local minimizer of f. For V a compact neighborhood of x sufficiently small, define

$$\varphi : V \to \mathbf{R} \quad \text{with } \varphi(z) = f(z) + |z-x|^2.$$

The function φ achieves its global minimum (on V) at x. Consider also the averaged functions

$$\varphi^\nu(z) = \int_{\mathbf{R}^n} \varphi(y-z)\psi^\nu(y)\,dy = f^\nu(z) + \beta^\nu(x,z),$$

where $\beta^\nu(x,z) = \int |y-z-x|^2 \psi^\nu(y)\,dy$. From Theorem 3.10, it follows that the functions φ^ν are continuously differentiable and Theorem 3.7 implies that they epi-converge to φ on V. Suppose φ^ν achieves its minimum at some point $z^\nu \in V$. It follows from Theorem 3.5 that $z^\nu \to x$, and thus

$$\nabla \varphi^\nu(z^\nu) = \nabla f^\nu(z^\nu) + \nabla \beta^\nu(x, z^\nu) = 0.$$

Hence

$$\nabla f^\nu(z^\nu) = -\nabla \beta^\nu(x,z^\nu) \longrightarrow 0 \quad \text{as} \quad \nu \to \infty,$$

and consequently as $0 \in \partial_\psi f(x)$. □

Although it is not really part of the objectives of this development to obtain expressions that can be manipulated by classical means, the following example might nevertheless help render more concrete some of the results and operations discussed so far.

Example 4.8. Consider the function

$$f(z) = \begin{cases} |z+1| & \text{if } z < 0, \\ (z-1) & \text{if } z \ge 0, \end{cases}$$

and the family of mollifiers $\{\psi_\theta, \theta > 0\}$, with

$$\psi_\theta(z) = \begin{cases} (1/4\theta^3)(z+2\theta)^2 & \text{if } z \in [\,-2\theta, -\theta\,], \\ (1/4\theta^3)(2\theta^2 - z^2) & \text{if } z \in [\,-\theta, \theta\,], \\ (1/4\theta^3)(z-2\theta)^2 & \text{if } z \in [\,\theta, 2\theta\,], \\ 0 & \text{otherwise.} \end{cases}$$

This is a family of smooth mollifiers and f is strongly lower semicontinuous. This means that the averaged functions f_θ are smooth (Proposition 3.10), and that they epi-converge to f (Theorem 3.7). Assuming that $\theta < 1/4$, from the formula for the gradient of f_θ in Proposition 3.10, we have

$$\nabla f_\theta(x) = \begin{cases} -1 & \text{if } x < -1 - 2\theta, \\ (1/6\theta^3)[\,x^3 + 3(1+2\theta)x^2 + 3(1+4\theta+4\theta^2)x & \\ \qquad + (1 + 6\theta + 12\theta^2 + 2\theta^3)\,] & \text{if } x \in [-1-2\theta, -1-\theta), \\ (1/6\theta^3)[\,-x^3 - 3x^2 + 3(2\theta^2 - 1)x + 6\theta^2 - 1\,] & \text{if } x \in [-1-\theta, -1+\theta), \\ (1/6\theta^3)[\,x^3 + 3(1-2\theta)x^2 + 3(1-4\theta+4\theta^2)x & \\ \qquad + (1 - 6\theta + 12\theta^2 - 2\theta^3)\,] & \text{if } x \in [-1+\theta, -1+2\theta), \\ 1 & \text{if } x \in [-1+2\theta, -2\theta), \\ (1/2\theta^3)[-x^2 - 4\theta x + 2\theta^2(\theta - 2)] & \text{if } x \in [-2\theta, -\theta), \\ (1/2\theta^3)[x^2 + 2\theta^2(\theta - 1)] & \text{if } x \in [-\theta, \theta), \\ (1/2\theta^3)[-x^2 + 4\theta x + 2\theta^2(\theta - 2)] & \text{if } x \in [\theta, 2\theta), \\ 1 & \text{if } x \geq 2\theta. \end{cases}$$

Letting $\theta \downarrow 0$ leads to the following expression for the mollifier subgradient associated with the family $\{\psi_\theta, \theta > 0\}$:

$$\partial_\psi f(x) = \begin{cases} -1 & \text{if } x < -1, \\ [-1, 1] & \text{if } x = 1, \\ 1 & \text{if } x \in (-1, 0), \\ [-\infty, 1] & \text{if } x = 0, \\ 1 & \text{if } x > 0. \end{cases}$$

The point $x = -1$ is a local minimizer, and $x = 0$ is a global minimizer. In both of these cases, $0 \in \partial_\psi f(x)$ as asserted by Theorem 4.7. Observe also that in the "convex" portions of the function f, the mollifier subgradient coincides with the subgradient from convex analysis; Remark 4.12 indicates that this is always the case when f is convex. □

In the remainder of this section we explore the relationship between the mollifier subgradient and some other subgradients notions.

For function $f : \mathbf{R}^n \to \mathbf{R}$ continuous on a neighborhood V of x, Warga [42]–[44] defines subgradients of f at x as follows: Let $\{f^k, k \in \mathbf{N}\}$ be a sequence of smooth functions converging uniformly to f on V, we refer to

$$\partial_W f(x) = \bigcap_{j=1}^{\infty} \bigcap_{\delta > 0} \text{cl} \left[\bigcup_{k \geq j, |x-y| \leq \delta} \nabla f^k(y) \right]$$

as the set of Warga-subgradients of f at x (cl denotes closure).

PROPOSITION 4.9. *For $f : \mathbf{R}^n \to \mathbf{R}$ be continuous on V a neighborhood of x and $\{f^k, k \in \mathbf{N}\}$ a sequence of smooth functions converging uniformly to f on V, then*

$$\partial_W f(x) = \text{Limsup}_{k\to\infty}\{\,\nabla f^k(x^k) \,|\, \forall x^k \to x\,\}.$$

Consequently, when f is continuous, $\partial_W f(x)$ coincides with $\partial_\psi f(x)$ if in the construction of $\partial_W f(x)$ the f^k are averaged functions generated by the sequence of smooth mollifiers $\{\psi^k\}$.

Proof. Let

$$D(x) = \mathrm{Limsup}_{k\to\infty}\{\,\nabla f^k(x^k)\,|\,\forall x^k \to x\,\}.$$

Let us first show that $D(x) \subset \partial_W f(x)$. Let $g \in D(x)$ be such that, passing to a subsequence if necessary, $g = \lim_k \nabla f^k(x^k)$ for some specific sequence $x^k \to x$. We have to show that for all j and $\delta > 0$,

$$g \in G_{j,\delta}(x) := \mathrm{cl}\left[\bigcup_{k\geq j, |x-y|\leq\delta} \nabla f^k(y)\right].$$

Obviously, if $k \geq j$ and $|x^k - x| \leq \delta$, then

$$\nabla f^k(x^k) \in G_{j,\delta}(x).$$

Since $G_{j,\delta}(x)$ is closed, each cluster point of the sequence $\{\nabla f^k(x^k)\}$ belongs to $G_{j,\delta}(x)$. Hence, $g \in \partial_W f(x)$ and $D(x) \subset \partial_W f(x)$.

To prove the converse inclusion, we must show that for each point g in $\partial_W f(x)$ we can find a sequence $x^k \to x$ such that $\nabla f^k(x^k) \to g$. By definition of ∂_W for all j and $\delta > 0$, $g \in G_{j,\delta}(x)$. Let us choose a sequence $\delta_j \downarrow 0$ as $j \to \infty$. Since $g \in G_{j,\delta_j}(x)$ for all j,

$$g \in \mathrm{cl}\{\,\nabla f^k(y)\ :\ k \geq j, |y - x| \leq \delta_j\,\}.$$

Thus in this set, there exists an element $g^j = \nabla f^{k_j}(y^j)$ such that $|g^j - g| < 1/j$. Clearly, $y^j \to x$, $k_j \to \infty$ and $g^j \to g$, so that $g \in D(x)$ and $\partial_W f(x) \subset D(x)$.

The equality between the Warga- and the ψ-mollifier subgradient sets then follow from the formula we just proved, and the definition of ψ-mollifier subgradients. □

In variational analysis, the *Clarke subderivative* of a function $f : \mathbf{R}^n \to \mathbf{R}$ is

$$(d_C f)(x; u) = \limsup_{y\to x, \lambda\downarrow 0} \frac{1}{\lambda}[\,f(y + \lambda u) - f(y)\,]$$

with the lim sup calculated with respect to all sequences $y \to x$, $\lambda \downarrow 0$. The set of *generalized* (*Clarke*) *subgradients* is

$$\partial_C f(x) = \{\,g \in \mathbf{R}^n\,|\,\langle g, u\rangle \leq d_C f(x; u),\ \forall\, u \in \mathbf{R}^n\,\}.$$

This notion was proposed by Clarke [6] for locally Lipschitz continuous functions; for lower semicontinuous functions this notion needs further adjustments (consult Rockafellar [31]).

PROPOSITION 4.10. *For $f : \mathbf{R}^n \to \mathbf{R}$ locally integrable, we have $f'_\psi(x;\cdot) \leq d_C f(x;\cdot)$. If f is also continuous, then $f'_\psi(x;\cdot) = d_C f(x;\cdot)$.*

Proof. By definition of $d_C f(x; u)$ it follows that for an arbitrary $\varepsilon > 0$, there exist δ_1, δ_2 such that whenever $|y - x| < \delta_1$ and $\lambda \in (0, \delta_2)$,

$$\frac{1}{\lambda}\,[\,f(y + \lambda u) - f(y)\,] < d_C f(x; u) + \varepsilon.$$

Let f^ν be the averaged function obtained as the convolution of f and the mollifier ψ^ν. Consider the finite differences

$$\triangle_\nu(y,u,\lambda) := \frac{1}{\lambda}\,[\,f^\nu(y+\lambda u) - f^\nu(y)\,] = \int_{\mathrm{R}^n} \frac{1}{\lambda}\,[\,f(y-z+\lambda u) - f(y-z)\,]\psi^\nu(z)\,dz.$$

If $|y-x| < \delta_1/2$, $\lambda < \delta_2/2$, and $|z| \le \delta_1/2$, then

$$\triangle_\nu(y,u,\lambda) \le (d_C f(x;u) + \varepsilon)\int_{|z|\le\delta_1/2} \psi^\nu(z)\,dz.$$

Thus for y close enough to x,

$$(f^\nu)'(y;u) = \lim_{\lambda\downarrow 0} \triangle_\nu(y,u,\lambda) \le (d_C f(x;u)+\varepsilon)\int_{|z|\le\delta_1/2}\psi^\nu(z)\,dz$$

from which, after letting $\varepsilon\downarrow 0$, it follows that $f'_\psi(x;u) \le d_C f(x;u)$.

We next set out to prove the reverse inequality, assuming that f is continuous. Let $x^\nu \to x$ and $\lambda_\nu \downarrow 0$ be such that

$$d_C f(x;u) = \lim_{\nu\to\infty} \frac{1}{\lambda_\nu}\,[\,f(x^\nu + \lambda_\nu u) - f(x^\nu)\,].$$

From Corollary 3.3, we know that when f is continuous, the averaged functions f^ν converge uniformly to f on some neighborhood, say V, of x. Thus, with $\varepsilon_\nu = \lambda_\nu/\nu$, we can always find k_ν such that

$$\sup_{y\in V} |f(y) - f^{k_\nu}(y)| < \varepsilon_\nu.$$

Now from the Mean Value Theorem follows the existence of $y^\nu := x^\nu + \tau_\nu u$, $\tau_\nu \in [\,0,\lambda_\nu\,]$ such that

$$\frac{1}{\lambda_\nu}\,[\,f^{k_\nu}(x^\nu+\lambda_\nu u) - f^{k_\nu}(x^\nu)\,] = (f^{k_\nu})'(y^\nu;u).$$

Thus for ν sufficiently large, with $x^\nu \in V$ and $x^\nu + \lambda_\mu u \in V$, we have

$$\begin{aligned} f(x^\nu+\lambda_\nu u) - f(x^\nu) &= [\,f^{k_\nu}(x^\nu+\lambda_\nu u) - f^{k_\nu}(x^\nu)\,] \\ &\quad + [\,f(x^\nu+\lambda_\nu u) - f^{k_\nu}(x^\nu+\lambda_\nu u)\,] - [\,f(x^\nu) - f^{k_\nu}(x^\nu)\,] \\ &\le \lambda_\nu((f^{k_\nu})'(y^\nu;u) + 2/\nu). \end{aligned}$$

Taking lim sup with respect to ν yields

$$d_C f(x;u) \le \limsup_{\nu\to\infty}\,(f^{k_\nu})'(y^\nu;u) \le f'_\psi(x;u),$$

which completes the proof. □

THEOREM 4.11. *If $f : \mathbf{R}^n \to \mathbf{R}$ is lower semicontinuous and locally integrable, then*

$$\operatorname{con}\partial_\psi f(x) \subset \operatorname{con}\partial_\Psi f(x) \subset \partial_C f(x).$$

If, in addition, f is locally Lipschitz continuous, then

$$\operatorname{con}\partial_\psi f(x) = \partial_\Psi f(x) = \partial_C f(x).$$

Proof. The relationship between $\partial_\psi f(x)$ and $\partial_\Psi f(x)$ yields the first inclusion, and the second inclusion follows from the preceding proposition. If f is locally Lipschitz, then also the averaged functions f^ν are equi-Lipschitz and $\partial_\psi f(x)$ is bounded. Equality then follows from Propositions 4.5 and 4.10. □

COROLLARY 4.12 (Gupal [15]). *If f is locally Lipschitz continuous, then for all $\alpha_\nu \downarrow 0$ and $x^\nu \to x$, all clusters points of the sequences $\{\nabla f_{\alpha_\nu}(x^\nu)\}$ belong to $\partial_C f(x)$.*

Remark 4.13. For the sake of completeness, let us also record the fact that for convex functions $f : \mathbf{R}^n \to \mathbf{R}$, any ψ-mollifier subgradient is equal to the subgradient of convex analysis *without* any need of taking convex hulls. More precisely, we have

$$\partial_\psi f(x) = \partial f(x) := \{\, g \in \mathbf{R}^n \mid f(z) \geq f(x) + \langle g, z - x\rangle,\ \forall\, z \in \mathbf{R}^n \,\}.$$

As is well known, convex functions of gradients can be characterized in terms of the expression on the right, cf. [32], for example. In view of the preceding theorem, it will thus be sufficient to show that if $g \in \partial f(x)$, then g is also included in $\partial_\psi f(x)$. Let us consider the function

$$\varphi(y) = f(y) + |y - x|^2 - f(x) - \langle g, y - z\rangle.$$

The function $\varphi \geq 0$ and attains its minimum $(= 0)$ at x; due to the strict convexity of φ, x is a unique minimizer of φ. Let

$$\begin{aligned}\varphi^\nu(y) &= \int_{\mathbf{R}^n} \varphi(y - z)\psi^\nu(z)\,dz \\ &= f^\nu(y) + \beta^\nu(x, y) - f(x) - \langle g, y - x\rangle - \int_{\mathbf{R}^n} \langle g, z\rangle \psi^\nu(z)\,dz\end{aligned}$$

be the averaged functions associated with φ by convolution with ψ^ν; here $\beta^\nu(x, y) = \int |y - z - x|^2 \psi^\nu(z)\,dz$. The averaged functions φ^ν uniformly converge to φ on some neighborhood V of x (Corollary 3.3). Due to the strict convexity of φ, for ν sufficiently large, the averaged functions ψ^ν have a (global) minimizer on V, say y^ν. Moreover, $y^\nu \to x$, since x is a unique minimizer of $\varphi = \text{e}-\text{lm}\,\varphi^\nu$ (Theorem 3.7). The averaged functions φ^ν, f^ν, and $\beta^\nu(x, \cdot)$ are smooth (Theorem 3.10), and thus

$$\nabla\varphi^\nu(y^\nu) = \nabla f^\nu(y^\nu) + \nabla_y \beta^\nu(x, y^\nu) - g,$$

$$\nabla_y \beta^\nu(x, y) = \int_{\mathbf{R}^n} \nabla_y |y - z - x|^2 \psi^\nu(z)\,dz = 2(y - x) - 2\gamma^\nu,$$

$$\gamma^\nu = \int_{\mathbf{R}^n} z\psi^\nu(z)\,dz.$$

From the conditions imposed on the mollifiers ψ^ν, it follows that $\gamma^\nu \to 0$, and hence $\nabla_y \beta^\nu(x, y^\nu) \to 0$, and

$$\nabla f^\nu(y^\nu) = g - \nabla_y \beta^\nu(x, y^\nu) \longrightarrow g \quad \text{as } \nu \to \infty,$$

which means that $g \in \partial_\psi f(x)$, as claimed. □

5. Numerical procedures. Let us consider the problem of minimizing a strongly lower semicontinuous φ on X, a compact subset of $\mathbf{R}^n$. Let

$$\mathbb{1}_X(x) = \begin{cases} 1 & \text{if } x \in X; \\ 0 & \text{if } x \notin X. \end{cases}$$

Then, instead of the original problem, we could work with one of the following unconstrained problems involving discontinuous penalty functions:

$$\text{minimize} f(x) := \varphi(x)\mathbb{I}_X(x) + \gamma(1 - \mathbb{I}_X(x))$$

or

$$\text{minimize} f(x) := \varphi(x)\mathbb{I}_X(x) + \gamma(1 - \mathbb{I}_X(x))d(x, X),$$

where $d(x, X) = \min\{\, |x - y| : y \in X \,\}$ and γ is sufficiently large.

If the function φ is bounded on X and $\gamma > \sup\{|\varphi(x)| : x \in X\}$, all local minima of φ on X are also local minima of the function f.

Assuming that f is also strongly lower semicontinuous, in view of Theorems 3.7 and 3.10, we can always find a sequence of smooth averaged functions f^ν (generated by mollifiers $\{\psi^\nu\}$) that epi-converge to f, and by Theorem 4.7, the condition $0 \in \partial_\psi f(x^*)$ is necessary for a point x^* to be a local minimizer of f

Let us now consider some optimization procedures for f making use of the approximating averaged function f^ν.

Method 5.1. Suppose a sequence $\{x^\nu\}$ of global minimizers of f^ν can be calculated. Then, according to Theorem 3.5 any cluster point of such a sequence is a (global) minimizer of f.

However finding global minimizers of the f^ν could be quite complicated. Let us thus consider the next method.

Method 5.2. Here a sequence of approximating solutions $\{x^\nu\}$ is built in accordance with the following rule. Each function f^ν is minimized—initiating the procedure at $x^{\nu-1}$—until a point x^ν is found such that $|\nabla f^\nu(x^\nu)| \le \varepsilon_\nu$, where $\varepsilon_\nu \downarrow 0$; the starting point x^0 is chosen arbitrarily. In this method, if $\bar{x}$ is a cluster point of the sequence $\{x^\nu\}$, then by the definition of $\partial_\psi f(\bar{x})$, passing to a subsequence if necessary,

$$\lim_{\nu\to\infty} \nabla f^\nu(x^\nu) = 0 \in \partial_\psi f(\bar{x}).$$

Moreover, this would also mean that $0 \in \partial_C f(\bar{x})$ (Theorem 4.11), i.e., $d_C f(x; u) \ge 0$ for all $u \in \mathbf{R}^n$.

This approach requires estimates of $|\nabla f^\nu(x^\nu)|$ during the iteration process. In general, this could be computationally expensive involving the calculation of multidimensional integrals. We can however, produce these estimates in parallel with the optimization process by a well-known averaging procedure (cf. Ermoliev [8]): Let

(i) x^0, z^0 be chosen arbitrarily in $\mathbf{R}^n$;
(ii) $x^{k+1} = x^k - \rho_k z^k, \quad k = 0, 1, \ldots;$
(iii) $z^{k+1} = z^k - \tau_k(z^k - \lambda_k(x^k)), \; k = 0, 1, \ldots;$

where x^k approximates $\operatorname{argmin} f^\nu$, z^k are averaged estimates of $\nabla f^\nu(x^k)$, $\lambda_\nu(x^k)$ are stochastic (finite-difference unbiased) estimates for $\nabla f^\nu(x^k)$ such that their mathematical expectation $E\{\lambda_\nu(x^k)\} = \nabla f^\nu(x^k)$ (see the observations that follow Proposition 3.11), and $\rho_k \ge 0$ and $\tau_k > 0$ are sequences such that

$$\sum_{k=0}^{\infty} \rho_k = \infty, \quad \sum_{k=0}^{\infty} \rho_k^2 < \infty, \quad \lim_{k\to\infty} \rho_k/\tau_k = 0.$$

PROPOSITION 5.3 (Ermoliev [8, theorem V.8]). *If the sequences $\{x^k\}$, $\{z^k\}$ are almost surely bounded, then almost surely*

$$\lim_{k\to\infty} |z^k - \nabla f^\nu(x^k)| = 0, \quad \textit{and } x^k \longrightarrow \{x \,|\, \nabla f^\nu(x) = 0\}.$$

Thus in Method 5.2, *we can proceed with the minimization of each f^ν until the estimate z^k of the gradient of $\nabla f^\nu(x^k)$ satisfies the condition $|z^k| \leq \varepsilon_\nu$.*

Method 5.4. A sequence of approximate solutions x^ν is generated by the following rule:

(i) $x^0 \in \mathbf{R}^n$ is chosen arbitrarily;

(ii) $x^{\nu+1} = x^\nu - \rho_\nu \lambda_\nu(x^\nu)$, $\nu = 0, 1. \dots,$

where $\boldsymbol{\lambda}_\nu(x^\nu)$ is a stochastic (finite-difference unbiased) estimator for $\nabla f^\nu(x^\nu)$ with expectation $E\{\boldsymbol{\lambda}_\nu(x^\nu)\} = \nabla f^\nu(x^\nu)$ (see the observations following Proposition 3.11 and Remark 3.12), and $\rho_\nu \geq 0$ is a deterministic sequence of multipliers.

This method combines ideas from the method of stochastic quasi-gradients with those of dynamic nonstationary optimization techniques, see Ermoliev and Nurminski [10] and Gaivoronski [13]. The following theorem is an example of the possible convergence results.

THEOREM 5.5 (Gupal and Norkin [17]). *Suppose the gradient estimates are those in Example* 3.12, *i.e., $\lambda_\nu(x) = \lambda_{\alpha_\nu \alpha_\nu}(x, \xi, \eta)$, the sequence $\{x^\nu\}$ belongs to some compact set, and $\rho_\nu \geq 0$, α_ν satisfy the conditions*

$$\sum_{\nu=1}^{\infty} \rho_\nu = \infty, \quad \sum_{\nu=1}^{\infty} \left(\frac{\rho_\nu}{\alpha_\nu^2}\right)^2 < \infty, \quad \lim_{\nu \to \infty} \alpha_\nu = \lim_{\nu \to \infty} \frac{\alpha_\nu - \alpha_{\nu+1}}{\alpha_\nu \rho_\nu} = 0.$$

Then, almost surely, the sequence $\{x^\nu\}$ admits a cluster point x^* such that $0 \in \partial_\psi f(x^*)$.

Example 5.6. Let us consider the following minimization of a probability function:

$$f(x) = \mathbf{P}\,[\,g(x, \omega) \geq 0\,].$$

We can express f as a mathematical expectation

$$f(x) = \int_\Omega \mathbb{1}_{\{g(x,\omega) \geq 0\}}(\omega)\, P(d\omega).$$

Since the function $\mathbb{1}_{\{.\}}$ is discontinuous, the function f will in general, not be differentiable. To estimate $f(x)$ and its "gradient," Tamm [41] and Lepp [21] proposed the use of Parzen–Rosenblatt kernel-type estimates [29], [35]:

$$f_\varepsilon(x) = \frac{1}{\varepsilon} \int_\Omega P(d\omega) \int_{-\infty}^{0} d\tau\, \psi\left(\frac{\tau + g(x, \omega)}{\varepsilon}\right),$$

$$\nabla f_\varepsilon(x) = \frac{1}{\varepsilon} \int_\Omega \psi\left(\frac{g(x, \omega)}{\varepsilon}\right) \nabla_x g(x, \omega)\, P(d\omega),$$

where ψ is some symmetric density function on $[-\infty, \infty]$; more recently Marti [23] has suggested a similar approach to deal with reliability constraints in structural optimization. The function f_ε can also be written as

$$f_\varepsilon(x) = \int_\Omega \psi_\varepsilon(g(x, \omega))\, P(d\omega),$$

where

$$\begin{aligned} \psi_\varepsilon(t) &= \frac{1}{\varepsilon} \int_{-\infty}^{t} \psi\left(\frac{\tau}{\varepsilon}\right) d\tau \\ &= \frac{1}{\varepsilon} \int_{-\infty}^{\infty} \mathbb{1}_{\{t-\tau \geq 0\}}(\tau) \psi\left(\frac{\tau}{\varepsilon}\right) d\tau = \frac{1}{\varepsilon} \int_{-\infty}^{\infty} \mathbb{1}_{\{t+\tau \geq 0\}}(\tau) \psi\left(-\frac{\tau}{\varepsilon}\right) d\tau. \end{aligned}$$

Thus ψ_ε is an averaged function (with base function $1\!\!1_{\{\cdot\geq 0\}}$). Instead of the original function f, we have a sequence of approximating functions f_ε constructed (indirectly) by means of averaged functions. Tamm [41] in the differentiable case, and Norkin [28] in the continuous nondifferentiable case, provided conditions under which f_ε converges uniformly to f, and they proposed methods, similar to Method 5.2, to minimize f making use of the approximating functions f_ε. Lepp [22] and Roenko [34] analyzed stochastic iterative methods, like Method 5.4, for the minimization f when it is differentiable, using statistical estimates for $\nabla f_\varepsilon(x)$.

REFERENCES

[1] Z. Artstein and R. J-B. Wets, *Stability results for stochastic programs and sensors, allowing for discontinuous objective functions*, SIAM J. Optim., 4 (1994), to appear.

[2] H. Attouch and R. J-B. Wets, *Approximation and convergence in nonlinear optimization*, in Nonlinear Programming 4, O. Mangasarian, R. Meyer, and S. Robinson, eds., Academic Press, New York, 1981, pp. 367–394.

[3] J.-P. Aubin, *Lipschitz behavior of solutions to convex minimization problems*, Math. Oper. Res., 8 (1984), pp. 87–111.

[4] J.-P. Aubin and H. Frankowska, *Set-Valued Analysis*, Birkhaüser, Basel, 1990.

[5] A. Bensoussan and J-L. Lions, *Control Impulsional et Inequations Quasi-Variationelles*, Bordas, Paris, 1982.

[6] F. H. Clarke, *Optimization and Nonsmooth Analysis*, John Wiley, New York, 1983.

[7] V. F. Demyanov and A. M. Rubinov, *Foundations of Nonsmooth Analysis and Quasi-Differential Calculus*, Nauka, Moscow, 1990. (In Russian.)

[8] Y. M. Ermoliev, *Methods of Stochastic Programming*, Nauka, Moscow, 1976. (In Russian.)

[9] Y. M. Ermoliev and A. A. Gaivoronski, *On optimization of discontinuous systems*, Working paper WP-91-41, International Institute for Applied System Analysis, Laxenburg, Austria, 1991.

[10] Y. M. Ermoliev and E. A. Nurminski, *Limit extremal problems*, Kibernetika, 4 (1973), pp. 130–132.

[11] H. Frankowska, *Inclusions adjointes associées aux trajectoires minimales d'inclusions différentielles*, Comptes Rendus de l'Académie des Sciences de Paris, 297 (1983), pp. 461–464.

[12] ———, *The first order necessary conditions for nonsmooth variational and control problems*, SIAM J. Control Optim., 22 (1984), pp. 1–12.

[13] A. A. Gaivoronski, *On nonstationary stochastic optimization problems*, Kibernetika, (1978), pp. 89–92.

[14] W. B. Gong and Y. C. Ho, *Smoothed (conditional) perturbations analysis of discrete event dynamic systems*, IEEE Trans. Automat. Control, 32 (1987), pp. 856–866.

[15] A. M. Gupal, On a method for the minimization of almost differentiable functions, Kibernetika, (1977), pp. 114–116.

[16] ———, *Stochastic Methods for Solving Stochastic Extremal Problems*, Naukova Dumka, Kiev 1979. (In Russian.)

[17] A. M. Gupal and V. I. Norkin, *An algorithm for the minimization of discontinuous functions*, Kibernetika, (1977), pp. 73–75.

[18] A. D. Ioffe, *Nonsmooth analysis: differential calculus of nondifferentiable mappings*, Trans. Amer. Math. Soc., 266 (1981), pp. 1–56.

[19] V. Y. Katkovnik, *Linear Estimators and Stochastic Optimization Problems*, Nauka, Moscow, 1976. (In Russian.)

[20] A. N. Kolmogorov, *Selected Works, Mathematics and Mechanics*, Nauka, Moscow, 1985 (In Russian.)

[21] R. Lepp, *The maximization of a probability function over simple sets*, Izv. Akad. Nauk Estonskoy SSR. Ser. Fiz.-Math., 28 (1979), pp. 303–309. (In Russian.)

[22] ———, *Stochastic approximation type algorithm for the maximization of a probability function*, Izv. Akad. Nauk Estonskoy SSR Ser Fiz.-Mat., 32 (1983), pp. 150–156. (In Russian.)

[23] K. Marti, *Stochastic optimization methods in structural mechanics*, Z. Angew. Math. Mech., 70 (1990), pp. T742–T745.

[24] D. Q. Mayne and E. Polak, *Nondifferential optimization via adaptive smoothing*, J. Optim. Theory Appl., 43 (1984), pp. 19–30.

[25] P. Michel and J-P. Penot, *Calcul sous-differentiel pour des fonctions lipschitziennes et non lipschitziennes*, Comptes Rendus de l'Académie des Sciences de Paris, 298 (1984), pp. 269–272.

[26] B. S. Mordukhovich, *Approximation Methods in Problems of Optimization and Control*, Nauka, Moscow, 1988. (In Russian.)

[27] N. D. Nikolaeva, *On an algorithm for solving convex programming problems*, Ekonom. i Mat. Metody, 10 (1974), pp. 941–946. (In Russian.)

[28] V. Norkin, *Optimization of probabilities*, Preprint 89-9, Glushkow Institut of Cybernetics, Kiev, 1989.

[29] E. Parzen, *On estimation of a probability density function and the mode*, Ann. Math. Statist., 33 (1962), pp. 1065–1076.

[30] B. Polyak, *Nonlinear programming methods in the presence of noise*, Math. Programming, 14 (1978), pp. 87–97.

[31] R. T. Rockafellar, *The Theory of Subgradients and its Application to Problems of Optimization: Convex and Nonconvex Functions*, Helderman Verlag, Berlin, 1981.

[32] ———, *Generalized subgradients in mathematical programming*, in Mathematical Programming: The State of the Art 1982, A. Bachem, M. Grötschel, and B. Korte, eds., Springer-Verlag, Berlin, 1983, pp. 368–380.

[33] R. T. Rockafellar and R. J-B. Wets, *Variational systems, an introduction*, Multifunctions and Integrands, G. Salinetti, ed., Lecture Notes in Mathematics 1091, Springer-Verlag, Berlin, 1984, pp. 1–54.

[34] N. V. Roenko, *Stochastic programming problems with integral functionals from multivalued mappings*, Abstract of Ph.D. thesis, Glushkov Institute of Cybernetics, Kiev, 1983.

[35] M. Rosenblatt, *Remarks on some nonparametric estimates of a density function*, Ann. Math. Statist., 27 (1966), pp. 832–835.

[36] R. Rubinstein, *How to optimize discrete-event systems from a single path by the score function method*, Ann. Oper. Res., 27 (1991), pp. 175–212.

[37] L. Schwartz, *Théorie des Distributions*, Hermann, Paris, 1966.

[38] S. L. Sobolev, *Some Applications of Functional Analysis in Mathematical Physics*, 3rd ed., Nauka, Moscow, 1988. (In Russian.)

[39] V. A. Steklov, *Sur les expressions asymptotiques de certaines fonctions définies par les équations différentielles du second ordre et leurs applications au problème du dévelopement d'une fonction arbitraire en séries procédant suivant les diverses fonctions*, Comm. Charkov Math. Soc., Serie 2, 10 (1907), pp. 97–199. (In Russian.)

[40] ———, *Main Problems of Mathematical Physics*, Nauka, Moscow, 1983. (In Russian.)

[41] E. Tamm, *On a probability function optimization*, Izv. Akad. Nauk Estonskoy SSR. Ser. Fiz.-Math., 28 (1979), pp. 17–24. (In Russian.)

[42] J. Warga, *Necessary conditions without differentiability assumptions in optimal control*, J. Differential Equations, 15 (1975), pp. 41–61.

[43] ———, *Derivative containers , inverse functions and controllability*, Calculus of Variations and Control Theory, D. Russell, 1976, Academic Press, New York, pp. 13–46.

[44] ———, *Fat homeomorphisms and unbounded derivative containers*, J. Math. Anal. Appl., 81 (1981), pp. 545–560.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 33, No. 1, pp. 168–184, January 1995

© 1995 Society for Industrial and Applied Mathematics
008

ON A GENERALIZATION OF A NORMAL MAP AND EQUATION*

JONG-SHI PANG[†] AND JEN-CHIH YAO[‡]

Abstract. The class of normal maps was recently investigated by Robinson and Ralph in connection with the study of a variational inequality defined on a polyhedral set. In this paper a generalization of such a map is considered, and the associated generalized normal equation is studied. The latter provides a unified formulation of several generalized variational inequality and complementarity problems. Using degree theory, some sufficient conditions for the existence of a zero of a generalized normal map are established and the stability of a generalized normal equation at a solution is analyzed. Specializations of the results to various applications are discussed.

Key words. nonsmooth equation, complementarity problem, variational inequality, degree theory, stability analysis

AMS subject classifications. 90C30, 90C33

1. Introduction. In the past two decades, the finite-dimensional variational inequality (VI) and nonlinear complementarity problem (NCP) have been studied extensively; a survey of results and applications can be found in the review article by Harker and Pang [10]. Since the early work of Eaves [4] and others, it has been known that the VI defined on a closed convex set is equivalent to a certain nonsmooth equation. Recently, Robinson [25], [26], [27] introduced the class of *normal maps* to describe such an equation and derived various properties of these maps when the underlying set is a convex polyhedron. Subsequent work can be found in [21], [22].

Motivated by several extensions of the VI and the NCP, we introduce a generalization of a normal map. Using degree theory, we establish some existence results for a generalized normal map to have a zero and discuss their applications. We also apply a recent sensitivity theory for parametric nonsmooth equations [20] to investigate the stability of a generalized normal equation at a given solution.

The rest of this paper is divided into three sections. The next section introduces the generalized normal equation and discusses its applications. In the third section, we derive the existence results using degree theory. Finally, in the fourth and final section, we study the stability of the generalized normal equation at a solution.

2. The generalized normal map and equation. Let f and g be two given mappings from R^n into itself. Let K be a nonempty closed convex set in R^n. The *generalized normal map* (GNM) associated with the triple (K, g, f) is defined to be the mapping $h : R^n \to R^n$ where

$$h(x) = g(x) - \Pi_K(g(x) - f(x)) \quad \text{for } x \in R^n,$$

and Π_K is the projection operator onto K under the Euclidean norm. When K is polyhedral and g is the identity map, then h becomes the *normal map* associated with the pair (K, f). Incidentally, the latter is not quite the same as Robinson's normal

* Received by the editors December 22, 1992; accepted for publication (in revised form) August 20, 1993.

[†] Department of Mathematical Sciences, Johns Hopkins University, Baltimore, Maryland 21218 (msc_wjp@jhuvms.hcf.jhu.edu). The research of this author was partially supported by National Science Foundation grant DDM-9104078 and by the National Science Council of Taiwan during his visit to the National Sun Yat-Sen University in the summer of 1992.

[‡] Department of Applied Mathematics, National Sun Yat-Sen University, Kaohsiung, Taiwan 804, People's Republic of China (yaojc@math.nsysu.edu.tw). The research of this author was partially supported by National Science Council grant NSC 82-0208-M-110-023.

map as defined in [25], which is $f(\Pi_K(x))+x-\Pi_K(x)$; with an abuse of language and no particular harm as far as the VI is concerned, we have extended the usage of the term "normal map" to include the special case of the GNM with g being the identity.

Associated with the GNM h is the *generalized normal equation* (GNE):

$$h(x) = 0. \tag{1}$$

The major objective of this paper is to derive some sufficient conditions for the solvability of this equation and to study its stability at a solution. The derived results will be specialized to various applications that can be modeled by such an equation.

As mentioned before, our consideration of the GNM was motivated by several generalizations of the VI and NCP. We now explain the latter problems. To begin, we note that a vector x is a zero of the GNM h defined above if and only if x satisfies the following conditions: (i) $g(x) \in K$, and (ii) for all $y \in K$,

$$(y - g(x))^T f(x) \geq 0.$$

Hence, by defining the set-valued map T as follows:

$$T(z) = z - g(z) + K \quad \text{for } z \in R^n,$$

we deduce that x is a zero of h if and only if x satisfies: (i) $x \in T(x)$, and (ii) for all $v \in T(x)$,

$$(v - x)^T f(x) \geq 0.$$

The latter is a quasi-variational inequality (QVI) defined by the pair (T, f); see [2]. Conversely, associated with an arbitrary QVI (T, f), where the set-valued map T is given by

$$T(x) = m(x) + K \tag{2}$$

with m being a (point-to-point) map from R^n into itself and K a closed convex set in R^n, is the equivalent generalized normal equation (1) corresponding to

$$g(x) = x - m(x).$$

Hence we see that the generalized normal equation provides a compact representation for a QVI with the map T having the special form (2).

As we can expect from the well-known relation between a VI and a complementarity problem, when K is a cone, the generalized normal equation associated with the triple (K, g, f) is equivalent to the following complementarity problem:

$$g(x) \in K, \quad f(x) \in K^*, \quad g(x)^T f(x) = 0, \tag{3}$$

where

$$K^* = \{y \in R^n : y^T x \geq 0 \text{ for all } x \in K\}$$

is the dual cone of K. The proof of this equivalence is standard and left to the reader. A special case of the latter complementarity problem is the implicit complementarity problem (ICP), which corresponds to the case where $g(x) = x - m(x)$ and K is the nonnegative orthant [18]. In turn, an interesting instance of the ICP is the case where

$$g(x) = \min(g^i(x) : i = 1, \ldots, m) \tag{4}$$

with each g^i being a mapping from R^n into itself and "min" is the componentwise minimium operator; this instance corresponds to what has been called the (finite-dimensional) *order complementarity problem* that has received an increasing amount of attention in the literature in recent years [1], [8], [11], [12]. Note that the mapping g in (4) is in general not F-differentiable even if the individual functions $\{g^i\}_{i=1}^m$ are.

It is interesting to note that with $f(x) = 0$ for all x, (1) reduces to the feasibility problem

$$g(x) \in K. \tag{5}$$

Some of our subsequent results are concerned with this special case.

It should be pointed out that in general, when the mapping g is a bijection, then the corresponding generalized normal equation can be turned into a nonsmooth equation defined by a normal map. Indeed, in this case, it is easy to see that (1) holds if and only if

$$u - \Pi_K(u - f \circ g^{-1}(u)) = 0$$

where $u = g(x)$. This equivalence clearly fails when g is not surjective. Hence, when considering the GNM associated with triple (K, g, f), we avoid assuming that g is surjective.

There is substantial literature on various generalizations of the VI; see [31], [32]. A large portion of this literature is concerned with the derivation of existence results for these problems. These results are established by means of either a fixed point argument or some minimax theory. With the exception of only two articles [14], [8] that deal with the ICP and the order linear complementarity problem, there is no study on the sensitivity of the QVI or the complementarity problem (3). The primary goal of this paper is to derive some new existence and sensitivity results for the latter problems based on the framework of a generalized normal equation.

The tool to be employed in our study is degree theory [15], [17]; hence our approach to the derivation of existence results (in the context of the generalized VIs) is different from the previous approaches. There are two basic motivations for using degree theory in a study of this kind. One is that degree theory is known to be a powerful tool for establishing existence results. (As a matter of fact, this theory was created for this purpose by L. E. J. Brouwer.) The second is that several recent studies [5], [6], [7], [8], [9] suggest that the same theory is useful for dealing with various sensitivity and stability issues of variational inequalities and complementarity problems. Our present research extends these studies and treats some similar issues for the generalized problems.

3. Existence results.

For the discussion in this section we make use of some basic results from degree theory. Those readers who are not familiar with this theory can consult two references [15], [17] for a review of these results and other related background material.

By means of the homotopy invariance property of the degree of a continuous mapping, we can very easily state a general principle for the derivation of existence results to a system of nonlinear equations. With reference to (1) where h is continuous, it suffices to exhibit a continuous homotopy $H : [0, 1] \times R^n \to R^n$ with the property that there exists a bounded open set $\Omega \subseteq R^n$ satisfying the following properties:

(a) the degree of $H(0, \cdot)$ at 0 with respect to Ω, denoted $\deg(H(0, \cdot), \Omega, 0)$, is defined and nonzero;

(b) $0 \notin H(t, \partial\Omega)$ for all $t \in [0,1]$ with ∂ denoting the boundary of a set; and

(c) $H(1,x) = h(x)$ for all $x \in \overline{\Omega}$ with the overline denoting the closure of a set.

If such a homotopy can be identified, then there exists a scalar $\varepsilon > 0$ such that for any function $\tilde{h}$ that is continuous in $\overline{\Omega}$ and satisfies

$$d(h, \tilde{h}) := \sup_{x \in \overline{\Omega}} \|h(x) - \tilde{h}(x)\| < \varepsilon,$$

the system

$$\tilde{h}(x) = 0, \qquad x \in \Omega$$

will have a solution. Note that the conclusion asserts not only the solvability of (1) but also that of all slightly perturbed systems.

Consequently, the task of proving the existence of a solution to (1) has been reduced to the identification of an appropriate homotopy H. In the sequel, we give two basic results of this nature that are specific to a GNM.

THEOREM 3.1. *Let K be a closed convex subset of R^n, and let f and g be two continuous functions from R^n into itself. Suppose there exist a continuous function e from R^n into itself and a bounded open set Ω in R^n such that*

(i) $\deg(k, \Omega, 0)$ *is defined and nonzero where*

$$k(x) = g(x) - \Pi_K(g(x) - e(x)) \quad \textit{for } x \in R^n;$$

$$\text{(ii)} \qquad \left.\begin{array}{r} x \in \partial\Omega \cap g^{-1}(K) \\ \mu > 0 \end{array}\right\} \quad \Rightarrow \quad f(x) + \mu e(x) \notin (K - g(x))^*.$$

Then there exists a solution to the system

$$g(x) - \Pi_K(g(x) - f(x)) = 0, \qquad x \in \overline{\Omega}.$$

Proof. Assume that the GNM

$$h(x) = g(x) - \Pi_K(g(x) - f(x))$$

has no zero in $\overline{\Omega}$. Consider the homotopy defined by

$$H(t,x) = g(x) - \Pi_K(g(x) - tf(x) - (1-t)e(x)).$$

Assumption (i) implies that condition (a) of the homotopy principle is satisfied. Note that our assumption about h implies that $H(1, \cdot)$ cannot have a zero on $\partial\Omega$. Suppose $H(t,x) = 0$ for some $t \in (0,1)$ and $x \in \partial\Omega$. For this vector x, we then have $g(x) \in K$ and

$$(y - g(x))^T (tf(x) + (1-t)e(x)) \geq 0$$

for all $y \in K$. Dividing by $t > 0$ in the last inequality yields a contradiction to assumption (ii) with $\mu = (1-t)/t$. □

The main point of the above result is that the solvability of (1) can be inferred from that of the GNE associated with the triple (K, g, e) under the stated assumptions. Presumably, the auxiliary function e is simpler than f, and the associated GNE is simpler than (1) as well. In the event that such a function e can not be identified easily, the next result might be useful.

THEOREM 3.2. *Let K be a closed convex subset of R^n, and let f and g be two continuous functions from R^n into itself. Suppose there exist a continuous function e from R^n into itself, a vector $a \in K$, and a bounded open set Ω in R^n such that* (i′) $\deg(e-a,\Omega,0)$ *is defined and nonzero*;

(ii′) *with* $\gamma_t(x)=g(x)+(1-t)(e(x)-g(x))$, *we have*

$$\left.\begin{aligned} x\in\partial\Omega\cap\gamma_t^{-1}(K) \\ t\in(0,1) \end{aligned}\right\} \quad\Rightarrow\quad (1-t)(e(x)-a)+tf(x)\notin(K-\gamma_t(x))^*.$$

Then the conclusion of Theorem 3.1 *holds.*

Proof. It suffices to use the homotopy

$$H(t,x)=tg(x)-\Pi_K(t(g(x)-f(x))+(1-t)a)+(1-t)e(x)$$

and apply the same argument as in the proof of the previous theorem. The details are omitted. □

Presumably, the difference between the two theorems is that in the former, assumption (i) is more restrictive, whereas in the latter, assumption (ii) becomes more so. In the sequel, applications of both results will be illustrated. We begin with a corollary that can be derived from either one of the two results.

COROLLARY 3.3. *Let K be a closed convex subset of R^n, and let f and g be two continuous functions from R^n into itself. Suppose there exist a vector $a\in K$ and a bounded open set Ω in R^n such that*

(iii) $\deg(g-a,\Omega,0)$ *is defined and nonzero*;

(iv) $$\left.\begin{aligned} x\in\partial\Omega\cap g^{-1}(K) \\ \mu>0 \end{aligned}\right\} \quad\Rightarrow\quad f(x)+\mu(g(x)-a)\notin(K-g(x))^*.$$

Then the conclusion of Theorem 3.1 *holds.*

Proof. It suffices to take the function e to be g in Theorem 3.2. □

The next result gives a sufficient condition for assumption (iv) to be satisfied.

COROLLARY 3.4. *Let K be a closed convex subset of R^n, and let f and g be two continuous functions from R^n into itself. Suppose there exist a vector $a\in K$ and a bounded open set Ω in R^n such that*

(iii) *deg* $(g-a,\Omega,0)$ *is defined and nonzero*;

(iv′) $(g(x)-a)^Tf(x)\geq 0$ *for all* $x\in\partial\Omega\cap g^{-1}(K)$.

Then the conclusion of Theorem 3.1 holds.

Proof. We verify assumption (iv). Let $x\in\partial\Omega\cap g^{-1}(K)$ and let $\mu>0$. Then $g(x)\neq a$ because of the well-definedness of $\deg(g-a,\Omega,0)$. Hence

$$(a-g(x))^T(f(x)+\mu(g(x)-a))<0$$

by assumption (iv′). Consequently, (iv) follows. □

We should point out that in the case where the set K is the Cartesian product of a finite number of sets of lower dimensions, say,

$$K=\Pi_{q=1}^N K_q, \tag{6}$$

where each K_q is a closed convex set in R^{n_q}, then Corollary 3.4 remains valid if condition (iv′) is replaced by the weaker assumption

(v) $\max_{1\leq q\leq N}(g_q(x)-a_q)^Tf_q(x)\geq 0$ for all $x\in\partial\Omega\cap g^{-1}(K)$.

Before discussing some general conditions on the triple (K,g,f) that will ensure the satisfaction of the assumptions in the above results, we give a simple one-dimensional example to illustrate the last corollary.

Example. Let

$$g(x) = \frac{x}{x^2+1}, \qquad x \in R,$$

and $K = [0, \infty)$. We claim that if f is any continuous function with $f(t) \geq 0$ for some $t > 0$, then the generalized normal equation associated with the triple (K, g, f) has a solution in the interval $[0, t]$. Since K is a (one-dimensional) cone, the discussion in §2 shows that this equation is equivalent to the complementarity problem (3). By an elementary argument, the reader can easily verify that the latter problem must have a solution. We now prove this simple fact using Corollary 3.4. Take $a = 0$ and $\Omega = (-t, t)$. Clearly, the only zero of g is $x = 0$ and $g'(0) = 1$. Hence $\deg(g - a, \Omega, 0) = 1$. Trivially, we have $\partial\Omega \cap g^{-1}(K) = \{t\}$ and $g(t)f(t) \geq 0$. Consequently, the two assumptions of Corollary 3.4 are satisfied and the desired conclusion follows.

3.1. Some known results. The theorems established above are general enough to include as special cases many well-known existence results from diverse applications. The discussion that follows serves to illustrate this point. As we shall see, such familiar results as the famous Farkas lemma in linear programming are amenable to a degree-theoretic treatment. In the next section, we derive some new existence results for various special problems.

We begin by considering the case where $K = R^n$. This, of course, corresponds to the classical problem of finding a zero of the function f (note that g plays no role). For this simple case, we obtain a solvability result for a system of nonlinear equations that is reminiscent of the famous Leray–Schauder fixed-point theorem [15], [17].

PROPOSITION 3.5. *Let f be a continuous function from R^n into itself. Suppose there exist a continuous function e from R^n into itself and a bounded open set Ω in R^n such that* $\deg(e, \Omega, 0)$ *is defined and nonzero, and for all $x \in \partial\Omega$ and all $\mu > 0$,*

$$f(x) + \mu e(x) \neq 0.$$

Then f has a zero in $\overline{\Omega}$.

Proof. It suffices to apply Theorem 3.1 with $K = R^n$ and g arbitrary. □

Our next application concerns the famous Farkas lemma, which states that the system of linear inequalities

$$Az = b, \qquad z \geq 0, \tag{7}$$

where $A \in R^{m \times n}$ and $b \in R^m$ are given, has a solution if and only if the implication below holds:

$$A^T x \geq 0 \quad \Rightarrow \quad b^T x \geq 0. \tag{8}$$

The nontrivial part of this result is the "if" part. Normally, the proof is based on the theory of linear inequalities. In what follows, we show this using Theorem 3.2.

PROPOSITION 3.6. *Suppose* (8) *holds. Then* (7) *is consistent.*

Proof. Let K be the convex cone in R^m generated by the columns of the matrix A. Then K is closed. This is a well-known fact in the theory of linear inequalities; for the sake of completeness, we give an elementary proof, which proceeds as follows. Take an arbitrary vector $y \in \overline{K}$. Then there exist sequences of vectors $\{v^k\} \subset R^m$ converging to zero and $\{z^k\} \subseteq R^n_+$ such that $y + v^k = Az^k$; moreover, for each k, we may choose z^k with the property that the columns of A corresponding to the positive components of z^k are linearly independent. The sequence $\{z^k\}$ must have a convergent subsequence with a nonnegative limit. Thus, $y \in K$.

It suffices to show that the given vector b belongs to K. Let $g(x)$ be the constant function equal to b and let $f(x)$ be the zero function; take e to be the identity function $a = 0$ and take Ω to be any open ball (in R^m) with center at the origin. Assumption (i′) in Theorem 3.2 is clearly satisfied because the degree of the identity map is equal to one. Suppose assumption (ii′) is violated. Then for some $t \in (0,1)$ and $x \in \partial\Omega \cap \gamma_t^{-1}(K)$, we have

$$x \in (K - \gamma_t(x))^*.$$

Since K is a convex cone and $\gamma_t(x) \in K$, it can easily be shown that $x \in K^*$ and $x^T\gamma_t(x) = 0$. The former implies $A^Tx \geq 0$; hence, by assumption, we have $b^Tx \geq 0$. Now, writing out $\gamma_t(x)$, we deduce

$$0 = x^T\gamma_t(x) = x^T(b + (1-t)(x-b)) = tx^Tb + (1-t)x^Tx > 0,$$

where the last inequality holds because $x \in \partial\Omega$ and $t \in (0,1)$. This is a contradiction. Consequently, Theorem 3.2 applies and the desired conclusion follows. □

Our third result concerns the VI that corresponds to the case where the function g is the identity. The following is one of many existence results for this problem; we have chosen it partly because of its generality among such results and partly to pave the way for subsequent generalization to the QVI.

PROPOSITION 3.7. *Let K be a closed convex set in R^n and f a continuous function from R^n into itself. Suppose that there exists a vector $a \in K$ such that the set*

$$\{x \in K : (x-a)^Tf(x) < 0\} \tag{9}$$

is bounded (possibly empty). Then there exists a solution to the VI *defined by the pair (K, f).*

Proof. It suffices to employ Corollary 3.4 with g being the identity map and Ω being an open bounded set containing (9). □

3.2. New existence results. In this section we derive some new existence results for the QVI and ICP. We start with one that concerns the feasibility problem (5) and extends Proposition 3.6. The assumption to be imposed is similar to the one in Proposition 3.7; the proof is similar to that of Proposition 3.6.

PROPOSITION 3.8. *Let K be a closed convex cone and g a continuous function from R^n into itself. Suppose that the set*

$$\{x \in K^* : x^Tg(x) < 0\} \tag{10}$$

is bounded (possibly empty). Then $g^{-1}(K)$ is nonempty.

Proof. Let Ω be an open ball with center at the origin and containing the set (10). Take e to be identity map $a = 0$ and f to be the zero function. We verify that the two assumptions in Theorem 3.2 are satisfied. It suffices to verify the second one. Assume the contrary. Let $t \in (0,1)$ and $x \in \partial\Omega \cap \gamma_t^{-1}(K)$ be such that $x \in (K - \gamma_t(x))^*$. Using the cone property of K we can easily show that $x \in K^*$ and $x^T\gamma_t(x) = 0$. Since $x \in \partial\Omega$ and Ω contains (10), it follows that $x^Tg(x) \geq 0$. As in Proposition 3.6, we can see that this contradicts the equation $x^T\gamma_t(x) = 0$. □

It is possible to establish a version of the above result that does not require K to be a cone. Indeed, if K is a closed convex set containing a vector a such that the set

$$\{x \in R^n : (x-a)^T(g(x)-a) < 0\}$$

is bounded and if g is continuous, then $g^{-1}(K)$ is nonempty. We omit the proof of this statement since it is similar to that of Proposition 3.8.

In the rest of this section we derive some specialized existence results for a generalized normal equation. These results are extensions of their counterparts for the (standard) VI. In principle, many existence results for the latter problem (see [10]) can all be extended to the present context. In the sequel, we choose to focus on the class of generalized normal equations with monotone pairs (g, f). A motivation for this choice is so that we can discuss the uniqueness of a solution to the generalized normal equation. The following definition makes the monotonicity concept precise.

DEFINITION. *The function f is said to be*

(i) monotone *with respect to g on the set K if*

$$g(x), g(y) \in K \quad \Rightarrow \quad (f(x)-f(y))^T(g(x)-g(y)) \geq 0;$$

(ii) strictly monotone *with respect to g on the set K if*

$$[g(x), g(y) \in K, \text{ and } x \neq y] \quad \Rightarrow \quad (f(x)-f(y))^T(g(x)-g(y)) > 0;$$

(iii) strongly monotone *with respect to g on the set K if there exists a constant $c > 0$ such that*

$$g(x), g(y) \in K \quad \Rightarrow \quad (f(x)-f(y))^T(g(x)-g(y)) \geq c\|x-y\|^2.$$

If f and g are both affine functions, say, $f(x) = a + Ax$ and $g(x) = b + Bx$, and if $K = R^n$, then f is monotone with respect to g if and only if the matrix A^TB is positive semidefinite, and f is strongly monotone with respect to g if and only if the same matrix is positive definite. If A is nonsingular, A^TB is positive (semi-)definite if and only if BA^{-1} is; a similar statement holds if B is nonsingular. More generally, if $K = R^n$ and $f(g)$ is a (nonlinear) bijective map, then f is (strictly, strongly) monotone with respect g if and only if $g \circ f^{-1}(f \circ g^{-1})$ is (strictly, strongly) monotone in the usual sense.

The above concept of monotonicity for a pair of affine functions is closely related to a certain generalized $\mathbf{P}_0$-property for a finite family of square matrices of the same order. The latter was introduced by Wilson [29] and its application to nonlinear networks was discussed in various articles in [30]. The recent paper [28] gives a renewed look at these matrix-theoretic properties and studies their role in some generalized linear complementarity problems.

PROPOSITION 3.9. *Let K be a closed convex subset of R^n, and let f and g be two continuous functions from R^n into itself with g being injective. Suppose there exist a vector $u \in g^{-1}(K)$ and positive scalars α and L such that for all $x \in g^{-1}(K)$ with $\|x\| \geq \alpha$, $\|g(x)-g(u)\| \leq L\|x-u\|$. If f is strongly monotone with respect to g on K, then there exists a unique vector $\bar{x}$ satisfying*

$$g(x) = \Pi_K(g(x) - f(x)). \tag{11}$$

Proof. Let $a = g(u) \in K$ and $c > 0$ be the constant associated with the strong monotonicity of f with respect to g on K. Let Ω be a bounded open set such that for all $x \in \partial\Omega$, we have $\|x\| \geq \alpha$ and

$$c\|x-u\| \geq L\|f(u)\|.$$

Since g is injective, it follows that $\deg(g-a, \Omega, 0) = \pm 1$ (see [15]). We verify condition (iv′) in Corollary 3.4. Let $x \in \partial\Omega \cap g^{-1}(K)$. We have

$$\begin{aligned}(g(x)-a)^T f(x) &= (g(x)-g(u))^T(f(x)-f(u)) + (g(x)-g(u))^T f(u)\\ &\geq c\|x-u\|^2 - L\|f(u)\|\|x-u\| \geq 0\end{aligned}$$

by the choice of Ω. Consequently, Corollary 3.4 implies the existence of the vector $\bar{x}$ with the desired property. It remains to establish the uniqueness of $\bar{x}$. Suppose x' is another such solution. Then we have

$$(g(x')-g(\bar{x}))^T f(\bar{x}) \geq 0 \quad \text{and} \quad (g(\bar{x})-g(x'))^T f(x') \geq 0.$$

Adding these two inequalities and using the strong monotonicity assumption immediately yield $\bar{x} = x'$. □

We can see from the above uniqueness proof that in general if f is strictly monotone with respect to g on K, then there can exist at most one solution to the generalized normal equation (11). Another noteworthy remark is that if f is strictly (in particular, strongly) monotone with respect to g on the whole space R^n, then both f and g must be injective (on R^n). Consequently, if we assume that f is strongly monotone with respect to g on R^n (and not just on K), then we may remove the injectivity assumption on g in the above theorem as it would become redundant.

The existence conclusion of Proposition 3.9 remains valid if the strong monotonicity assumption on the entire set K is weakened to the strong monotonicity with respect to the vector u. Indeed, from the above proof, we see that if for all $x \in g^{-1}(K)$ with $\|x\|$ sufficiently large,

$$(f(x)-f(u))^T(g(x)-g(u)) \geq c\|x-u\|^2,$$

then perhaps with a larger set Ω, Corollary 3.4 is still applicable; hence the existence of a solution to (11) follows. Note that with this weakened monotonicity assumption, the uniqueness of $\bar{x}$ is no longer guaranteed.

Taking into account the remark following the proof of Corollary 3.4, which concerns the case of a set K with a Cartesian product structure, we may derive an analogous existence result for the complementarity problem (3) where K is the nonnegative orthant. The proof is similar to the one above and consists of verifying condition (v). We do not repeat this verification.

PROPOSITION 3.10. *Let f and g be two continuous functions from R^n into itself with g being injective. Suppose that there exists a vector u satisfying $g(u) \geq 0$ and there are positive scalars α and L such that for all x with $g(x) \geq 0$ and $\|x\| \geq \alpha$, $\|g(x)-g(u)\| \leq L\|x-u\|$. If there exists a scalar $c > 0$ such that, for all x with $g(x) \geq 0$ and $\|x\|$ sufficiently large,*

$$\max_{1 \leq i \leq n}(f_i(x)-f_i(u))(g_i(x)-g_i(u)) \geq c\|x-u\|^2, \tag{12}$$

then there exists a vector $\bar{x}$ satisfying

$$g(x) \geq 0, \quad f(x) \geq 0, \quad g(x)^T f(x) = 0. \tag{13}$$

We have mentioned that in §2 that for a generalized normal equation associated with the triple (K, g, f), the case where g is bijective corresponds essentially to Robinson's normal equation. In what follows, we give an example of a one-dimensional function g satisfying the assumptions of the above proposition (see also Proposition 3.9) that is not surjective.

Example. Let

$$g(x) = \frac{1 - e^x}{1 + e^x}, \qquad x \in R.$$

Since $|g(x)| \leq 1$, g is not surjective. It is easy to verify that g is injective and satisfies $|g(x) - g(0)| \leq |x|$ for $|x| \geq 1$.

When g is the identity map, (12) reduces to the assumption that f is a *uniform P-function* ([16]). As mentioned in the Introduction, the order complementarity problem (in a finite-dimensional vector lattice) is an instance of the problem (13) where the function g is of the special form (4). When f and each function g^i in the latter equation are all affine, this order complementarity problem has recently been studied extensively by Gowda and Sznajder [8]. Among other things, they give a necessary and sufficient conditions for such a *generalized order linear complementarity problem* to have a unique solution "for all constant vectors" in terms of a consistent determinantal sign property (see also [26], [21]). In the Appendix we show how the latter property is related to (12) specialized to certain piecewise affine functions f and g.

The final result in this section concerns a weakening of the strong monotonicity assumption in Proposition 3.9. In its place, we assume that f is monotone with respect to g on K. We also need two other assumptions, one of which involves the following concept.

DEFINITION. *A function $g : R^n \to R^n$ is said to be* proper *with respect to a set D if for every bounded subset S of D, $g^{-1}(S)$ is bounded.*

In the next result, "int S" and "0^+S" denote, respectively, the interior and recession cone of a set S.

PROPOSITION 3.11. *Let K be a closed convex set in R^n, and let f and g be continuous functions from R^n into itself with g being injective. Suppose there exists a vector $u \in g^{-1}(K)$ with $f(u) \in \operatorname{int}(0^+K)^*$. If f is monotone with respect to g on K and g is proper with respect to K, then* (11) *has a solution.*

Proof. Let $a = g(u)$. We claim that the set

$$\{x \in g^{-1}(K) : (g(x) - a)^T f(x) < 0\}$$

is bounded. Assume the contrary. Then there exists an unbounded sequence $\{x^k\} \subseteq g^{-1}(K)$ such that $(g(x^k) - g(u))^T f(x^k) < 0$ for each k. Since f is monotone with respect to g on K, we have

$$(g(x^k) - g(u))^T f(u) < 0$$

for each k. Since g is proper on K, the sequence $\{g(x^k)\}$ is unbounded. Without loss of generality, we may assume that the normalized sequence $\{(g(x^k) - g(u))/\|g(x^k)\|\}$ converges to a limit v that must belong to $0^+K \setminus \{0\}$ because $\{g(x^k)\} \subseteq K$ and $g(u) \in K$. Moreover, we have $v^T f(u) \leq 0$; but this contradicts the assumption that $f(u) \in \operatorname{int}(0^+K)^*$. This establishes the claim stated at the beginning of the proof. Invoking Corollary 3.4, we easily deduce the desired conclusion of the proposition. □

Remark. When K is a cone, the assumption on the vector u is a "strict feasibility" condition for the complementarity problem (3).

4. Sensitivity analysis. In this section we derive some sensitivity results for the parametric generalized normal equation

$$H(x,p) := G(x,p) - \Pi_K(G(x,p) - F(x,p)) = 0, \tag{14}$$

where $G, F : R^n \times R^m \to R^n$ and K is a closed convex subset of R^n. In this problem, $x \in R^n$ is the primary variable and $p \in R^m$ is the parameter. At a given value p^*, a zero of the function $H(\cdot, p^*)$, say x^*, is known. We are interested in analyzing the behavior of this solution when p is perturbed around p^*. The tool we employ for this analysis is the recent results in [20]. We divide the discussion into two cases. The first case assumes no particular differentiability assumption on the functions F and G and no special structure on K; whereas in the second case, the set K is assumed to be a polyhedron and the functions F and G are assumed to have some differentiability properties.

4.1. The nonsmooth case. In this case we assume that the functions $G(\cdot, p^*)$ and $F(\cdot, p^*)$ have continuous *first-order approximations* at x^* [24]; specifically, we assume that there exist continuous functions g and f such that $g(x^*) = G(x^*, p^*)$, $f(x^*) = F(x^*, p^*)$, and

$$\lim_{x \to x^*} \frac{G(x,p^*) - g(x)}{\|x - x^*\|} = 0, \quad \text{and} \quad \lim_{x \to x^*} \frac{F(x,p^*) - f(x)}{\|x - x^*\|} = 0.$$

Clearly, the function

$$h(x) = g(x) - \Pi_K(g(x) - f(x)) \tag{15}$$

is then a first-order approximation of $H(\cdot, p^*)$ at x^* by the nonexpansiveness of the projection operator. Moreover, h is continuous. By imposing a certain strong monotonicity assumption on the pair of approximating functions (g, f) at the vector x^*, we may derive the following stability result for the solution x^*.

THEOREM 4.1. *In the above setting, suppose that the functions g and f are locally Lipschitzian at x^* and that there exist an open neighborhood V of x^* and positive scalars α and β such that* (i) *g is injective in V and* (ii) *for all $x \in V$ and all p sufficiently close to p^*,*

$$(f(x) - f(x^*))^T(g(x) - g(x^*)) \geq \alpha \|x - x^*\|^2, \tag{16}$$

and

$$\max(\|F(x,p) - F(x,p^*)\|, \|G(x,p) - G(x,p^*)\|) \leq \beta \|p - p^*\|.$$

Then x^* is an isolated zero of $H(\cdot, p^*)$; moreover, there exist a scalar $L > 0$ and open neighborhoods U of p^* and W of x^* such that for all vectors $p \in U$, the system

$$H(x,p) = 0, \qquad x \in W$$

has a solution and for any such solution x,

$$\|x - x^*\| \leq L \|p - p^*\|.$$

Proof. Without loss of generality, we may assume that f and g are Lipschitzian in V. According to the results in [20], it suffices to verify three things: (i) x^* is the only zero of h in V, (ii) the inverse function h^{-1} is pseudo-upper Lipschitzian at 0 relative to x^*, and (iii) the index of h at x^* is nonzero. The "isolated" requirement (i) is easy by (16) and the uniqueness proof in Proposition 3.9. To show (ii), it suffices to verify that there exists a constant $\lambda > 0$ such that if $h(x) = y$ and $x \in V$, then $\|x - x^*\| \leq \lambda\|y\|$. Let x be any such vector. Then we have $g(x) - y \in K$ and

$$(u - g(x) + y)^T(f(x) - y) \geq 0 \quad \text{for all } u \in K.$$

Since $g(x^*) \in K$, we derive

$$(g(x^*) - g(x) + y)^T(f(x) - y)) \geq 0.$$

On the other hand, we also have

$$(g(x) - y - g(x^*))^T f(x^*) \geq 0$$

because $g(x) - y \in K$ and $h(x^*) = 0$. Adding the last two inequalities and rearranging terms, we obtain

$$(g(x) - g(x^*))^T(f(x) - f(x^*)) \leq -y^T(f(x^*) - f(x) + g(x^*) - g(x)).$$

Hence, since $x \in V$, by (16), the local Lipschitzian assumption of f and g at x^*, and the Cauchy–Schwartz inequality, we can easily deduce the existence of the desired constant λ. Finally, we need to verify that the index of h at x^* is nonzero. For this purpose, we consider the homotopy

$$\gamma(t, x) := th(x) + (1 - t)(g(x) - g(x^*)).$$

We claim that for all $t \in [0, 1]$, x^* is the only zero of $\gamma(t, \cdot)$ in V. When $t = 0$, this is true by the injectiveness assumption of g; as noted above, the same is true when $t = 1$. Consider a $t \in (0, 1)$ and suppose $\gamma(t, x) = 0$. Then we have

$$t^{-1}(g(x) - (1 - t)g(x^*)) = \Pi_K(g(x) - f(x)).$$

By an algebraic manipulation similar to the one used above, we can easily deduce that $x = x^*$. This establishes our claim. By the homotopy invariance property of the degree, we deduce that

$$\deg(\gamma(1, \cdot), V', 0) = \deg(\gamma(0, \cdot), V', 0),$$

where V' is any open neighborhood of x^* contained in V; the right-hand degree is equal to ±1 by the injectiveness of g. This completes the proof of the theorem. □

We note that similar to several previous results, if K is a Cartesian product of the form (6), then (16) can be weakened to

$$\max_{1 \leq q \leq N} (f_q(x) - f_q(x^*))^T(g_q(x) - g_q(x^*)) \geq \alpha\|x - x^*\|^2,$$

and the same conclusion of the theorem still holds. Conditions such as this last one and (16) are rather strong; in the next section we see how they can be weakened under a structural assumption on K and a differentiability property of G and F.

4.2. The differentiable and polyhedral case. Consider the parametric equation (14) where K is a polyhedron and the functions $G(\cdot,p^*)$ and $F(\cdot,p^*)$ have Fréchet derivatives at x^*. In this case, the approximating functions g and f can be taken to be

$$g(x) = G(x^*,p^*)+\nabla_x G(x^*,p^*)(x-x^*) \quad \text{and} \quad f(x) = F(x^*,p^*)+\nabla_x F(x^*,p^*)(x-x^*).$$

Moreover, instead of using the function given in (15) as a first-order approximation for $H(\cdot,p^*)$ at x^*, we may use

$$(17)\quad h(x) = \nabla_x G(x^*,p^*)(x-x^*) - \Pi_L((\nabla_x G(x^*,p^*) - \nabla_x F(x^*,p^*))(x-x^*)),$$

where L is the *critical cone* of K at $G(x^*,p^*) - F(x^*,p^*)$. We refer the reader to [19], [24] for an explanation of this cone and a demonstration of why the latter function h is a first-order approximation of $H(\cdot,p^*)$ at x^* in this case (recall that $H(x^*,p^*) = 0$). As a matter of fact, this function h is also the B-derivative of $H(\cdot,p^*)$ at x^* (see the cited references for details). With L being a polyhedral cone, h is a piecewise linear function. Hence h^{-1} is necessarily pseudo-upper Lipschitzian by a result of Robinson [23]. Consequently, the second requirement (ii) in the proof of Theorem 4.1 is easily satisfied. We now address the other two requirements (i) and (iii). The equation $h(x) = 0$ corresponds to the following generalized linear complementarity problem on the cone L:

$$\nabla_x G(x^*,p^*)(x-x^*) \in L, \qquad \nabla_x F(x^*,p^*)(x-x^*) \in L^*,$$
$$(x-x^*)^T \nabla_x G(x^*,p^*)^T \nabla_x F(x^*,p^*)(x-x^*) = 0.$$

Hence x^* is an isolated zero of h if and only if the following system has $v = 0$ as the unique solution:

$$(18)\ \nabla_x G(x^*,p^*)v \in L, \quad \nabla_x F(x^*,p^*)v \in L^*, \quad (\nabla_x G(x^*,p^*)v)^T \nabla_x F(x^*,p^*)v = 0.$$

We now come to the final requirement (iii). We claim that (a) if (18) has $v = 0$ as the unique solution, (b) if the matrix $\nabla_x G(x^*,p^*)$ is nonsingular, and (c) if

$$(19)\qquad \nabla_x G(x^*,p^*)v \in L \quad \Rightarrow \quad v^T \nabla_x G(x^*,p^*)^T \nabla_x F(x^*,p^*)v \geq 0,$$

(under (b), (c) says that the matrix $\nabla_x F(x^*,p^*)(\nabla_x G(x^*,p^*))^{-1}$ is copositive on the cone L), then the index of h given by (17) at x^* is nonzero. Indeed, by the nonsingularity of $\nabla_x G(x^*,p^*)$, the affine mapping

$$a(x) = \nabla_x G(x^*,p^*)(x-x^*)$$

has an index at x^* equal to the sign of the determinant of $\nabla_x G(x^*,p^*)$ that is nonzero; moreover,

$$\gamma(t,x) = th(x) + (1-t)a(x)$$

is a homotopy connecting h and a, and $\gamma(t,\cdot)$ has $x = x^*$ as the unique zero for all $t \in [0,1]$. Consequently, the index of h at x^* equals that of a at x^*.

Summarizing the above discussion, we present a stability result for the parametric equation (14) under the special assumptions on K, F, and G. Since we have already verified the three conditions stated at the beginning of the proof of Theorem 4.1, the proof of the following result is essentially complete.

THEOREM 4.2. *Let K be a polyhedron in R^n and H be defined by* (14). *Suppose $H(x^*,p^*) = 0$, $F(\cdot,p^*)$ and $G(\cdot,p^*)$ are Fréchet differentiable at x^*, and there exists a constant $\beta > 0$ such that for all (x,p) sufficiently close to (x^*,p^*),*

$$\max(\|F(x,p) - F(x,p^*)\|, \|G(x,p) - G(x,p^*)\|) \leq \beta\|p - p^*\|.$$

Assume further that the Jacobian matrix $\nabla_x G(x^*,p^*)$ is nonsingular, (18) *has $v = 0$ as the unique solution, and* (19) *holds. Then the same conclusion of Theorem* 4.1 *is valid for* (14).

5. Appendix. In this appendix we relate the strong monotonicity condition (12) to a consistent determinantal sign property of matrices. The latter property has been used by several authors [13], [8] to characterize the global homeomorphism of a piecewise affine mapping and the global solvability of a generalized order linear complementarity problem. We are not concerned with these characterizations here; instead, we provide some matrix-theoretic criteria for the satisfaction of (12) when the functions f and g are the componentwise mininum of certain affine mappings.

DEFINITION. *Let $(A_i : i = 1, \ldots, m)$ be a family of m square matrices of order n. A* representative matrix *of this family is an $n \times n$ matrix B whose ith row is that of one of the A_j.*

The following lemma is concerned with the case of two matrices.

LEMMA 5.1. *Let A and B be two square matrices of order n. The following statements are equivalent.*

(a) *All representative matrices of the pair (A, B) have the same nonzero determinantal sign.*

(b) *All matrices of the form D_1A+D_2B, where D_1 and D_2 are nonnegative diagonal matrices with $(D_1)_{ii} + (D_2)_{ii} > 0$ for all i, have the same nonzero determinantal sign.*

(c) *Both A and B are nonsingular and BA^{-1} is a* P*-matrix.*

(d) *There exists a scalar $c > 0$ such that*

$$\max_{1 \leq i \leq n} (Az)_i(Bz)_i \geq c\|z\|^2$$

for all vectors $z \in R^n$.

Proof. (a) $\Rightarrow$ (b). Clearly, (a) implies that det A is nonzero. Moreover, by a suitable renaming of the matrices A and B if necessary, it suffices to show that the determinant of a matrix of the form $A + DB$ where D is an arbitrary nonnegative diagonal matrix has the same sign as det A. In turn, the latter can be proved by induction on the number of nonzero diagonal entries of D. The key inductive step follows rather easily from (a) and the fact that the determinant of a matrix is a linear function of any given row.

(b) $\Rightarrow$ (c). Clearly, (b) implies that both A and B are nonsingular. Suppose that BA^{-1} is not a P-matrix. Then by a sign reversal property of a P-matrix [3, Thm. 3.3.4], it follows that there exists a nonzero vector u such that

$$u_i(BA^{-1}u)_i \leq 0 \quad \text{for all } i.$$

By letting $A^{-1}u = z$ we obtain

$$(Az)_i(Bz)_i \leq 0 \quad \text{for all } i.$$

Hence defining

$$(D_1)_{ii} = \begin{cases} 0 & \text{if } (Bz)_i = 0, \\ 1 & \text{if } (Bz)_i \neq 0, \end{cases} \quad \text{and} \quad (D_2)_{ii} = \begin{cases} 1 & \text{if } (Bz)_i = 0, \\ -\frac{(Az)_i}{(Bz)_i} & \text{if } (Bz)_i \neq 0, \end{cases}$$

and letting D_1 and D_2 be the respective (nonnegative) diagonal matrices, we see that $D_1 + D_2$ is a positive diagonal matrix and

$$(D_1A + D_2B)z = 0,$$

which contradicts the nonsingularity of the matrix $D_1A + D_2B$.

(c) $\Rightarrow$ (d). The above proof already reveals this implication (using the characterization of a P-matrix in terms of the sign reversal property).

(d) $\Rightarrow$ (a). Let us call two representative matrices C_1 and C_2 of the pair (A, B) *complementary* if for each i, $\{i\text{th row of } C_1, i\text{th row of } C_2\} = \{i\text{th row of } A, i\text{th row of } B\}$. Clearly, if C_1 and C_2 are any two complementary representative matrices, then

$$\max_{1 \leq i \leq n} (C_1z)_i(C_2z)_i \geq c\|z\|^2$$

for all z. This implies that all complementary representative matrices are nonsingular and $C_2C_1^{-1}$ is a P-matrix. Hence, any two complementary representative matrices have the same (nonzero) determinantal sign. It remains to show that any representative matrix has the same nonzero determinantal sign as A. For this purpose, we first show that if C is the (representative) matrix obtained by replacing one row of A with the corresponding row of B, then det CA^{-1} is positive. Without loss of generality, we may take this to be the first row. Then it suffices to show that $(B_1A^{-1})_1 > 0$, where B_1 is the first row of B. Let z be the first column of A^{-1}. Then Az is the first coordinate vector. Hence, by (d), we have

$$0 < c\|z\|^2 \leq \max_{1 \leq i \leq n} (Az)_i(Bz)_i = (B_1A^{-1})_1,$$

as desired. To complete the proof, we use an inductive argument. Suppose that C is a representative matrix with $k(\geq 2)$ rows coming from B (which we take to be the first k rows) and the remaining $n - k$ rows from A (which therefore are the last $n - k$). Let C' be the representative matrix whose first $k - 1$ rows come from B and the remaining $n - k + 1$ rows from A. By an inductive hypothesis, the determinantal signs of C' and A are the same. Let $\tilde{C}$ be the complementary representative matrix of C'. Clearly, C is obtained from C' by replacing its kth row with that of $\tilde{C}$; hence the proof of the case $k = 1$ implies that C and C' have the same determinantal sign. The inductive proof is now complete. □

We now generalize the above lemma to a family of matrices.

PROPOSITION 5.2. *Let $(A_1, \ldots, A_m)$ be a family of square matrices of order n. Consider the following statements.*

(a) *All representative matrices of this family have the same nonzero determinantal sign.*

(b) *All matrices of the form $\sum_{j=1}^{m} D_jA_j$, where D_j are nonnegative diagonal matrices satisfying $\sum_{j=1}^{m}(D_j)_{ii} > 0$ for all i, have the same nonzero determinantal sign.*

(c) *All representative matrices of the family $(A_j : j = 1, \ldots, m)$ are nonsingular and $B_1B_2^{-1}$ is a* P*-matrix for any two such representative matrices B_1 and B_2.*

(d) *For arbitrary vectors* $(a^1, \ldots, a^m)$ *and* u, *there exist scalars* $c, \alpha > 0$ *such that for any two subsets* J_1 *and* J_2 *of* $\{1, \ldots, m\}$,

$$\max_{1 \le i \le n} (f(z) - f(u))_i (g(z) - g(u))_i \ge c\|z - u\|^2$$

for all vectors $z \in R^n$ *with* $\|z\| \ge \alpha$, *where*

$$f(z) = \min(A_j z + a^j : j \in J_1), \qquad g(z) = \min(A_j z + a^j : j \in J_2).$$

It holds that (a) ⇔ (b) ⇔ (c) ⇒ (d).

Proof. (a) ⇒ (b) ⇒ (c). The proofs of these implications are easy extensions of the corresponding proofs for the case of $m = 2$.

(c) ⇒ (a). Fix $B_1 = A_1$. It then follows that all representative matrices have the same determinantal sign as A_1.

(c) ⇒ (d). According to Lemma 5.1 and by the fact that there are only finitely many representative matrices, it follows that there exists a constant $c_1 > 0$ such that for any two representative matrices B_1 and B_2, we have

$$\max_{1 \le i \le n} (B_1 z)_i (B_2 z)_i \ge c_1 \|z\|^2$$

for all vectors z. For the functions f and g as given in (d), it is clear that we have

$$(f(z) - f(u))_i (g(z) - g(u))_i = (A_{j_1} z)_i (A_{j_2} z)_i + \text{a linear term in } z + \text{ a constant},$$

where the two integers $j_1 \in J_1$ and $j_2 \in J_2$ depend on z, u, and i. Let B_1 and B_2 be the representative matrices with the ith row equal to that of A_{j_1} and A_{j_2}, respectively. Then we have

$$\begin{aligned} &\max\nolimits_{1 \le i \le n} (f(z) - f(u))_i (g(z) - g(u))_i \\ &\ge \max\nolimits_{1 \le i \le n} (B_1 z)_i (B_2 z)_i + \text{ a piecewise linear function of } z \\ &\ge c_1 \|z\|^2 + \text{ a piecewise linear function of } z. \end{aligned}$$

From this, the existence of the constants c and α follows easily. □

Note that in the last proposition, the reverse implication (d) ⇒ (a) is not claimed. At this time, we are not certain whether this will hold, hence the omission. Finally, we mention that many results related to the ones given in this Appendix can be found in [29], [30], [28].

Acknowledgments. Several people have made useful comments concerning this paper. M.S. Gowda pointed out the connection between the monotonicity concepts defined in §3 and the generalized $\mathbf{P}_0$-properties of matrices introduced by Wilson; he also provided references [29], [30]. Danny Ralph brought reference [27] to our attention. S. M. Robinson reminded us that the closedness of the set K was needed in Proposition 3.6. The referees and the editors also have made a number of constructive comments. We thank them all.

REFERENCES

[1] J. M. Borwein and M. A. H. Dempster, *The linear order complementarity problem*, Math. Oper. Res., 14 (1989), pp. 534–558.

[2] D. Chan and J. S. Pang, *The generalized quasi-variational inequality*, Math. Oper. Res., 7 (1982), pp. 284–313.

[3] R. W. Cottle, J. S. Pang, and R. E. Stone, *The Linear Complementarity Problem*, Academic Press, New York, 1992.

[4] B. C. Eaves, *On the basic theorem of complementarity*, Math. Programming, 1 (1971), pp. 68–75.

[5] M. S. Gowda, *Applications of degree theory to linear complementarity problems*, Math. Oper. Res., 18 (1993), pp. 868–879.

[6] M. S. Gowda and J. S. Pang, *Stability analysis of variational inequalities and nonlinear complementarity problems, via the mixed linear complementarity problem and degree theory*, Math. Oper. Res., 20 (1995), to appear.

[7] ———, *On the boundedness and stability of solutions to the affine variational inequality problem*, SIAM J. Control Optim., 32 (1994), pp. 421–441.

[8] M. S. Gowda and R. Sznajder, *The generalized order linear complementarity problem*, SIAM J. Matrix Anal. Appl., 15 (1994), pp. 779–795.

[9] C. D. Ha, *Application of degree theory in stability of the complementarity problem*, Math. Oper. Res., 12 (1987), pp. 368–376.

[10] P. T. Harker and J. S. Pang, *Finite-dimensional variational inequality and nonlinear complementarity problems: A survey of theory, algorithms, and applications*, Math. Programming, B, 48 (1990), pp. 161–220.

[11] G. Isac and D. Goeleven, *The implicit general order complementarity problem, models and iterative methods*, Ann. Oper. Res., 44 (1993), pp. 63–92.

[12] G. Isac and M. M. Kostreva, *The generalized order complementarity problem*, J. Optim. Theory Appl., 71 (1991), pp. 517–534.

[13] D. Kuhn and R. Löwen, *Piecewise affine bijections of R^n, and the equation $Sx^+ - Tx^- = y$*, Linear Algebra Appl., 96 (1987), pp. 109–129.

[14] J. Kyparisis and C. M. Ip, *Solution behavior for parametric implicit complementarity problems*, Math. Programming, 56 (1992), pp. 65–70.

[15] N. G. Lloyd, *Degree Theory*, Cambridge University Press, Cambridge, UK, 1978.

[16] J. J. Moré, *Coercivity conditions in nonlinear complementarity problems*, SIAM Rev., 16 (1974), pp. 1–16.

[17] J. M. Ortega and W. C. Rheinboldt, *Iterative Solution of Nonlinear Equations in Several Variables*, Academic Press, New York, 1970.

[18] J. S. Pang, *The implicit complementarity problem*, in Nonlinear Programming 4, O. L. Mangasarian, R. R. Meyer, and S. M. Robinson, eds., Academic Press, New York, 1981, pp. 487–518.

[19] ———, *Newton's method for B-differentiable equations*, Math. Oper. Res., 15 (1990), pp. 311–341.

[20] ———, *A degree-theoretic approach to parametric nonsmooth equations with multivalued perturbed solution sets*, Math. Programming ser. B, 62 (1993), pp. 359–383.

[21] D. Ralph, *A new proof of Robinson's homeomorphism theorem for piecewise linear maps*, Linear Algebra Appl., 178 (1993), pp. 249–260.

[22] ———, *On branching numbers of normal manifolds*, Math. Oper. Res., to appear.

[23] S. M. Robinson, *Some continuity properties of polyhedral multifunctions*, Math. Programming Stud., 14 (1981), pp. 206–214.

[24] ———, *An implicit-function theorem for a class of nonsmooth functions*, Math. Oper. Res., 16 (1991), pp. 292–309.

[25] ———, *Normal maps induced by linear transformations*, Math. Oper. Res., 17 (1992), pp. 691–714.

[26] ———, *Homeomorphism conditions for normal maps of polyhedra*, in Optimization and Nonlinear Analysis, A. Ioffe, M. Marcus, and S. Reich, eds., Pitman Research Notes in Mathematics No. 244, Longman, Harlow, U.K., 1992, pp. 240–248.

[27] ———, *Nonsingularity and symmetry for linear normal maps*, Math. Programming Ser. B, 62 (1993), pp. 415–426.

[28] R. Sznajder and M. S. Gowda, *Generalization of* $\mathbf{P}_0$- *and* $\mathbf{P}$*-properties*; *Extended vertical and horizontal LCPs*, Linear Algebra Appl., (1994), to appear.

[29] A. N. Wilson, Jr., *A useful generalization of the* $\mathbf{P}_0$*-matrix concept*, Numer. Math., 17 (1971), pp. 62–70.

[30] ———, ed., *Nonlinear Networks: Theory and Analysis*, IEEE Press, New York, 1974.

[31] J. C. Yao, *Generalized quasi-variational inequality and implicit complementarity problems*, Ph.D. thesis, Department of Operations Research, Stanford University, Stanford, CA, June 1990.

[32] ———, *The generalized quasi-variational inequality problem with applications*, J. Math. Anal. Appl., 158 (1991), pp. 139–160.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 33, No. 1, pp. 185–207, January 1995

© 1995 Society for Industrial and Applied Mathematics

009

NONLINEAR H^∞ OPTIMIZATION: A CAUSAL POWER SERIES APPROACH*

CIPRIAN FOIAS†, CAIXING GU†, AND ALLEN TANNENBAUM‡

Abstract. In this paper, using a power series methodology a design procedure applicable to analytic nonlinear plants is described. The technique used is a generalization of the linear H^∞ theory. In contrast to previous work on this topic ([*Indiana J. Math.*, 36 (1987), pp. 693–709], [*Oper. Theory Adv. Appl.*, 41 (1989), pp. 255–277], [*SIAM J. Control Optim.*, 27 (1989), pp. 842–860]), the authors are now able to incorporate explicitly a causality constraint into the theory. In fact, it is shown that it is possible to reduce a causal optimal design problem (for nonlinear systems) to a classical interpolation problem solvable by the commutant lifting theorem [*Harmonic Analysis of Operators on Hilbert Space*, North-Holland, Amsterdam, 1970], [*The Commutant Lifitng Approach to Interpolation Problems*, Birkhäuser, Boston, 1990].

Key words. nonlinear systems, H^∞ optimization, causality, commutant lifting theorem, interpolation theory, Volterra series

AMS subject classifications. primary 47A20, secondary 47A99, 93B35, 93C05

1. Introduction. In this paper, we continue our work on finding a suitable, implementable nonlinear extension of the powerful linear H^∞ design methodology. In what follows, we will just consider discrete-time systems, even though the techniques elucidated below carry over to the continuous-time setting as well.

Our approach is based on previous work ([14], [15]) in which we considered systems described by analytic input/output operators. A key idea here involved the expression of each n-linear term of a suitable Taylor expansion of the given operator as an equivalent linear operator acting on a certain associated tensor space, which allowed us to iteratively apply the classical commutant lifting theorem in designing a compensator. (Our class of operators includes Volterra series [9].)

More precisely, in such an approach we are reduced to applying the classical (linear) commutant lifting theorem to an H^2-space defined on some D^n (where D denotes the unit disc). Now when we apply the classical result to D^n ($n \geq 2$), even though time-invariance is preserved (that is, commutation with the appropriate shift), causality may be lost. Indeed, for systems described by analytic functions on the disc D (these correspond to stable, discrete-time, one-dimensional (1D) systems), time-invariance (that is, commutation with the unilateral shift) implies causality. For analytic functions on the n-disc ($n > 1$), this is not necessarily the case. For dynamical system control design and for any physical application, this is, of course, a major drawback for such an approach. (The compensators we obtained were "weakly causal" and causal approximations were discussed.)

Hence for a dilation result in $H^2(D^n)$ we must include the causality constraint explicitly in the set-up of the dilation problem. It is precisely this problem that motivated the mathematical operator-theoretic work of [16] and [13], which incorporated Arveson theory [1] into the dilation, commutant lifting framework.

* Received by the editors August 24, 1992; accepted for publicaiton (in revised form) August 24, 1993. This work was supported in part by a grant from the Research Fund of Indiana University, and by National Science Foundation grants DMS-8811084 and ECS-9122106, Air Force Office of Scientific Research grants AFOSR-90-0024 and F49620-94-1-00S8DEF, Army Research Office grants DAAL03-91-G-0019, DAAH04-94-G-0054, and DAAH04-93-G-0032, and by Image Evolutions, Ltd.

† Department of Mathematics, Indiana University, Bloomington, Indiana 47405.

‡ Department of Electrical Engineering, University of Minnesota, Minneapolis, Minnesota 55455.

While the general method explicated in this paper is based on a causal extension of the commutant lifting theorem, for the purposes of the operators and spaces that appear in control we will give a direct simple method for finding the optimal causal compensators. In fact, we will show that *the computation of an optimal causal nonlinear compensator may be reduced to a classical interpolation problem.*

We now briefly outline the contents of this paper. In §2, we define *causality* and *time-invariance* as applied to analytic mappings. We show in particular that while in the linear case, time-invariance and boundedness imply causality, this is not true in general in the nonlinear setting. In §3, we formulate the causal optimization problem to be studied. In §4, we discuss the Fourier representation of certain Hilbert spaces, a technique that we apply throughout the paper. In §5, we prove the main theoretical result of this paper in which we show how to reduce a causal optimization problem to a problem solvable via the classical commutant lifting theorem [25]. This is summarized in a computational algorithm in §§6 and 7. Sections 8 and 9 are then concerned with our formulation of the nonlinear generalization of the H^∞ sensitivity minimization problem, which is then solved via a *causal iterative commutant lifting* method in §10. Section 11 is devoted to a natural control interpretation of our optimization procedure, while §12 is connected to computational aspects of our work, namely a nonlinear notion of *rationality* that reduces our work to finite-dimensional skew Toeplitz calculations. We illustrate our methods with an example in §13, and finally in §14, we make some concluding remarks.

We conclude this section by noting that there have been other approaches to nonlinear H^∞. These include a nonlinear commutant lifting theorem [3], [4], and a very promising nonlinear game-theoretic approach [7] as well as a nonlinear version of Ball–Helton theory [6], and the recent work in [26].

Once again, we will just consider discrete-time systems in what follows.

2. Causal analytic mappings. In this section, we will define the class of nonlinear input/output operators that we will study in this paper. To do this, we will first need to discuss a few standard results about analytic mappings on Hilbert spaces. See [3], [4], [14], [15], [21] and the references therein for complete details.

Let $\mathcal{G}$ and $\mathcal{H}$ denote complex separable Hilbert spaces. Set

$$B_{r_o}(\mathcal{G}) := \{g \in \mathcal{G} : \|g\| < r_o\}$$

(the open ball of radius r_o in $\mathcal{G}$ about the origin). Then we say that a mapping $\phi : B_{r_o}(\mathcal{G}) \to \mathcal{H}$ is *analytic* if the complex function $(z_1, \ldots, z_n) \mapsto \langle \phi(z_1 g_1 + \ldots + z_n g_n), h\rangle$ is analytic in a neighborhood of $(1, 1, \ldots, 1) \in \mathbf{C}^n$ as a function of the complex variables $z_1, \ldots, z_n$ for all $g_1, \ldots, g_n \in \mathcal{G}$ such that $\|g_1 + \ldots + g_n\| < r_o$, for all $h \in \mathcal{H}$, and for all $n > 0$.

We will now assume that $\phi(0) = 0$. It is easy to see that if $\phi : B_{r_o}(\mathcal{G}) \to \mathcal{H}$ is analytic, then ϕ admits a convergent Taylor series expansion ([21, p. 97]), i.e.,

$$\phi(g) = \phi_1(g) + \phi_2(g, g) + \cdots + \phi_n(g, \cdots, g) + \cdots,$$

where $\phi_n : \mathcal{G} \times \cdots \times \mathcal{G} \to \mathcal{H}$ is an n-linear map. Clearly, without loss of generality we may assume that the n-linear map $(g_1, \cdots, g_n) \to \phi(g_1, \ldots, g_n)$ is symmetric in the arguments $g_1, \ldots, g_n$. This assumption will be made throughout this paper for the various analytic maps which we consider. For ϕ a Volterra series, ϕ_n is basically the nth-Volterra kernel.

Now set

$$\hat{\phi}_n(g_1 \otimes \cdots \otimes g_n) := \phi_n(g_1, \ldots, g_n).$$

Then $\hat{\phi}_n$ extends in a unique manner to a dense set of $\mathcal{G}^{\otimes n} := \mathcal{G} \otimes \cdots \otimes \mathcal{G}$ (tensor product taken n times). Note by $\mathcal{G}^{\otimes n}$ we mean the Hilbert space completion of the algebraic tensor product of the $\mathcal{G}$'s. Clearly if $\hat{\phi}_n$ has finite norm on this dense set, then $\hat{\phi}_n$ extends by continuity to a bounded linear operator $\hat{\phi}_n : \mathcal{G}^{\otimes n} \to \mathcal{H}$. By abuse of notation, we will set $\phi_n := \hat{\phi}_n$. (Recall that an n-linear map on $G \times G \times \cdots \times G$ (product taken n times) becomes linear on the tensor product $\mathcal{G}^{\otimes n}$. For details about the construction of the tensor product, see [2, pp. 24–27].)

We now recall the following standard definitions.

DEFINITION 1. (i) *Notation as above. By a* majorizing sequence *for the analytic map ϕ, we mean a positive sequence of numbers α_n $n = 1, 2, \ldots$ such that $\|\phi_n\| < \alpha_n$ for $n \geq 1$. Suppose that $\rho := \limsup {\alpha_n}^{1/n} < \infty$. Then it is completely standard that the Taylor series expansion of ϕ converges at least on the ball $B_r(\mathcal{G})$ of radius $r = 1/\rho$* ([21, p. 97]).

(ii) *If ϕ admits a majorizing sequence as in* (i), *then we will say that ϕ is* majorizable.

Let $H^2_K(D^n)$ denote the standard Hardy space of $\mathbf{C}^K$-valued analytic functions on the n-disc D^n (D denotes the unit disc) with square integrable boundary values. We set $H^2_K := H^2_K(D)$ and and $H^2 := H^2_1$. We denote the shift on $H^2_K(D^n)$ by $S_{(n)}$. Note that $S_{(n)}$ is defined by multiplication by the function $(z_1 \cdots z_n)$. On H^2_K we set $S_{(1)} =: U$ (U is given by multiplication by z).

We now consider an analytic map ϕ with $\mathcal{G} = \mathcal{H} = H^2_k$. Note that

$$H^2_k \otimes \cdots \otimes H^2_k = (H^2_k)^{\otimes n} \cong H^2_K(D^n) \quad \text{with } K = k^n, \tag{1}$$

where we map $1 \otimes \cdots \otimes z \otimes \cdots \otimes 1$ (z in the ith place) to z_i, $i = 1, \ldots, n$. Clearly, $S_{(n)}$ corresponds to $U^{\otimes n}$ under this identification.

We will identify ϕ_n as a bounded linear map from $H^2_K(D^n) \to H^2_k$ via the canonical isomorphism (1). Then we say that ϕ is *time-invariant* if

$$\phi_n S_{(n)} = U \phi_n \quad \forall n \geq 1. \tag{2}$$

(We will also say each ϕ_n is *time-invariant.*) Equivalently, this means that $U\phi = \phi \circ U$ on some open ball about the origin in which ϕ is defined.

Now set

$$P^{(j)}_{(n)} := I - S^j_{(n)} S^{*j}_{(n)}, \qquad j \geq 1, \quad n \geq 1.$$

Note

$$P^{(j)} := P^{(j)}_{(1)} = I - U^j U^{*j}.$$

Then we say that ϕ is *causal* if

$$P^{(j)} \phi_n = P^{(j)} \phi_n P^{(j)}_{(n)}, \qquad j \geq 1, \quad n \geq 1. \tag{3}$$

(We also say each ϕ_n is *causal.*) Equivalently, $\phi_n : H^2_K(D^n) \to H^2_k$ is causal if for $F(z_1, \ldots, z_n) \in H^2_K(D^n)$,

$$F(z_1, \ldots, z_n) = \sum_{i_1, \ldots, i_n \geq 0} F_{i_1, \ldots, i_n} z_1^{i_1} \ldots z_n^{i_n}, \qquad \phi_n(F)(z) := \sum_{m \geq 0} f_m z^m,$$

each f_m only depends on

$$\{F_{i_1,\ldots,i_n} : 0 \leq i_1, \ldots, i_n \leq m\}.$$

This means that for

$$F(z_1, \ldots, z_n) = \sum_{\max\{i_1,\ldots,i_n\} \geq m} F_{i_1,\ldots,i_n} z_1^{i_1} \ldots z_n^{i_n},$$

we have that

$$(I - U^m U^{*m})\phi_n(F(z_1, \ldots, z_n)) = 0. \tag{4}$$

We would now like to discuss the relationship between time-invariance and causality. For simplicity, we assume $k = 1$, i.e., we work with single-input/single-output (SISO) systems. Let $\phi : H^2 \to H^2$ be linear and time-invariant (i.e., intertwines with the shift). Then it is easy to see that ϕ is causal. Indeed, $\phi U = U\phi$ implies

$$\begin{aligned} U^m U^{*m} \phi U^m U^{*m} &= U^m U^{*m} U^m \phi U^{*m} \\ &= U^m \phi U^{*m} \\ &= \phi U^m U^{*m}, \end{aligned}$$

which immediately implies

$$P^{(m)} \phi P^{(m)} = P^{(m)} \phi \quad \forall m \geq 1,$$

that is, ϕ is causal.

In the nonlinear setting however, time-invariance may not imply causality. As a concrete example, let $\phi_o : (H^2)^{\otimes 2} \to H^2$ be a linear operator such that $U^{\otimes 2}\phi_o = \phi_o U$, defined by

$$(\phi_o(f \otimes g))(z) := \sum_{m=0}^{\infty} (f_{m+1} f_m + f_m g_m + f_m g_{m+1}) z^m,$$

where

$$f(z) = \sum_{m=0}^{\infty} f_m z^m, \qquad g(z) = \sum_{m=0}^{\infty} g_m z^m.$$

Now set

$$\phi(f) := \phi_o(f \otimes f), \qquad f \in H^2.$$

Then ϕ is an analytic, time-invariant map. (In fact ϕ is a homogeneous polynomial of degree 2.) But ϕ is not causal. Indeed,

$$\begin{aligned} (P^{(1)}\phi(f))(z) &= 2 f_1 f_0 + f_0^2, \qquad z \in D \\ (P^{(1)}\phi(P_{(2)}^{(1)} f))(z) &= f_0^2, \qquad z \in D. \end{aligned}$$

Thus $P^{(1)}\phi(f) \neq P^{(1)}\phi(P_{(2)}^{(1)} f)$, for example for $f(z) := 1 + z$ for $z \in D$. (Note that under the identification (1), $P_{(2)}^{(1)}$ corresponds to $P^{(1)} \otimes P^{(1)}$.)

3. Causal optimization problem. One of the key techniques in this paper will be to reduce a nonlinear generalization of the H^∞ sensitivity minimization problem to a series of *linear causal optimization problems.* (This will be done in §§8–10 below.) In this section, we will formulate this new causal problem.

As above, we let $S_{(n)}$ denote the unilateral shift on $H^2_K(D^n)$ given by multiplication by $(z_1 \cdots z_n)$. Since $H^2_K(D^n)$ will be fixed in the discussion we will let $S := S_{(n)}$. As above, U will denote the unilateral shift on H^2_k given by multiplication by z, and $\Theta \in H^\infty_{k\times k}$ will be an inner $k \times k$ matrix-valued H^∞ function (i.e., a $k \times k$ inner matrix with entries H^∞ scalar functions). Finally $W : H^2_K(D^n) \to H^2_k$ will denote a causal, time-invariant bounded linear operator (in the sense of (2) and (3) above).

We can now state the *causal H^∞-optimization problem* (COP): Find

$$\sigma := \inf\{\|W - \Theta Q\| : Q : H^2_K(D^n) \to H^2_k,\ Q \text{ causal, time-invariant}\}. \tag{5}$$

Moreover, we want to compute an optimal, causal, time-invariant Q_{opt} such that

$$\sigma = \|W - \Theta Q_{\text{opt}}\|. \tag{6}$$

If we drop the causality constraint the solution to problem (5) is provided by the classical commutant lifting theorem [25]. With the causality constraint, the solution to (COP) is abstractly provided by a causal commutant lifting theorem [16], [13].

In this paper, based on this work we will provide a simple solution to the problem (COP) without directly referring to the operator theoretic results of [16] and [13]. In fact, we will show how to directly reduce the computation of σ to a classical interpolation problem handled by the ordinary commutant lifting theorem, a computational procedure for which was given in [14] and [15]. We will also describe how to get the corresponding optimal parameter Q_{opt}.

Our technique will be based on a *reduction theorem* stated in §5. To formulate this result, we will first discuss the Fourier representation, which we do in the next section.

4. Fourier representation. In what follows we must use the Fourier representation of elements of $H^2_K(D^n)$. We refer the reader to [25] for all the details.

We first precisely define all the relevant spaces. First we denote by

$$\ell^2(H^2_K) := \bigoplus_{i=1}^{\infty} H^2_K,$$

the Hilbert space of all column vectors

$$f(z) = [f_1(z), f_2(z), \ldots, f_n(z), \ldots]', \tag{7}$$

($'$ stands for tranpose) such that

$$\|f\|^2 := \sum_{i=1}^{\infty} \|f_i\|^2, \tag{8}$$

is finite. ($\|\ \|$ is our generic symbol for a Hilbert space norm (2-norm) as well as the induced operator norm. So for example in (8), it stands for the usual norm on H^2_K as well as the associated norm on $\ell^2(H^2_K)$.) Thus $\ell^2(H^2_K)$ is a vector-valued Hardy space. Indeed, if $f(z)$ is given by (7), then we may write

$$f(z) = \sum_{m=0}^{\infty} a_m z^m, \tag{9}$$

where each a_m is an infinite column vector with components in $\mathbf{C}^K$, and

$$a_m = \frac{1}{m!}[f_1^{(m)}(0), \ldots, f_j^{(m)}(0), \ldots]'.$$

Clearly,

$$\|f\|^2 = \sum_{m=0}^{\infty} \|a_m\|^2.$$

Conversely, if $f(z) \in \ell^2(H_K^2)$ is given in the form (9) for

$$a_m = [a_{m1}, \ldots, a_{mj}, \ldots]',$$

then $f(z)$ can be written in the form (7), i.e.,

$$f(z) = [f_1(z), \ldots, f_j(z), \ldots]',$$

where

$$f_j(z) = \sum_{m=0}^{\infty} a_{mj} z^m.$$

In what follows, we will either use representation (7) or (9). The context should always make the meaning clear.

Next we let $S_\Phi : \ell^2(H_K^2) \to \ell^2(H_K^2)$ denote the unilateral shift defined by multiplication by z. Then the *Fourier representation of* $H_K^2(D^n)$ is given by the (linear, bounded) operator

$$\Phi := \Phi_n : H_K^2(D^n) \to \ell^2(H_K^2),$$

which is defined by

$$\begin{aligned} f(z) &:= \Phi(F(z_1, z_2, \ldots, z_n)) \\ &:= \sum_{m=0}^{\infty} z^m \begin{bmatrix} F_{m,m,\ldots,m} \\ F_{m,\ldots,m,m+1} \\ F_{m,\ldots m+1,m+1} \\ \vdots \\ F_{m+i_1,m+i_2,\ldots,m+i_n} \\ \vdots \end{bmatrix}, \end{aligned} \tag{10}$$

where

$$F(z_1, \ldots, z_n) = \sum_{j_1,\ldots,j_n \geq 0} F_{j_1,\ldots,j_n} z_1^{j_1} \cdots z_n^{j_n},$$

and $(i_1, \ldots, i_n) \in I_n$ for

$$I_n := \{(i_1, \ldots, i_n) : i_1, \ldots, i_n \geq 0, \ \min\{i_1, \ldots, i_n\} = 0\}. \tag{11}$$

We order the set I_n in the following manner. We have $(i_1, \ldots, i_n) < (i_1', \ldots, i_n')$, if $\max\{i_1, \ldots, i_n\} < \max\{i_1', \ldots, i_n'\}$. Thus

$$I_n = \bigcup_{k \geq 0} I_n^{(k)},$$

where

$$I_n^{(k)} := \{(i_1, \dots, i_n) \in I_n : \max\{i_1, \dots, i_n\} = k\}.$$

Each $I_n^{(k)}$ is then ordered by the lexicographical order.

Note that we are taking $f(z)$ in the form (9) in the above representation. Moreover, note that

$$H_K^2(D^n) = \left\{ F(z_1, \dots, z_n) = \sum_{j_1,\dots,j_n \geq 0} F_{j_1,\dots,j_n} z_1^{j_1} \cdots z_n^{j_n} : \sum_{j_1,\dots,j_n \geq 0} \|F_{j_1,\dots,j_n}\|^2 < \infty \right\}.$$

We can also write

$$f(z) = [f_{0,\dots,0}(z), f_{0,\dots,1}(z), \dots, f_{i_1,\dots,i_n}(z), \dots]', \tag{12}$$

where

$$f_{i_1,\dots,i_n}(z) := \sum_{m=0}^{\infty} F_{i_1+m,\dots,i_n+m} z^m, \tag{13}$$

and $(i_1, \dots, i_n) \in I_n$.

Next, it is easy to see that $\Phi : H_K^2(D^n) \to \ell^2(H_K^2)$ is an isometry. Indeed, using (10), (12), and (13), we have

$$\begin{aligned} \|\Phi(F)\|^2 &= \|f\|^2 \\ &= \sum_{i_1,\dots,i_n \in I_n} \|f_{i_1,\dots,i_n}\|^2 \\ &= \sum_{i_1,\dots,i_n \in I_n} \|F_{i_1+m,\dots,i_n+m}\|^2 \\ &= \sum_{j_1,\dots,j_n \geq 0} \|F_{j_1,\dots,j_n}\|^2 \\ &= \|F\|^2. \end{aligned}$$

A similar computation shows that the adjoint of Φ is also an isometry, so that Φ is an unitary operator. We next show that

$$\Phi S = S_\Phi \Phi. \tag{14}$$

Indeed, we see that

$$\begin{aligned} \Phi S(F) &= \Phi(z_1 \cdots z_n F(z_1, \dots, z_n)) \\ &= \Phi(\sum_{j_1,\dots,j_n \geq 0} F_{j_1,\dots,j_n} z_1^{j_1+1} \cdots z_n^{j_n+1}) \\ &= \sum_{m=0}^{\infty} z^{m+1} \begin{bmatrix} F_{m,\dots,m} \\ F_{m,\dots,m,m+1} \\ F_{m,\dots,m+1,m+1} \\ \vdots \\ F_{m+i_1,m+i_2,\dots,m+i_n} \\ \vdots \end{bmatrix} \\ &= z\Phi(F) \\ &= S_\Phi \Phi(F). \end{aligned}$$

By (14), we see that if $W : H^2_K(D^n) \to H^2_k$ is such that $WS = UW$, then the operator $W\Phi^* : \ell^2(H^2_K) \to H^2_k$ satisfies

$$(W\Phi^*)S_\Phi = WS\Phi^* = U(W\Phi^*);$$

that is, $W\Phi^*$ intertwines the shifts S_Φ and U. Consequently, it is standard (see, e.g., [12], or [25, p. 277]) that $W\Phi^*$ is represented by a row vector

$$[W_{0,\dots,0}(z), W_{0,\dots,1}(z), \dots, W_{i_1,\dots,i_n}(z), \dots], \tag{15}$$

for $(i_1, \dots, i_n) \in I_n$. Specifically, for any

$$f(z) = [f_{0,\dots,0}(z), f_{0,\dots,1}(z), \dots, f_{i_1,\dots,i_n}(z), \dots]' \in \ell^2(H^2_K),$$

we have

$$(W\Phi^*)f(z) = \sum_{i_1,\dots,i_n \in I_n} W_{i_1,\dots,i_n}(z) f_{i_1,\dots,i_n}(z). \tag{16}$$

We will write that

$$W\Phi^* \cong [W_{0,\dots,0}(z), W_{0,\dots,1}(z), \dots, W_{i_1,\dots,i_n}(z), \dots], \tag{17}$$

in the sense expressed by (15) and (16).

We would like to make this representation a bit more precise now. Note that the action of $W\Phi^*$ is determined by its action on

$$\ker S^*_\Phi = \{a \in \ell^2(H^2_K) : a \text{ is a column vector with components in } \mathbf{C}^K\}.$$

(This follows from the fact that

$$\ell^2(H^2_K) \cong \bigoplus_{j=0}^{\infty} S^j_\Phi(\ker S^*_\Phi),$$

and that $W\Phi^*$ intertwines the shifts S_Φ and U.) Thus we need only to compute the action of W on

$$\Phi^* \ker S^*_\Phi = \left\{ F(z_1, \dots, z_n) \in H^2_K(D^n) : F(z_1, \dots, z_n) = \sum_{i_1,\dots,i_n \in I_n} F_{i_1,\dots,i_n} z_1^{i_1} \cdots z_n^{i_n} \right\}.$$

(See (11) for the definition of I_n.) By linearity,

$$W\left(\sum_{i_1,\dots,i_n \in I_n} F_{i_1,\dots,i_n} z_1^{i_1} \cdots z_n^{i_n} \right) = \sum_{i_1,\dots,i_n \in I_n} F_{i_1,\dots,i_n} W(z_1^{i_1} \cdots z_n^{i_n}).$$

So by (10) and (16) we have

$$W\Phi^* \cong [W_{0,\dots,0}(z), W_{0,\dots,1}(z), \dots, W_{i_1,\dots,i_n}(z), \dots], \tag{18}$$

where

$$W_{i_1,\dots,i_n}(z) = W(z_1^{i_1} \cdots z_n^{i_n}) \qquad (i_1, \dots, i_n) \in I_n. \tag{19}$$

The above discussion used only the time-invariance for W. In the next proposition, we will write down an explicit expression for the row vector of (18) and (19) associated with $W\Phi^*$ in case W is causal.

PROPOSITION 4.1. *Let $W : H^2_K(D^n) \to H^2_k$ be time-invariant. Then W is causal if and only if*

$$W_{i_1,\ldots,i_n}(z) = z^{\max\{i_1,\ldots,i_n\}} W^c_{i_1,\ldots,i_n}(z) \quad \forall (i_1, \ldots, i_n) \in I_n,$$

where $W^c_{i_1,\ldots,i_n}(z) \in H^\infty_{k\times K}$ (the space of $k \times K$ matrix-valued H^∞ functions).

Proof. By definition, for all $(i_1, \ldots, i_n) \in I_n$ with $\max\{i_1, \ldots, i_n\} = m$, and for all $v \in \mathbf{C}^k$, we have by the causality condition (4) that

$$(I - U^m U^{*m}) W\Phi^*(\Phi(z_1^{i_1} \cdots z_n^{i_n} v)) = (I - U^m U^{*m}) W_{i_1,\ldots,i_n}(z) v = 0.$$

Thus

$$W_{i_1,\ldots i_n}(z) = z^m W^c_{i_1,\ldots i_n}(z) = z^{\max\{i_1,\ldots,i_n\}} W^c_{i_1,\ldots i_n}(z) \quad \forall (i_1, \ldots, i_n) \in I_n, \tag{20}$$

for some $W^c_{i_1,\ldots i_n}(z) \in H^\infty_{k\times K}$, as required. □

By the above discussion (in particular, Proposition 4.1), we see that for W, Θ as in the (COP) problem (5), we have

$$\begin{aligned}
\sigma &= \inf\{\|W - \Theta Q\| : QS = UQ, Q \text{ causal, time-invariant}\} \\
&= \inf\{\|W\Phi^* - \Theta Q\Phi^*\| : (Q\Phi^*)S_\Phi = U(Q\Phi^*), Q \text{ causal, time-invariant}\} \\
&= \inf\{\|W_1 - \Theta Q_1\| : W_1, Q_1 : \ell^2(H^2_K) \to H^2_k,\ W_1 = W\Phi^*, \\
Q_1 &\cong [q_{0,\ldots,0}(z), zq_{0,\ldots,1}(z), \ldots, zq_{1,\ldots,1,0}(z), z^2 q_{0,\ldots,2}(z), \ldots]\}.
\end{aligned}$$

From now on (unless explicitly stated otherwise), we will just work with the operators $W_1, Q_1 : \ell^2(H^2_K) \to H^2_k$. Essentially, via the unitary equivalence Φ, we are identifying the spaces $H^2_K(D^n)$ and $\ell^2(H^2_K)$. In particular, we identify W with $W_1 = W\Phi^*$, and Q with $Q_1 = Q\Phi^*$. For simplicity of notation, we will denote

$$W = W_1, \qquad Q = Q_1.$$

The context should always make the meaning clear.

We now translate the notions of causality and time-invariance for an operator $W : \ell^2(H^2_K) \to H^2_k$. We will say that W is *time-invariant* if $WS_\Phi = UW$, that is,

$$W \cong [W_{0,\ldots,0}(z), W_{0,\ldots,1}(z), \ldots, W_{i_1,\ldots,i_n}(z), \ldots].$$

Moreover, we say that W is *causal* if the operator $W\Phi : H^2_K(D^n) \to H^2_k$ is causal, which means (see Proposition 4.1) that

$$W \cong [W^c_{0,\ldots,0}(z), zW^c_{0,\ldots,1}(z), \ldots, zW^c_{1,\ldots,1,0}(z), z^2 W^c_{0,\ldots,2}(z), \ldots],$$

for some

$$\{W^c_{i_1,\ldots,i_n}(z) \in H^\infty_{k\times K} : (i_1, \ldots, i_n) \in I_n\}.$$

Motivated by the above discussion, for $W : \ell^2(H^2_K) \to H^2_k$ time-invariant and causal, we introduce the operator

$$\begin{aligned}
W_c &\cong [W^c_{0,\ldots,0}(z), W^c_{0,\ldots,1}(z), \ldots, W^c_{1,\ldots,1,0}(z), W^c_{0,\ldots,2}(z), \ldots] \\
&= [W_{0,\ldots,0}(z), W_{0,\ldots,1}(z)/z, \ldots, W_{1,\ldots,1,0}(z)/z, W_{0,\ldots,2}(z)/z^2, \ldots].
\end{aligned} \tag{21}$$

We conclude this section by noting that to solve the (COP) problem (5), we can equivalently solve the following problem: Given $W : \ell^2(H^2_K) \to H^2_k$ time-invariant and causal as above, find

$$\sigma = \inf\{\|W - \Theta Q\| : QS_\Phi = UQ,\ Q \text{ causal}\}. \tag{22}$$

Thus we must solve the optimization problem (COP) on the Fourier transformed operators. This we will show how to explicitly do via a reduction theorem in the next section.

5. Reduction theorem. In this section, we formulate and prove our main result which will allow us to reduce the computation of a causal dilation to an ordinary one based on the classical commutant lifting theorem, i.e., interpolation in H^∞. In what follows $\mathcal{H}$, $\mathcal{K}$, $\mathcal{H}_i, i \geq 1$ will denote (complex, separable) Hilbert spaces.

To prove the result we will need two elementary lemmas.

LEMMA 5.1. *Let $A : \mathcal{K} \to \mathcal{H}$ be a bounded linear operator, and let T and S^* be isometries on $\mathcal{H}$ and $\mathcal{K}$, respectively. Then*

$$\|TAS\| = \|A\|.$$

Proof. By hypothesis, $T^*T = I$, and $SS^* = I$, and so

$$\begin{aligned}\|A\|^2 &= \|A^*A\| = \|A^*T^*TA\| \\ &= \|(TA)(TA)^*\| = \|TASS^*(TA)^*\| \\ &= \|(TAS)(TAS)^*\| = \|TAS\|^2,\end{aligned}$$

as required. □

LEMMA 5.2. *Let*

$$A = [A_1, A_2, \ldots] : \bigoplus_{i=1}^{\infty} \mathcal{H}_i \to \mathcal{H},$$

where

$$A(\oplus_{i=1}^\infty h_i) := \sum_{i=1}^{\infty} A_i h_i.$$

Further, let U_i^* be an isometry on $\mathcal{H}_i$ for $i \geq 1$. Then

$$\|A\| = \|[A_1, A_2, \ldots]\| = \|[A_1U_1, A_2U_2, \ldots]\|.$$

Proof. Note that

$$[A_1U_1, A_2U_2, \ldots, A_nU_n, \ldots] = [A_1, A_2, \ldots, A_n, \ldots] \begin{bmatrix} U_1 & 0 & 0 & \ldots & \ldots \\ 0 & U_2 & 0 & \ldots & \ldots \\ \vdots & \ddots & \ddots & \vdots & \vdots \\ 0 & \ldots & \ddots & U_n & \vdots \\ \vdots & \ldots & \ldots & \ddots & \ddots \end{bmatrix}.$$

However, if we set $S := \oplus_{i=1}^\infty U_i$, by hypothesis S^* is an isometry on $\oplus_{i=1}^\infty \mathcal{H}_i$, and so by Lemma 5.1, we are done. □

THEOREM 5.3 (Reduction theorem). *Notation as above. Then*

$$
\begin{aligned}
&(23)\quad \sigma = \inf\{\|W - \Theta Q\|\} : QS = UQ,\ Q\ \textit{causal}\}\\
&(24)\quad\ \ = \inf\{\|[W_{0,\dots,0}(z) - \Theta q_{0,\dots,0}(z), z(W_{0,\dots,1}(z) - \Theta q_{0,\dots,1}(z)), \dots]\| :\\
&\qquad\qquad [q_{0,\dots,0}(z), \dots, q_{i_1,\dots,i_n}(z), \dots] \in \mathcal{L}(\ell^2(H^2_K), H^2_k),\ (i_1, \dots, i_n) \in I_n\}\\
&(25)\quad\ \ = \inf\{\|W_c - \Theta Q\| : QS = UQ\}.
\end{aligned}
$$

(*Note in* (24) *the norm is the operator norm in* $\mathcal{L}(\ell^2(H^2_K), H^2_k)$. *In general, for Hilbert spaces* $\mathcal{H}$ *and* $\mathcal{K}$, $\mathcal{L}(\mathcal{H}, \mathcal{K})$ *denotes the space of bounded linear operators from* $\mathcal{H}$ *to* $\mathcal{K}$.)

Proof. The second equality (24) follows from Proposition 4.1. To prove the third equality (25), it is enough to prove that for any causal, time-invariant operator

$$\Omega \cong [\omega_{0,\dots,0}(z), \omega_{0,\dots,1}(z), \dots, \omega_{i_1,\dots,i_n}(z), \dots],$$

we have $\|\Omega\| = \|\Omega_c\|$. (See (21) above.)

Now since

$$
\begin{aligned}
\|\Omega\| &= \text{ess sup}\{\|[\omega_{0,\dots,0}(\zeta), \omega_{0,\dots,1}(\zeta), \dots, \omega_{i_1,\dots,i_n}(\zeta), \dots]\| : |\zeta| = 1\},\\
\|\Omega_c\| &= \text{ess sup}\{\|[\omega_{0,\dots,0}(\zeta), \omega^c_{0,\dots,1}(\zeta), \dots, \omega^c_{i_1,\dots,i_n}(\zeta), \dots]\| : |\zeta| = 1\},
\end{aligned}
$$

we need to prove that for any fixed $\zeta \in \partial D$ that

$$\|[\omega_{0,\dots,0}(\zeta), \omega_{0,\dots,1}(\zeta), \dots, \omega_{i_1,\dots,i_n}(\zeta), \dots]\| = \|[\omega_{0,\dots,0}(\zeta), \omega^c_{0,\dots,1}(\zeta), \dots, \omega^c_{i_1,\dots,i_n}(\zeta), \dots]\|.$$

However, by Proposition 4.1,

$$\omega_{i_1,\dots,i_n}(\zeta) = \omega^c_{i_1,\dots,i_n}(\zeta)\zeta^{\max\{i_1,\dots,i_n\}} I_{\mathbf{C}^K},$$

where $I_{\mathbf{C}^K}$ is the identity on $\mathbf{C}^K$. Hence by Lemma 5.2 with $\mathcal{H}_i := \mathbf{C}^K$ and $U_i := \zeta^{\max\{i_1,\dots,i_n\}} I_{\mathbf{C}^K}$ $(i \geq 1)$, we are done. □

6. Algorithm for computation of σ. We would like to summarize the above discussion with a high-level algorithm for the computation of the optimal causal performance σ, and corresponding causal optimal interpolant Q_{opt} in (5) and (6).

First, using the notation of Theorem 5.3, let us denote

$$(26)\qquad \sigma_o := \inf\{\|W_c - \Theta Q\| : QS = UQ\}.$$

(See equation (25).) Then Theorem 5.3 guarantees that

$$\sigma = \sigma_o.$$

This means that a causal optimization problem can be reduced to a classical generalized interpolation problem in H^∞.

We can summarize the procedure as follows:

(i) Let W, Θ be as in (5). (Thus $W : H^2_K(D^n) \to H^2_k$ here.) We compute $W(z_1^{i_1} \cdots z_n^{i_n})$ where $(i_1, \dots, i_n) \in I_n$. By (18) and (19), we get

$$W\Phi^* \cong [W_{0,\dots,0}(z), W_{0,\dots,1}(z), \dots, W_{i_1,\dots,i_n}(z), \dots],$$

and then by (20) we obtain the row matrix

$$[W_{0,\dots,0}(z), W^c_{0,\dots,1}(z), \dots, W^c_{i_1,\dots,i_n}(z), \dots].$$

(ii) The row matrix represents an operator (see (17)) $W_c : \ell^2(H^2_K) \to H^2_k$. Let $\Pi : H^2_k \to H^2_k \ominus \Theta H^2_k$ denote orthogonal projection. Using skew Toeplitz theory ([8], [17], [20]), we can compute the norm of the operator

$$\Lambda(W, \Theta) := \Pi W_c. \tag{27}$$

This norm is σ, the optimal causal performance.

(iii) Using the classical commutant lifting theorem and skew Toeplitz theory, we can compute the optimal dilation $B_c : \ell^2(H^2_K) \to H^2_k$ of $\Lambda(W, \Theta)$. Recall this means that

$$B_c S_\Phi = U B_c, \ \ \Pi B_c = \Lambda(W, \Theta), \ \ \|B_c\| = \|\Lambda(W, \Theta)\| = \sigma.$$

We can then write

$$B_c = W_c - \Theta Q_{opt,c}.$$

Then from (21), we can find the *optimal causal dilation*

$$B = W\Phi^* - \Theta Q_{opt}\Phi^*.$$

Note that B and B_c are related as in (21), and similarly for $Q_{\text{opt},c}$ and $Q_{\text{opt}}\Phi^*$. $Q_{\text{opt}} : H^2_K(D^n) \to H^2_k$ is the *optimal causal interpolant*, i.e.,

$$\sigma = \|W - \Theta Q_{\text{opt}}\|.$$

In the next section, we will give an explicit procedure for the computation of Q_{opt} in the SISO case.

7. Maximal vectors and optimal dilations. We use the notation of the previous section. We want to show how to compute the optimal dilation for

$$A := \Lambda(W, \Theta) : \ell^2(H^2) \to H^2.$$

(We are only considering SISO systems here.)

Our discussion will be based on [15], which generalizes a well-known result of Sarason [24]. We recall that a *maximal vector* of A, $h_o \neq 0$, is a vector such that $\|Ah_o\| = \|A\|\|h_o\|$.

Given $h \in \ell^2(H^2)$,

$$h = [h_1, h_2, \ldots]',$$

we write

$$h^* = [\bar{h}_1, \bar{h}_2, \ldots].$$

Moreover, we set

$$T := \Pi U | H^2 \ominus \Theta H^2,$$

where $\Pi : H^2 \to H^2 \ominus \Theta H^2$ denotes orthogonal projection. As above, U is the unilateral shift on H^2, and S_Φ denotes the shift on $\ell^2(H^2)$.

With this notation, we can now state the following result.

PROPOSITION 7.1. *Notation as above. Let $A : \ell^2(H^2) \to H^2 \ominus \Theta H^2$ be as above (so that $AU = TA$). Suppose moreover that that A has a maximal vector h_o. Let $B_c : \ell^2(H^2) \to H^2$ be the minimal intertwining dilation of A, i.e., $\Pi B_c = A$, $B_c U = S_\Phi B_c$, and $\|A\| = \|B_c\|$. Then if we let $\lambda := \|A\|^2$, we have that*

$$B_c = \frac{\lambda h_o^*}{A h_o}.$$

Proof. We sketch the proof following [15]. First, given $h_o \in H$, we represent h_o as a column vector with components h_j, $j \geq 1$ as above. Let

$$B_c \cong [b_1, b_2, \ldots].$$

Then we have that

$$(B_c h_o)(z) = \sum_{j \geq 1} b_j(z) h_j(z)$$

(for $z \in D$), and

$$\|B_c\| = \text{ess sup} \left\{ \left(\sum_{j=1}^{\infty} |b_j(\zeta)|^2 \right)^{1/2} : |\zeta| = 1 \right\}.$$

However,

$$\|A\|^2 \|h_o\|^2 = \|A h_o\|^2 \leq \|B_c h_o\|^2 \leq \|B_c\|^2 \|h_o\|^2 = \|A\|^2 \|h_o\|^2.$$

Thus $\|A h_o\|^2 = \|B_c h_o\|^2$, and since $\Pi B_c h_o = A h_o$, we have that $A h_o = B_c h_o$. Next note that

$$\sum_{j \geq 1} |b_j(e^{it})|^2 \leq \lambda$$

almost everywhere, and

$$\frac{1}{2\pi} \int_0^{2\pi} \left(\lambda \sum_{j=1}^{\infty} |h_j(e^{it})|^2 - |\sum_{j=1}^{\infty} b_j(e^{it}) h_j(e^{it})|^2 \right) dt = 0.$$

(This follows from the fact that $\lambda \|h_o\|^2 = \|B_c h_o\|^2$.) Using the Cauchy–Schwarz inequality, the expression under the integral sign is nonnegative. Thus

$$\begin{aligned} \lambda \sum_{j \geq 1} |h_j(e^{it})|^2 &= |\sum_{j \geq 1} b_j(e^{it}) h_j(e^{it})|^2 \\ &\leq \left(\sum_{j \geq 1} |b_j(e^{it})|^2 \right) \left(\sum_{j \geq 1} |h_j(e^{it})|^2 \right) \leq \lambda \sum_{j \geq 1} |h_j(e^{it})|^2 \end{aligned}$$

almost everywhere, which implies that

$$\sum_{j \geq 1} |b_j(e^{it})|^2 = \lambda$$

almost everywhere, and

$$h_j = \phi(e^{it})\overline{b_j(e^{it})}$$

almost everywhere for all $j \geq 1$, and for some function $\phi \in H^2$ satisfying

$$Ah_o = B_c h_o = \lambda\phi.$$

Thus for

$$B_c(e^{it}) \cong [b_1(e^{it}), b_2(e^{it}), \ldots]$$

we have

$$B_c(e^{it})\overline{Ah_o(e^{it})} = \lambda h_o(e^{it})^*$$

almost everywhere, as required. □

Remarks. (i) As remarked above, from the optimal dilation B_c, we can solve for $Q_{\mathrm{opt},c}$ via

$$B_c = W_c - \Theta Q_{\mathrm{opt},c}.$$

The optimal causal interpolant is then derived as described as in the last section.

(ii) In some cases it may be more convenient to derive the optimal dilation from a maximal vector of A^*. A similar proof to the one just given shows that

$$B_c = \frac{\overline{A^* h_1}}{\overline{h_1}}, \tag{28}$$

where $h_1 \in H^2 \ominus \Theta H^2$ is a maximal vector for A^*.

8. Nonlinear Control Problem. We will now describe the physical control problem in which we are interested. In our treatment that follows, we will add the causality constraint to the results of [15], and thereby derive a physically realizable nonlinear optimization procedure. First, we will need to consider the precise kind of input/output operator we will be considering. As above, H_k^2 denotes the standard Hardy space of $\mathbf{C}^k$-valued functions on the unit disc. We now make the following definition.

Then we say an analytic input/output operator $\phi : H_k^2 \to H_k^2$ is *admissible* if it is causal, time-invariant, majorizable, and $\phi(0) = 0$. We denote

$$\mathcal{C}_l := \{\text{space of admissible operators}\}.$$

Since the theory we are considering is local, the notion of admissibility is sufficient for all of the applications we have in mind.

We now begin to formulate our control problem. Referring to Fig. 1, P represents a physical plant that we assume is modeled by an admissible operator. In our problem, we are required to design a feedback compensator C in such a way as to attentuate the effect of the filtered disturbances (filtered by the "weight" W) d. The unfiltered disturbances v are assumed to have energy (i.e., 2-norm) bounded by some fixed constant. This leads to following kind of mathematical problem. See [14] and [15] for more details.

Let P, W denote admissible operators, with W invertible. Then we say that the feedback compensator C *stabilizes* the closed loop if the operators $(I + P \circ C)^{-1}$ and

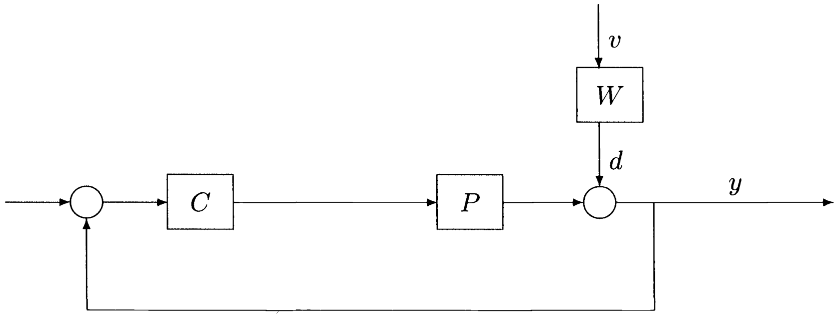


FIG. 1. *Standard feedback configuration.*

$C \circ (I + P \circ C)^{-1}$ are well defined and admissible. We can show that C stabilizes the closed loop if and only if

$$C = \hat{q} \circ (I - P \circ \hat{q})^{-1} \tag{29}$$

for some $\hat{q} \in \mathcal{C}_l$. (See [14], [15] and the references therein.) Note that the *weighted sensitivity* $(I + P \circ C)^{-1} \circ W$ can be written as $W - P \circ q$, where $q := \hat{q} \circ W$. This is precisely the operator relating the disturbance v to the output y. (Since W is invertible, the data q and $\hat{q}$ are equivalent.) In this context, we will call such a q, a *compensating parameter*. Note that from the compensating parameter q, we get a stabilizing compensator C via the formula (29).

As in [15], the problem we would like to solve here is a nonlinear version of the classical disturbance attenuation problem. This corresponds to the "minimization" of the "sensitivity" $W - P \circ q$ taken over all admissible q. To formulate a precise mathematical problem, we need to say in what sense we want to minimize $W - P \circ q$. This we will do in the next section, where we will propose a notion of "sensitivity minimization" which seems quite natural to analytic input/output operators. For the linear case of sensitivity minimization see [10], [18] and the references therein.

9. Nonlinear sensitivity function. This section follows very closely the set-up of [15]. However, now we explicitly put in the causality constraint.

We begin by defining a fundamental object, namely a nonlinear version of *sensitivity*. We should note that while the optimal H^∞ measure of performance is a real number in the linear case [18], the measure of performance that seems to be more natural in this nonlinear setting is a certain function defined in a real interval. This new kind of performance criterion is one of the keys concepts developed in [14] and [15]. See also §11 for a further analysis of the physical meaning of our nonlinear sensitivity function.

To define our notion of sensitivity, we will first have to partially order germs of analytic mappings. All of the input/output operators here will be admissible. We also follow here our convention that for given $\phi \in \mathcal{C}_l$, ϕ_n will denote the bounded linear map on the space $(H^2_k)^{\otimes n} \cong H^2_K(D^n)$ (with $K = k^n$) associated to the n-linear part of ϕ, which we also denote by ϕ_n (and which we always assume without loss of generality is symmetric in its arguments). The context will always make the meaning of ϕ_n clear.

We can now state the following definition.

DEFINITION 2. (i) *For $W, P, q \in \mathcal{C}_l$ (W is the weight, P the plant, and q the compensating parameter), we define the* sensitivity function $S(q)$,

$$S(q)(\rho) := \sum_{n=1} \rho^n \|(W - P \circ q)_n\|$$

for all $\rho > 0$ such that the sum converges. Note that for fixed P and W, for each $q \in \mathcal{C}_l$, we get an associated sensitivity function.

(ii) *We write $S(q) \preceq S(\tilde{q})$, if there exists a $\rho_o > 0$ such that $S(q)(\rho) \leq S(\tilde{q})(\rho)$ for all $\rho \in [0, \rho_o]$. If $S(q) \preceq S(\tilde{q})$ and $S(\tilde{q}) \preceq S(q)$, we write $S(q) \cong S(\tilde{q})$. This means that $S(q)(\rho) = S(\tilde{q})(\rho)$ for all $\rho > 0$ sufficiently small, i.e., $S(q)$ and $S(\tilde{q})$ are equal as germs of functions.*

(iii) *If $S(q) \preceq S(\tilde{q})$, but $S(\tilde{q}) \not\preceq S(q)$, we will say that q* ameliorates $\tilde{q}$. *Note that this means $S(q)(\rho) < S(\tilde{q})(\rho)$ for all $\rho > 0$ sufficiently small.*

Now with Definition 2, we can define a notion of "optimality" relative to the sensitivity function.

DEFINITION 3. (i) *$q_o \in \mathcal{C}_l$ is called* optimal *if $S(q_o) \preceq S(q)$ for all $q \in \mathcal{C}_l$.*

(ii) *We say $q \in \mathcal{C}_l$ is* optimal with respect to its nth term q_n, *if for every n-linear $\hat{q}_n \in \mathcal{C}_l$, we have*

$$S(q_1 + \cdots + q_{n-1} + q_n + q_{n+1} \ldots) \preceq S(q_1 + \cdots + q_{n-1} + \hat{q}_n + q_{n+1} + \cdots).$$

If $q \in \mathcal{C}_l$ is optimal with respect to all of its terms, then we say that it is partially optimal.

10. Iterative causal commutant lifting method. In this section, we discuss a construction from which we will derive both partially optimal and optimal compensators relative to the sensitivity function given in Definition 2 above. As before, P will denote the plant, and W the weighting operator, both of which we assume are admissible. We always suppose that P_1 (the linear part of P) is an isometry, i.e., P_1 is a $k \times k$ inner matrix-valued H^∞ function. (P_1 corresponds to Θ of §6.)

We begin by noting the following key relationship:

$$(W - P \circ q)_l = W_l - \sum_{1 \leq j \leq l} \sum_{i_1 + \cdots + i_j = l} P_j(q_{i_1} \otimes \cdots \otimes q_{i_j}) \quad \forall l \geq 1.$$

Note that once again for ϕ admissible, ϕ_n denotes the n-linear part of ϕ, as well as the associated linear operator on $H^2_K(D^n)$.

We are now ready to formulate the *iterative causal commutant lifting procedure.* Let $\Pi : H^2_k \to H^2_k \ominus P_1 H^2_k$ denote orthogonal projection. Using the above (see (27)) we may choose q_1 causal such that

$$\|W_1 - P_1 q_1\| = \|\Lambda(W_1, P_1)\|.$$

Now given this q_1, we choose a causal q_2 such that

$$\|W_2 - P_2(q_1 \otimes q_1) - P_1 q_2\| = \|\Lambda(W_2 - P_2(q_1 \otimes q_1)), P_1)\|.$$

Inductively, given $q_1, \ldots, q_{n-1}$, set

$$\hat{W}_n := \left(W_n - \sum_{2 \leq j \leq n} \sum_{i_1 + \cdots + i_j = n} P_j(q_{i_1} \otimes \cdots \otimes q_{i_j}) \right) \tag{30}$$

for $n \geq 2$. Then we may choose q_n such that

$$\|\hat{W}_n - P_1 q_n\| = \|\Lambda(\hat{W}_n, P_1)\|. \tag{31}$$

Note that in each step of the procedure, the new "weight" $\hat{W}_n$ is determined by the n-linear part W_n of the original weight, and the optimal causal parameters chosen previously (namely, $q_1, \ldots, q_{n-1}$). The "plant" P_1 remains fixed throughout the procedure. Thus if P_1 is rational, the iterative causal commutant lifting procedure takes place on the finite dimensional space $H_k^2 \ominus P_1 H_k^2$, and may therefore be reduced to *finite matrix computations*. This will be illustrated with an example in §13.

The following facts can be proven just as in [14] and [15], to which we refer the reader for the proofs. (See in particular [15, pp. 849–853].) First the causal iterative commutant lifting procedure converges:

PROPOSITION 10.1. *With the above notation, let $q^{(1)} := q_1 + q_2 + \cdots$. Then $q^{(1)} \in \mathcal{C}_l$.*

Next given any $q \in \mathcal{C}_l$, we can apply the causal iterative commutant lifting procedure to $W - P \circ q$. Now set

$$S_C(q)(\rho) := \sum_{n=1} \rho^n \|\Lambda(\hat{W}_n, P_1)\|.$$

Then we have the following result.

PROPOSITION 10.2. *Given $q \in \mathcal{C}_l$, there exists $\tilde{q} \in \mathcal{C}_l$, such that $S(\tilde{q}) \equiv S_C(q)$. Moreover $\tilde{q}$ may be derived from the causal iterated commutant lifting procedure.*

Moreover, as in [15] we have the following results.

PROPOSITION 10.3. *q is partially optimal if and only if $S(q) \cong S_C(q)$.*

THEOREM 10.4. *For given P and W as above, any $q \in \mathcal{C}_l$ is either partially optimal or can be ameliorated by a partially optimal compensating parameter.*

Finally we have the following result.

THEOREM 10.5. *Let P and W be single-input/single-output admissible operators. If the linear part of P is rational, then the partially optimal compensating parameter q_{opt} constructed by the iterated causal commutant lifting procedure is optimal.*

The proof of this last result is based on the uniqueness of the optimal interpolant in the case when $k = 1$, and when the space $H^2 \ominus P_1 H^2$ is finite-dimensional. In fact, the conclusion of Theorem 10.5 remains valid under the hypotheses that the operators ΠW_j, $j \geq 1$ and ΠP_i, $i \geq 2$ are compact (and $k = 1$). See [15].

11. Control interpretation of iterated lifting. We would like to mention here what we believe to be a very natural way of looking at the optimization procedure discussed above. For convenience, we will only treat SISO systems here.

We refer again to Fig. 1. We consider the problem of finding

$$\mu_\delta := \inf_C \sup_{\|v\| \leq \delta} \|[(I + P \circ C)^{-1} \circ W]v\|, \tag{32}$$

where we assume all the operators involved are admissible. Thus we are looking at a worst case disturbance attenuation problem where the energy of the signals v is required to be bounded by some prespecified level δ. (Of course in the linear case since everything scales, we can always without loss of generality take $\delta = 1$. For nonlinear systems, we must specify the energy bound a priori.) Again with the assumptions made in §8, we see that (32) is equivalent to the problem of finding

$$\mu_\delta = \inf_{q \in \mathcal{C}_l} \sup_{\|v\| \leq \delta} \|(W - P \circ q)v\|. \tag{33}$$

The iterated causal commutant lifting procedure gives an approach for approximating a solution to such a problem. Briefly, the idea is that we write

$$\begin{aligned} W &= W_1 + W_2 + \cdots, \\ P &= P_1 + P_2 + \cdots, \\ q &= q_1 + q_2 + \cdots, \end{aligned}$$

where W_j, P_j, q_j are homogeneous polynomials of degree j. Note that

$$\mu_\delta = \delta \inf_{q_1 \in H^\infty} \|W_1 - P_1 q_1\| + O(\delta^2), \tag{34}$$

where the latter norm is the operator norm (i.e., H^∞ norm). From the classical commutant lifting theorem we can find an optimal (linear, causal, time-invariant) $q_{1,\mathrm{opt}} \in H^\infty$ such that

$$\mu_\delta = \delta \|W_1 - P_1 q_{1,\mathrm{opt}}\| + O(\delta^2). \tag{35}$$

Now the iterative procedure gives a way of giving higher-order corrections to this linearization. Let us illustrate this now with the second-order correction. Indeed, having fixed now the linear part $q_{1,\mathrm{opt}}$ of q in (33), we note that

$$\begin{aligned} &W(v) - P(q(v)) - (W_1 - P_1 q_{1,\mathrm{opt}})(v) \\ &\quad = W_2(v) - P_2(q_{1,\mathrm{opt}}(v)) - P_1 q_2(v) + \text{higher-order terms.} \end{aligned}$$

Regarding $\hat{W}_2, P_2, q_2$ as linear operators on $H^2 \otimes H^2 \cong H^2(D^2, \mathbf{C})$ as above, we see that

$$\sup_{\|v\| \le \delta} \|(W - P \circ q)(v) - (W_1 - P_1 q_{1,\mathrm{opt}})v\| \le \delta^2 \|\hat{W}_2 - P_1 q_2\| + O(\delta^3),$$

where the "weight" $\hat{W}_2$ is given as in (30). The point of the iterative causal commutant lifting procedure is now to pick an optimal admissible $q_{2,\mathrm{opt}}$, and so on.

In short, instead of simply designing a linear compensator for a linearization of the given nonlinear system, this methodology allows us to explicitly take into account the higher-order terms of the nonlinear plant, and therefore increase the ball of operation for the nonlinear controller.

12. Rationality. A nice feature of the iterated procedure described above is that if we start out with rational data, then we derive compensating parameters at each step that are also rational. Thus the whole procedure is amenable to digitable implementation in such cases. Let us briefly review the notion of rationality in this context. See [14] for all the details.

Let $W : H^2_K(D^n) \to H^2_k$ be time-invariant and admit the row vector representation

$$W\Phi^* \cong [W_{0,\ldots,0}(z), W_{0,\ldots,1}(z), \ldots, W_{i_1,\ldots,i_n}(z), \ldots], \; (i_1, \ldots, i_n) \in I_n.$$

Then we say that W is *rational* if there exists a numerical polynomial $q(z) \neq 0$ such that

$$q(z)[W_{0,\ldots,0}(z), W_{0,\ldots,1}(z), \ldots, W_{i_1,\ldots,i_n}(z), \ldots]$$

is a row of matrix-valued polynomials of bounded degree. Moreover if W is causal, we say that W is *causal rational* if

$$W_c \cong [W_{0,\dots,0}(z), W^c_{0,\dots,1}(z), \dots, W^c_{i_1,\dots,i_n}(z), \dots]$$

is rational in the above sense.

The following result may be derived exactly as in [15, (see Thm. 8.7)].

THEOREM 12.1. *Notation as above. Suppose that the linear part of the plant is rational. Then the class of causal rational input/output operators is preserved under the causal iterated commutant lifting procedure.*

Hence for this important class of systems, we are reduced to rational finite-dimensional operations in carrying out our optimization procedure.

13. Example. In this section, we will give an example of our nonlinear design procedure. In what follows below, we set $H_{D^2} := H^2(D^2)$, the space of $\mathbf{C}$-valued analytic functions on the bidisc D^2 with square integrable boundary values. We should note that this example was first worked in [15] without the causality constraint that we impose now.

We let

$$W(z) = \frac{1-z}{2}$$

and $P = P_1 + P_2$ where P_1 is the operator given by multipication by z^2 (in the discrete Fourier domain), and

$$P_2(F) = \frac{1}{2\pi i}\int_{|\zeta|=1} F(z\zeta^{-1}, \zeta)\frac{d\zeta}{\zeta}$$

for $F \in H_{D^2} \cong H^2 \otimes H^2$. More precisely, as we explained above, we can regard a bilinear map P_2 on $H^2 \times H^2$ as a linear map on $H^2 \otimes H^2$, and then we identify $H^2 \otimes H^2$ with H_{D^2}. (The identification is given by $z \otimes 1 \to z_1$ and $1 \otimes z \to z_2$.) Notice that in the discrete-time domain, P_2 is just discrete Fourier transform of the "squaring" map, i.e., given the square integrable sequence $\{a_n\}$, we have that P_2 is the Fourier transform of the mapping $\{a_n\} \to \{a_n^2\}$. Thus it is clear that P_2 is causal.

We now apply our procedure to the weight W and the plant P. By slight abuse of notation, we let $W : H^2 \to H^2$ denote the operator defined by multiplication by W, and let $\Pi : H^2 \to H^2 \ominus P_1 H^2 =: H_1$ be orthogonal projection. We set $A_o := \Pi W | H_1$. Note that $H_1 \cong \mathbf{C}^2$, and that via this isomorphism, we have the identification

$$A_o = \begin{bmatrix} \frac{1}{2} & 0 \\ -\frac{1}{2} & \frac{1}{2} \end{bmatrix}.$$

However,

$$A_o^* A_o = \begin{bmatrix} \frac{1}{2} & -\frac{1}{4} \\ -\frac{1}{4} & \frac{1}{4} \end{bmatrix},$$

from which we get that $\|A_o\| = (\sqrt{5}+1)/2$, and that a maximal vector h_o (i.e., a vector such that $\|A_o h_o\| = \|A_o\|\|h_o\| \neq 0$) is given by

$$h_o := \begin{bmatrix} 1 \\ -\beta \end{bmatrix},$$

where $\beta := (\sqrt{5}-1)/2$. Using then the Sarason formula [24], we can compute that the optimal compensating parameter is

$$q_1 := \frac{\beta}{2(1-\beta z)}.$$

Of course, the above computation was based on standard linear H^∞-optimization theory. We now want to show how to get the optimal *causal* second-order compensating parameter.

For $F \in H_{D^2}$, let

$$F(z_1, z_2) = \sum_{j,k=0}^{\infty} F_{jk} z_1^j z_2^k.$$

Note that the action of the operator (see (30))

$$-\hat{W}_2 := \frac{4}{\beta^2} P_2(q_1 \otimes q_1)$$

on F is determined by its action on

$$F_{00} + \sum_{j=1}^{\infty} F_{j0} z_1^j + \sum_{k=1}^{\infty} F_{0k} z_2^k.$$

Thus to compute the row vector representing $-\hat{W}_2$, we need only compute

$$\begin{aligned}
&(-\hat{W}_2)(F_{00} + \sum_{j=1}^{\infty} F_{j0} z_1^j + \sum_{k=1}^{\infty} F_{0k} z_2^k) \\
&= \frac{1}{2\pi} \int_{|\zeta|=1} (\sum_{m\geq 0} \beta^m z^m \zeta^{-m})(\sum_{n\geq 0} \beta^n \zeta^n)(\sum_{\min\{j,k\}=0} F_{jk} z^j \zeta^{k-j}) \frac{d\zeta}{\zeta} \\
&= \sum_{\min\{j,k\}=0} (F_{jk}(\beta z)^{\max\{j,k\}})/(1-\beta^2 z).
\end{aligned}$$

We identify as above an operator $\Omega : H_K^2(D^n) \to H_k^2$ and its Fourier transformed version $\Omega\Phi^* : \ell^2(H_K^2) \to H_k^2$.

Therefore (under this identification),

$$\begin{aligned}
-\hat{W}_2 &\cong \frac{1}{1-\beta^2 z}[1, \beta z, \beta z, \beta^2 z^2, \ldots, \beta^n z^n, \ldots], \\
-\hat{W}_{2,c} &\cong \frac{1}{1-\beta^2 z}[1, \beta, \beta, \beta^2, \ldots, \beta^n, \ldots],
\end{aligned}$$

and

$$\|\hat{W}_2\| = \|\hat{W}_{2,c}\| \approx 2.4195.$$

Set $A = \Pi(-\hat{W}_{2,c})$, where $\Pi : H^2 \to H^2 \ominus z^2 H^2 =: H(z^2) \cong \mathbf{C}^2$ denotes orthogonal projection. Note that the compressed shift T on $H(z^2) \cong \mathbf{C}^2$ is given by the truncated Toeplitz operator

$$T = \begin{bmatrix} 0 & 0 \\ 1 & 0 \end{bmatrix}.$$

Using skew Toeplitz theory ([8], [17], [20]), we compute the norm of A and the corresponding optimal vector. Accordingly, we let $r(z) := 1 - \beta^2 z$. Then for $\rho > 0$, and for

$$\lambda := \frac{2-\beta}{(2\beta - 1)\rho^2},$$

we compute that

$$\begin{aligned} r(T)(\rho^2 I_{\mathbf{C}^2} - AA^*)r(T)^* &= \rho^2 r(T)r(T)^* - (1 + 2\sum_{i=1}^{\infty} \beta^{2i}) I_{\mathbf{C}^2} \\ &= (1-\beta)\rho^2 \begin{bmatrix} 1 + 1/\beta - \lambda & -1 \\ -1 & 3 - \lambda \end{bmatrix}. \end{aligned}$$

$\|A\|$ is given by the largest ρ such that the latter matrix is singular. Thus we see that

$$\|\Pi(-\hat{W}_{2,c})\| = \|A\| \approx 1.8079,$$

which is the optimal *causal* performance. If we drop the causality requirement, then we get that

$$\|\Pi(-\hat{W}_2)\| \approx 1.4314.$$

(Of course, with the additional constraint the norm of the optimal dilation increases.)

Let

$$y_o(z) := 1 + \left(1 + \frac{1}{\beta} - \lambda\right) z \in H(z^2),$$

so that we may regard

$$y_o = \begin{bmatrix} 1 \\ 1 + \frac{1}{\beta} - \lambda \end{bmatrix}$$

under the identification $H(z^2) \cong \mathbf{C}^2$. Then it is easy to compute that

$$r(T)(\|A\|^2 I_{\mathbf{C}^2} - AA^*)r(T)^* y_o = 0.$$

Therefore $r(T)^* y_o$ is a maximal vector of A^*. But from the previous section (see (28)), the optimal dilation $B_{\text{opt},c}$ of A is

$$\begin{aligned} B_{\text{opt},c} &\cong \frac{\overline{A^* r(T)^* y_o}}{\overline{r(T)^* y_o}} \\ &= \frac{(3-\lambda)z + 1}{(1 + \frac{1}{\beta} - \lambda)z + 1} [1, \beta, \beta, \beta^2, \beta^2, \ldots]. \end{aligned}$$

Thus the optimal *causal* dilation B_{opt} of $\Pi(-\hat{W}_2)$ is

$$B_{\text{opt}} \cong \frac{(3-\lambda)z + 1}{(1 + \frac{1}{\beta} - \lambda)z + 1} [1, \beta z, \beta z, \beta^2 z^2, \ldots].$$

The optimal causal interpolant q_2 is derived from

$$-\frac{4}{\beta^2} P_2(q_1 \otimes q_1) - z^2 q_2 = -B_{opt},$$

which gives that

$$q_2 \cong \frac{(\lambda - 3)\beta^2}{(1-\beta^2 z)((1+1/\beta-\lambda)z+1)}[1, \beta z, \beta z, \beta^2 z^2, \ldots].$$

Now set $q^{(2)} := q_1 + q_2$, the optimal second-order compensating parameter, and $\hat{q}^{(2)} := q^{(2)}W^{-1}$. The resulting controller is given by $C^{(2)} = \hat{q}^{(2)} \circ (I - P \circ \hat{q}^{(2)})^{-1}$. Note that it is not necessary to explicitly compute $C^{(2)}$, since it can be implemented in a feedback loop with components P and $\hat{q}^{(2)}$ as in [27].

14. Concluding remarks. In this paper, we have given an iterative approach for the construction of optimal causal compensators for input/output operators described by analytic mappings. Our procedure generalizes weighted sensitivity H^∞ minimization in a straightforward natural way. Hence, it may be regarded as a weighted nonlinear inversion procedure.

In contrast to our previous work using power series approaches ([3], [4], [14], [15]), we can now *guarantee causality* a priori. Moreover, the computation of a causal compensator can be reduced to classical dilation theory, and in fact the skew Toeplitz techniques of [8], [17], and [20] provide an explicit computational methodology.

The example which we have worked out here, has been given just for the purpose of illustrating our procedure. We plan to work out a more complicated and realistic problem, the details of which will be given in an upcoming report.

REFERENCES

[1] W. Arveson, *Interpolation problems in nest algebras*, J. Funct. Anal., 20 (1975), pp. 208–233.

[2] M. Atiyah and I. MacDonald, *Introduction to Commutative Algebra*, Addison-Wesley, New York, 1969.

[3] J. Ball, C. Foias, J. W. Helton, and A. Tannenbaum, *On a local nonlinear commutant lifting theorem*, Indiana J. Math., 36 (1987), pp. 693–709.

[4] ———, *Nonlinear interpolation theory in H^∞*, in Modelling, Robustness, and Sensitivity in Control Systems, Ruth Curtain, ed., Springer-Verlag, New York, 1987.

[5] ———, *A Poincaré-Dulac approach to a nonlinear Beurling-Lax-Halmos theorem*, J. Math. Anal. Appl., 139 (1989), pp. 496–514.

[6] J. Ball and J. W. Helton, *Sensitivity bandwidth optimization for nonlinear feedback systems*, Tech. Report, Dept. of Math., Univ. of California at San Diego, 1988.

[7] ———, *H^∞ control for nonlinear plants: connections with differential games*, IEEE Conference on Decision and Control, Tampa, Florida, 1989, pp. 956–962.

[8] H. Bercovici, C. Foias, and A. Tannenbaum, *On skew Toeplitz operators*. I, Oper. Theory Adv. Appl., 29 (1988), pp. 21–44.

[9] S. Boyd and L. Chua, *Fading memory and the problem of approximating nonlinear operators with Volterra series*, IEEE Trans. Circuits Systems, CAS-32 (1985), pp. 1150–1161.

[10] J. Doyle, B. Francis, and A. Tannenbaum, *Feedback Control Theory*, MacMillan, New York, 1991.

[11] C. Foias, *Contractive intertwining dilations and waves in layered media*, Proc. Internat. Congress Mathematicians, Helsinki, 2 (1978), pp. 605–613.

[12] C. Foias and A. Frazho, *The Commutant Lifting Approach to Interpolation Problems*, Birkhauser-Verlag, Boston, 1990.

[13] C. Foias, C. Gu, and A. Tannenbaum, *Intertwining dilations, intertwining extensions and causality*, Acta Sci. Math. (Szeged), 57 (1993), pp. 101–123.

[14] C. Foias and A. Tannenbaum, *Iterated cummutant lifting for systems with rational symbol*, *Oper. Theory Adv. Appl.*, 41 (1989), pp. 255–277.

[15] ———, *Weighted optimization theory for nonlinear systems*, SIAM J. Control Optim., 27 (1989), pp. 842–860.

[16] ———, *Causality in commutant lifting theory*, J. Funct. Anal., 118 (1993), pp. 407–441.

[17] C. Foias, A. Tannenbaum, and G. Zames, *Some explicit formulae for the singular values of certain Hankel operators with factorizable symbol*, SIAM J. Math. Anal., 19 (1988), pp. 1081–1089.

[18] B. Francis, *A Course in H^∞ Control Theory*, McGraw-Hill, New York, 1981.

[19] B. Francis and A. Tannenbaum, *Generalized interpolation theory in control*, Math. Intelligencer, 10 (1988), pp. 48–58.

[20] C. Gu, *Eliminating the genericity conditions in the skew Toeplitz operator algorithm for H^∞ optimization*, SIAM J. Math. Anal., 23 (1992), pp. 1623–1636.

[21] E. Hille and R. Phillips, *Functional Analysis and Semigroups*, AMS Colloquium Publications, Vol. 31, American Mathematical Society, Providence, Rhode Island, 1957.

[22] J. W. Helton, *Broadbanding: gain equalization directly from data*, IEEE Trans. Circuits Systems, CAS-28 (1981), pp. 1125–1137.

[23] S. Parrott, *On the quotient norm and the Sz. Nagy-Foias lifting theorem*, J. Funct. Anal., 30 (1970), pp. 311–328.

[24] D. Sarason, *Generalized interpolation in H^∞*, Trans. Amer. Math. Soc., 127 (1967), pp. 179–203.

[25] B. Sz.-Nagy and C. Foias, *Harmonic Analysis of Operators on Hilbert Space*, North-Holland, Amsterdam, 1970.

[26] A. J. van der Schaft, *L_2 gain analysis of nonlinear systems and nonlinear state feedback H_∞ control*, IEEE Trans. Automat. Control, 37 (1992), pp. 770–784.

[27] G. Zames, *Feedback and optimal sensitivity: model reference transformations, multiplicative seminorms, and approximate inverses*, IEEE Trans. Automat. Control, AC-26 (1981), pp. 301–320.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 33, No. 1, pp. 208–226, January 1995

© 1995 Society for Industrial and Applied Mathematics
010

MODEL UNCERTAINTY IN DISCRETE EVENT SYSTEMS*

STANLEY YOUNG[†] AND VIJAY K. GARG[‡]

Abstract. Earlier work concerning control of discrete event systems usually assumed that a correct model of the system to be controlled was available. A goal of this work is to provide an algorithm for determining the correct model from a set of models. The result of the algorithm is a finite language that can be used to test for the correct model or notification that the remaining models cannot be controllably distinguished. We use the finite state machine model with controllable and uncontrollable events presented by Ramadge and Wonham.

Key words. discrete event systems, system identification

AMS subject classifications. 93A, 93B

1. Introduction.

A discrete event system (DES) is one that responds to distinct events occurring at asynchronous times [7]. Examples of such systems include computer networks, manufacturing systems, and other dynamic systems that require high-level coordinated control. There has been some success recently in developing a theory for the control of such systems (see [13] and the references therein). Most of this work has assumed that an accurate model for the system of interest is available.

The motivation for this work is the desire to control systems in the presence of uncertainty in the model of the system and environment in which the system operates. Part of this work is an extension of learning and inference theory [6], [2], [16] to the domain of discrete event systems. This work is also related to recent results concerning the determination of a system model when certain assumptions are made about the model and type of experiments [14], [15]. In both the learning theory and system determination work, an assumption is that all events are controllable. The uncontrollability of certain events figures prominently in this work. The approach taken in this paper is similar to the approach used for system identification in [17] in that any model that is falsified is dropped from consideration as a correct model.

There are many different types of uncertainty that might occur in a system model. To discuss such uncertainties, a model representation must be chosen. In this work, we investigate uncertainty in a deterministic finite state machine. An example of such uncertainty is an uncertainty in the transitions of a system that can be described as a state that has a single event specified as providing transitions to at least two different resulting states; however, only one of the transitions is actually present in the system. Other examples are discussed in §3. Such uncertainty results in multiple models of the system that might potentially be correct. The goal is to specify conditions and algorithms that enable the identification of the correct model in a finite number of transitions despite the presence of uncontrollable actions. In particular, an algorithm provides either notification that no more models may be controllably distinguished or a finite distinguishing language that can be used to remove an incorrect model.

* Received by the editors September 25, 1991; accepted for publication (in revised form) August 24, 1993. This work was supported in part by National Science Foundation grant CCR-911065, and in part by the University Research Institute, University of Texas at Austin. A preliminary version of this paper was presented at the 1991 Institute of Electrical and Electronics Engineers International Symposium on Intelligent Control.

[†] Applied Research Laboratories, University of Texas, Austin, Texas 78712 (syoung@titan.tsd.arlut.utexas.edu).

[‡] Department of Electrical and Computer Engineering, University of Texas, Austin, Texas 78712 (vijay@pine.ece.utexas.edu).

Section 2 describes the method used to model the plant and the relevant controllability results. Section 3 gives some examples of how a set of potentially correct models for a system might arise. Section 4 describes the concepts and techniques used to identify a correct model from a given set of models. Section 5 provides example applications of the results.

2. Description of the model.

We use the deterministic finite state machine as a model for system behavior. In what follows, only the main features of the finite state machine model related to this work are covered in a condensed manner. A more complete development related to the finite state machine model can be found in [5], [9]. More complete descriptions of the controllability and related results can be found in [13], [3], [11].

2.1. Finite state machines and regular languages.

A finite state machine is represented by either a four or five tuple. Specifically, if M is a finite state machine (FSM), then we write $M = (Q, A, \delta, q_0)$ or $M = (Q, A, \delta, q_0, Q_m)$, where

$$\begin{array}{lcl} Q & = & \text{a finite set of states,} \\ A & = & \text{a finite set of transition labels or events,} \\ \delta & = & \text{the transition function, } \delta : Q \times A \to Q, \\ q_0 & = & \text{the initial state, } q_0 \in Q, \text{ and} \\ Q_m & = & \text{the marked states, } Q_m \subseteq Q. \end{array}$$

The transition function δ is, in general, a partial function: $\delta(q, \sigma)!$ denotes that the transition event σ is defined from state q. The marked states signify a subset of the state set that is used to determine acceptance of a given string. A string is *accepted* if the machine executing the string stops in a marked state.

A finite state machine, $P = (Q, A, \delta, q_0, Q_m)$, can also be represented as a directed graph $M = (Q, T)$, where Q and T are the sets of nodes and arcs, respectively, or states and transitions in this instance [1], [4]. Q is the set of states in the machine and $T \subseteq Q \times A \times Q$ is the set of transitions. If $q_1, q_2 \in Q$ and $\delta(q_1, \sigma) = q_2$, then one denotes the transition by the three tuple (q_1, σ, q_2).

A^* is used to denote the set of all finite sequences of symbols from the alphabet A. A language is a set of strings of elements from an alphabet. If $u \in A^*$, then $|u|$ denotes the length of u, $u(j)$ denotes the jth element of the string, and the set $pr(u)$ denotes the set of all strings that are prefixes of u, i.e. for $u \in A^*$

$$pr(u) = \{s \in A^* | s = u(1) \dots u(k), 0 < k \leq |u|\}.$$

The notation $s \leq u$ is used to denote that s is a prefix of u. Note that the empty string ε is the length zero prefix of all strings. The concept of prefix can be extended to a language in the following manner. The *prefix closure* of a language L is defined by

$$\overline{L} = \{w \in A^* | \exists u \in L : w \leq u\}.$$

We use the notation $s \in ppr(u)$ or $s < u$ to signify that s is a proper prefix of u, i.e., that $s \leq u$ and $s \neq u$. This concept is extended to a language, $L \subseteq A^*$, in the following manner:

$$ppr(L) = \{s \in A^* | \exists u \in L : s < u\}.$$

For this work, we restrict our attention to the class of regular languages that is a strict subset of the class of formal languages. A basic result relates regular languages and finite state machines: a language $L \subseteq A^*$ is regular if and only if it is generated by a finite state machine [9]. Language $L_m(M)$ is the language *marked* or *recognized* by machine $M = (Q, A, \delta, q_0, Q_m)$ if

$$w \in L_m(M) \Leftrightarrow \delta(q_0, w) \in Q_m,$$

where δ is extended in the usual manner, $\delta : Q \times A^* \longrightarrow Q$. A language $L(M)$ is the language *generated* by machine M if

$$w \in L(M) \Leftrightarrow \delta(q_0, w)!.$$

The product machine is a single machine that can be used to represent the synchronous behavior of two original machines. If machines $M_1 = (Q_1, A, \delta_1, q_{1,0}, Q_{1,m})$ and $M_2 = (Q_2, A, \delta_2, q_{2,0}, Q_{2,m})$ have the same event set A then the product of the two machines is denoted

$$M_1 \| M_2 = (Z, A, \delta_{\|}, z_0, Z_m),$$

where

$$Z = Q_1 \times Q_2 \text{ and } z_0 = (q_{1,0}, q_{2,0}),$$

$$\delta_{\|}((q_1, q_2), \sigma) = \begin{cases} (\delta_1(q_1, \sigma), \delta_2(q_2, \sigma)) & \text{if defined} \\ \text{undefined} & \text{otherwise,} \end{cases}$$

and

$$Z_m = Q_{1,m} \times Q_{2,m}.$$

The languages generated and marked by the product machine have a specific relation to the languages of the machines from which they are composed. If $M_{\|} = M_1 \| M_2$ then

$$L(M_{\|}) = L(M_1) \cap L(M_2) \quad \text{and} \quad L_m(M_{\|}) = L_m(M_1) \cap L_m(M_2).$$

2.2. Control of discrete event systems. The event set A can be partitioned into two sets A_c and A_u, representing controllable and uncontrollable events, respectively. A language K is *controllable* with respect to language L if

$$\overline{K}.A_u \cap L \subseteq \overline{K},$$

where $a.b$ denotes concatenation and is often denoted by ab. A supervisor for a plant, modeled with finite state machine P, where $L = L(P)$, is a map

$$f : L \rightarrow 2^A,$$

which specifies a set of inputs enabled by the supervisor that can be applied as a function of the string in L of events that the plant has previously executed. The closed-loop system consisting of a supervisor f and plant P has the closed-loop behavior denoted by L_f, and is defined as follows.

1. $\varepsilon \in L_f$,
2. $w\sigma \in L_f$ if and only if $w \in L_f, \sigma \in f(w)$, and $w\sigma \in \overline{L}$.

A supervisor f is *complete* with respect to a given plant P, with $L = L(P)$, if all uncontrolled actions of the plant are respected, i.e., if $x \in L_f$ and $x\sigma_u \in L$, then $x\sigma_u \in L_f$, where $\sigma_u \in A_u$ and L_f is the closed-loop behavior as discussed above. The following result is a basic theorem relating these concepts.

THEOREM 2.1 ([12]). *For nonempty $K \subseteq L$, there exists a complete supervisor f such that $L_f = K$ if and only if K is prefix closed and controllable.*

The region of weak attraction, as discussed in [3], [11], can be directly related to distinguishing different machines. The region of weak attraction for a specified set of states can be described informally as the set of states from which the system can be controlled so as to enter the set of specified states in a finite number of transitions.

The *region of weak attraction* $\Omega_M(G)$ for a machine $M = (Q, A, \delta, q_0)$, or in graph notation, $M = (Q, T)$, and a specified subset of states $G \subseteq Q$ can be determined by the algorithm in [3]. For a specific calculation of the region of weak attraction of a given set of states G, the transitions used in its construction are denoted by $T_\Omega(G)$. This algorithm builds the region of weak attraction starting from G. Each iteration of the algorithm adds states to the region defined in the previous iteration. A state is added to the region of weak attraction only if there is an event σ that describes a transition into the region defined in the previous iteration and there does not exist an uncontrolled event to a state not in the region defined by the previous iterations of the algorithm. The transition labeled by this σ is added to $T_\Omega(G)$ as are the uncontrolled transitions from this state. The states in $\Omega_M(G)$ are well defined; as discussed in [3], the transitions chosen for $T_\Omega(G)$ are not necessarily uniquely defined. The algorithm is guaranteed to terminate by the finite state description of the machine. An efficient algorithm in [11] computes the region of weak attraction in $O(|Q| \cdot |A|)$ time.

The characteristics of the region of weak attraction are most easily described by certain conditions on the directed graph that describes the finite state machine. Let the machine be described by the graph $M = (Q, T)$ with $G \subseteq Q$. The region of weak attraction satisfies three main criteria as described in the following proposition.

PROPOSITION 2.2 ([3]). *$M' = (\Omega_M(G), T_\Omega(G)) \subseteq (Q, T)$ if and only if*

(1) *M' is G-connected,*

(2) *M' is realizable,*

(3) *$M' - G$ is acyclic.*

A graph $M = (Q, T)$ is *F-connected* if $F \subseteq Q$ and from every state in Q there exists a path to a state in F. A subgraph $M' = (Q', T') \subseteq (Q, T)$ is *realizable* if

$$(((q_1, \sigma, q_2) \in T) \wedge (q_1 \in Q') \wedge (\sigma \in A_u)) \Rightarrow ((q_1, \sigma, q_2) \in T').$$

A realizable subgraph includes all uncontrollable arcs that are defined from any state in the state set of the subgraph.

If the initial state q_0 for the machine M is in the region of weak attraction, i.e., $q_0 \in \Omega_M(G)$, then we also define a machine based on the region of weak attraction $M_\Omega(G)$, where $M_\Omega(G)$ is formally defined by the tuple:

$$M_\Omega(G) = (Q, A, \delta_\Omega, q_0, G).$$

And $\delta_\Omega(q_1, \sigma) = q_2$ if and only if (q_1, σ, q_2) is an arc defined in the construction of the region of weak attraction, i.e., $(q_1, \sigma, q_2) \in T_\Omega(G)$. With M_Ω defined, the language recognized by the resulting machine is denoted by $L(M_\Omega(G))$. As mentioned above and in [3], $M_\Omega(G)$, and hence $L(M_\Omega(G))$, is not unique.

3. Model uncertainty. Uncertainties in the plant model provide a set of models that are potentially correct models of the plant. Each model in this set is obtained by assuming that the uncertainty results from a specific lack of knowledge about the structure of the plant.

Example 3.1. Consider an automatic guided vehicle system guided by wires in the floor of a manufacturing facility. The model of the guidance system may contain errors. Each error will produce an uncertain model of the correct system. For instance, two branch nodes in the wiring may be combined into a single node in the model, an extra branch may be in the model that is not installed in the plant, or the model may be lacking a branch that is installed in the plant. Each of these errors generates an uncertain model that can be used to define a set of potentially correct models.

Example 3.2. Model (A) in Fig. 1 gives an example of a system with uncertainty in the transitions. For this transition uncertainty, there is a single state q_0 in the model that has "b" transitions defined to k different states, $q_1, \ldots, q_k$; yet, in the actual system, only one of these "b" transitions is defined.

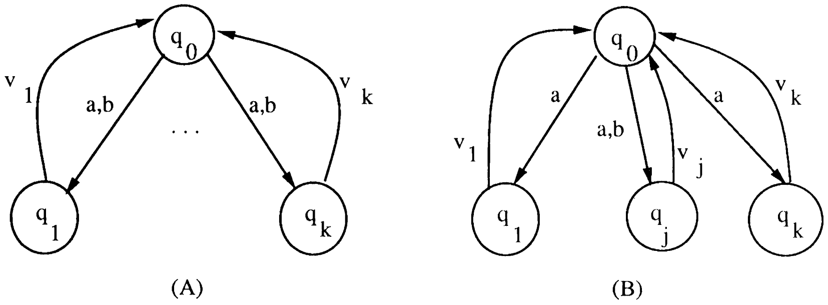


FIG. 1. *Model for transition uncertainty.* (A) *Model with transition uncertainty.* (B) *One of the possibly correct models.*

Example 3.3. Assume that the set of events that a system can accomplish is known and that there is a known upper bound on the size of the state space. Using these two assumptions, we can construct all possible models for the system. After all unique models have been constructed, a technique is required to generate tests that can distinguish the correct model.

4. Distinguishing between models. We present deterministic techniques that provide an easily checked condition and an algorithm for correctly removing inconsistent models from consideration and identifying the correct model.

Certain concepts will provide a unified framework for the development that follows. For the following definitions, we assume that there are models $M = (Q, A, \delta, q_0)$, $M_1 = (Q_1, A, \delta_1, q_{1,0})$, and $M_2 = (Q_2, A, \delta_2, q_{2,0})$ that have states $q_1 \in Q_1$, $q_2 \in Q_2$ and an event $\sigma \in A$.

DEFINITION 4.1. *Given M_1 and M_2, the predicate* different(z, w) *holds if*

$$(\delta_1(q_1, w)! \wedge \neg\delta_2(q_2, w)!) \vee (\neg\delta_1(q_1, w)! \wedge \delta_2(q_2, w)!),$$

where

$$z = (q_1, q_2) \quad \textit{and} \quad w \in A^*.$$

This predicate depends on both the state in the product machine and the string chosen to differentiate the states that make up the product state.

Example 4.1. For the machines defined in Example 3.2, let M_i and M_j be the possible models that have the b transition defined to states q_i and q_j, respectively. $\text{different}((q_{i0}, q_{j0}), bv_j)$ holds, whereas $\text{different}((q_{i0}, q_{j0}), b)$ does not hold, where (q_{i0}, q_{j0}) is the product initial state.

In the following $A \triangle B$ denotes the symmetric difference of the two sets A and B, i.e., $A \triangle B = (A \cup B) \setminus (A \cap B)$.

DEFINITION 4.2. *A string w is a* distinguishing string *for languages L_1 and L_2 if*

$$(w \in L_1 \triangle L_2).$$

DEFINITION 4.3. *A string w is a* minimally distinguishing string *for languages L_1 and L_2 if*

$$(w \in L_1 \triangle L_2) \wedge (ppr(w) \subseteq L_1 \cap L_2).$$

A minimally distinguishing string is minimal in the sense that no substring is also a distinguishing string.

DEFINITION 4.4. *A nonempty language L is a* distinguishing language *for L_1 and L_2 if $w \in L$ implies that w is a minimally distinguishing string for L_1 and L_2. Hence, a string in a distinguishing language is one that uses the last event to distinguish between L_1 and L_2. Note that, in general, there is not a unique distinguishing language for two machines M_1 and M_2.*

Example 4.2. Let $L_1 = a^*$ and $L_2 = a^*b^*$ be two languages that describe the behavior of two possible models of a black box machine. For these languages, $L_1 \cup L_2 = a^*b^*$ and $L_1 \cap L_2 = a^*$; consequently, $L_1 \triangle L_2 = a^*b^*b$. For a distinguishing language L to satisfy $(ppr(L) \subseteq L_1 \cap L_2)$, we must have that $L \subseteq a^*b$.

Observe that if the machine executes the final b, then L_2 is the correct language, and if not, then L_1 is the correct language.

For languages generated by a state machine, we have the following result.

PROPOSITION 4.5. *Let L_1 and L_2 be languages generated by the machines M_1 and M_2. There exists a distinguishing language L for L_1 and L_2 if and only if*

$$(\exists w \in A^* : \text{different}(z_0, w)), \text{ where } z_0 = (q_{1,0}, q_{2,0}).$$

Proof. $\Longrightarrow$ Let $w \in L$. Since $w \in L_1 \triangle L_2$, we immediately have that $\text{different}(z_0, w)$ holds. $\Longleftarrow$ Let w be the string that satisfies $\text{different}(z_0, w)$. From the definition of $\text{different}(\cdot, \cdot)$, there is some $v \in L_1 \cap L_2$ such that $v < w$ and a $\sigma \in A$ such that $v\sigma \notin L_1 \cap L_2$ and $v\sigma \leq w$. Set $L = \{v\sigma\}$. It is clear that L is a distinguishing language for L_1 and L_2. □

Example 4.3. For the machines in Example 4.1, $\text{different}((q_{i0}, q_{j0}), bv_i)$ holds; consequently, there is a distinguishing language, $L = bv_i + bv_j$, where $a + b = \{a, b\}$.

4.1. Distinguishing between two models. Assume that machine M has an uncertainty that causes the set of potentially correct models to consist of the models M_1 and M_2. We assume that one of these models is correct. The models are specified by the following tuples:

$$M_1 = (Q_1, A, \delta_1, q_{1,0}) \quad \text{and} \quad M_2 = (Q_2, A, \delta_2, q_{2,0}).$$

The languages generated by these models are referenced by $L(M_1)$ and $L(M_2)$. We also refer to the standard synchronous product machine:

$$M_{||} = M_1 || M_2 = (Z, A, \delta_{||}, z_0).$$

The set of states in the product machine that can be used to controllably distinguish the two models is defined in the following manner.

DEFINITION 4.6. *G is the* controllably distinguishing set *of states for M_1 and M_2 if*

$$G = \{z \in Z| \ (\exists \sigma \in A: \ \text{different}(z,\sigma)) \wedge (\neg\exists \sigma_u \in A_u : \delta_{||}(z,\sigma_u)!) \ \},$$

where $M_1 \| M_2 = (Z, A, \delta_{||}, z_0)$.

A particular event that can be used to distinguish two states is called a *controllably distinguishing event.* Hence, an event σ is a controllably distinguishing event for z if *different*(z,σ) holds and there is not an uncontrolled event σ_u defined in the product machine from z.

Example 4.4. For the machines given in Fig. 2, let $A_u = \{b\}$. The controllably distinguishing state set is $G = \{(q_1, q), (q_2, q)\}$. In this example, the set of controllably distinguishing states is the entire product space.

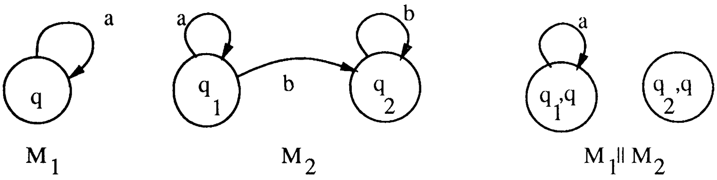


FIG. 2. *Machines for controllably distinguishing state set calculation.*

Example 4.5. Consider the same machines as Example 4.4, but let $A_u = \{a\}$. In this instance, the set of controllably distinguishing states is $G = \{(q_2, q)\}$.

To controllably distinguish states in the product machine, a string must be found which leads to a state in G. Note that G is a superset of states that can be used to distinguish states in the product machine and reached from the initial state. This inclusion is a result of defining G to be all states in the product machine that have controllably distinguishing events defined without consideration of reachability constraints.

Proposition 4.7 states that there is a finite controllable method for distinguishing between two finite state machines if and only if the initial state of the product machine is in the region of weak attraction of the set of states that can be used to distinguish between the two machines.

PROPOSITION 4.7. *Let* $M_1 = (Q_1, A, \delta_1, q_{1,0})$ *and* $M_2 = (Q_2, A, \delta_2, q_{2,0})$, *be two machines. There exists a finite language L that satisfies*

1. *L is controllable with respect to* $L(M_1) \cap L(M_2)$,
2. *L is a distinguishing language for* $L(M_1)$ *and* $L(M_2)$,

if and only if

$$z_0 \in \Omega_{M_1 \| M_2}(G),$$

where

1. $z_0 = (q_{1,0}, q_{2,0})$,
2. *G is the controllably distinguishing set of states, and*
3. $\Omega_{M_1\|M_2}(G)$ *is the region of weak attraction of G in* $M_1\|M_2$. *for* M_1 *and* M_2.

Proof. $\Longleftarrow$ We will demonstrate a language L that satisfies Definition 4.4 with respect to $L(M_1)$ and $L(M_2)$, and is controllable with respect to $L(M_1) \cap L(M_2)$.

Since $z_0 \in \Omega_{M_1\|M_2}(G)$, we can define a machine $M_\Omega(G)$, as discussed in §2.2, and let $L_m(M_\Omega)$ be the language marked by this machine with G as the marked states. Define $X \subseteq A$ such that if $t \in L_m(M_\Omega)$ then there exists $\sigma \in X$ to satisfy the definition of G, i.e., that $(\delta_{\|}(z_0,t) = z) \wedge (z \in G) \Rightarrow \neg\delta_{\|}(z,\sigma)!$. Such a symbol is defined for every string in $L_m(M_\Omega)$ by the definition of G. Further define $L = L_m(M_\Omega)X$ by $t\sigma \in L$ if $t \in L_m(M_\Omega)$, $z = \delta_{\|}(z_0,t)$, and σ is a controllably distinguishing event for z.

We must show that this language is finite, satisfies the definition for controllability, and satisfies the definition of a distinguishing language.

1. L finite:

$$\begin{array}{ll}
 & (\Omega_{M_{\|}}(G), T_\Omega(G)) - G \text{ acyclic}\\
\Rightarrow & \qquad \{\text{property of } \Omega_M(G)\}\\
 & \text{no cycles in } M_\Omega - G\\
\Rightarrow & \\
 & \neg\exists s \in L_m(M_\Omega) : |s| \geq |\Omega_{M_{\|}}(G)|\\
\Rightarrow & \\
 & L \text{ finite.}
\end{array}$$

2. L controllable with respect to $L(M_1) \cap L(M_2)$: Let $(Z_\Omega, T_\Omega) = (\Omega_{M_{\|}}(G), T_\Omega(G))$ and assume that $(t \in \overline{L}) \wedge (\sigma \in A_u) \wedge (t\sigma \in L(M_1) \cap L(M_2))$.

$$\begin{array}{ll}
 & t \in \overline{L} \wedge z_0 \in \Omega_{M_{\|}}(G)\\
\Rightarrow & \qquad \{L = L_m(M_\Omega).X \text{ and definition of } L_m(M_\Omega)\}\\
 & \delta_{\|}(z_0,t) \in Z_\Omega\\
\Rightarrow & \qquad \{t\sigma \in L(M_1) \cap L(M_2)\ ,\ \Omega_M(G) \text{ realizable, } z = \delta_{\|}(z_0,t)\}\\
 & \exists z' \in Z_\Omega : (z,\sigma,z') \in T_\Omega\\
\Rightarrow & \qquad \{\text{definition of } L_m(M_\Omega)\}\\
 & t\sigma \in L(M_\Omega)\\
\Rightarrow & \\
 & t\sigma \in \overline{L}.
\end{array}$$

Hence L is controllable with respect to $L(M_1) \cap L(M_2)$.

3. L non-empty:

$$\begin{array}{ll}
 & z_0 \in \Omega_{M_{\|}}(G)\\
\Rightarrow & \qquad \{M_\Omega \text{ is } G\text{-connected}\}\\
 & \exists v \in A^* : \delta_{\|}(z_0,v) \in G\\
\Rightarrow & \qquad \{\text{definition of } G \text{ and } X\}\\
 & \exists t = v\sigma \in A^* : t \in L_m(M_\Omega).X\\
\Rightarrow & \\
 & L \neq \emptyset.
\end{array}$$

4. $ppr(L) \subseteq L(M_1) \cap L(M_2)$:

$$\begin{array}{ll}
 & L = L_m(M_\Omega).X\\
\Rightarrow & \\
 & ppr(L) = L(M_\Omega)\\
\Rightarrow & \qquad \{\text{definition of } M_\Omega \text{ and } M_{\|}\}\\
 & ppr(L) \subseteq L(M_1\|M_2)\\
\Rightarrow & \qquad \{\text{property of regular languages and FSM}\}\\
 & ppr(L) \subseteq L(M_1) \cap L(M_2).
\end{array}$$

5. $L \subseteq L(M_1)\triangle L(M_2)$:

$$\begin{array}{ll} & (t \in L) \wedge (t = v\sigma) \\ \Rightarrow & \qquad \{\text{definition of } G \text{ and } L\} \\ & z = \delta_{\|}(z_0, v) \in G \\ \Rightarrow & \qquad \{\text{definition of } G \text{ and } X\}, \\ & \textit{different}(z, \sigma) \\ \Rightarrow & \qquad \{\text{definition of } G\} \\ & t \in L(M_1)\triangle L(M_2). \end{array}$$

The last three items combine to show that L satisfies the conditions for being a distinguishing language.

$\Longrightarrow$: Let (V, T) denote the graph representation of the product machine $M_1\|M_2$. Consider the subgraph (V_0, T_0) of (V, T) obtained by running all strings in L on (V, T). Define (V_0, T_0) by

$$\begin{array}{l} V_0 = \{z \in Q_1 \times Q_2 : (\exists w \in \overline{L} : \delta_{\|}(z_0, w) = z)\}, \\ T_0 = \{(z_1, \sigma, z_2) \in T : (\exists w \in \overline{L} : (\delta_{\|}(z_0, w) = z_1) \wedge (w\sigma \in \overline{L}))\}. \end{array}$$

Any state in the product machine reached by a string in the closure of L is included in V_0 and any transition traversed by a string in the closure of L is included in T_0. Since L is finite and $ppr(L) \subseteq L_1 \cap L_2$, it is clear that (V_0, T_0) is well defined and that a simple algorithm will generate this subgraph.

Consider the states reached by string in L/A. (Recall that L/A denotes strings in L with the last symbol removed.) The following lemma follows easily from the definition of a distinguishing language and controllability.

LEMMA 4.8. *With L as assumed in the statement of the proposition,*

$$\forall w \in L/A : \delta_{\|}(z_0, w) \in G$$

where z_0 and G are as defined in the statement of Proposition 4.7.

The following lemma provides the realizability of (V_0, T_0) with respect to (V, T) and A_u.

LEMMA 4.9. *(V_0, T_0) is realizable with respect to (V, T) and A_u.*

Proof. Let $z_1 = \delta_{\|}(z_0, w)$ where $w \in ppr(L)$, $(z_1, \sigma, z_2) \in T$, and $\sigma \in A_u$. From these facts, we must show that $(z_1, \sigma, z_2) \in T_0$. Note that from the construction of V_0, we have that $w \in ppr(L)$ implies that $z_1 \in V_0$.

$$\begin{array}{ll} & (z_1, \sigma, z_2) \in T \\ \Rightarrow & \\ & \delta_{\|}(z_0, w\sigma)! \\ \Rightarrow & \qquad \{L \text{ controllable with respect to } L_1 \cap L_2\} \\ & w\sigma \in ppr(L) \\ \Rightarrow & \\ & (z_1, \sigma, z_2) \in T_0. \qquad \Box \end{array}$$

By Lemma 4.8, (V_0, T_0) is G-connected. By Lemma 4.9, (V_0, T_0) is realizable. If there are no cycles in (V_0, T_0), then by Proposition 2.2 we have that $z_0 \in \Omega_{M_1\|M_2}(G)$ as desired.

If there are cycles, more work is required. The idea is to disable a cycle and show that the remaining subgraph is still G-connected and realizable. Since (V_0, T_0)

is a finite graph, there are only finitely many cycles; consequently, after disabling all cycles, we have a subgraph remaining that is G-connected, realizable, and acyclic. By Proposition 2.2, we have that $z_0 \in \Omega_{M_1 \| M_2}(G)$ as desired.

Now we must show that cycles may be removed while retaining the connectedness and realizability of the subgraph (V_0, T_0).

We start by classifying all transitions in T_0. A transition is included in class $\mathcal{C}$ if it must be included in T_0 for controllability reasons, i.e., if the subgraph would lose the realizability characterization by not having a specific transition, then that transition is included in $\mathcal{C}$. Hence,

$$(z_1, \sigma, z_2) \in \mathcal{C} \Leftrightarrow ((z_1, \sigma, z_2) \in T_0 \wedge \sigma \in A_u).$$

A transition is included in class $\mathcal{R}$ if it must be included in T_0 for reachability reasons, i.e., if a node would no longer be G-connected without the presence of a specific transition, then that transition is included in $\mathcal{R}$. Hence, $(z_1, \sigma, z_2) \in \mathcal{R}$ if and only if there does not exist a $w \in A^*$ such that there is a path labeled by w from z_1 to G in the subgraph (V_0, T_0), where $w(1) \neq \sigma$.

Assume that there is a cycle in (V_0, T_0). If there is a transition (z_1, σ, z_2) on the cycle that is not in $\mathcal{C} \cup \mathcal{R}$, then we can clearly remove this transition and retain the G-connectivity and realizability. This fact follows from the facts that any transition (z_1, σ, z_2) not in $\mathcal{C} \cup \mathcal{R}$ is controllable and there is another path in the subgraph from z_1 to G that does not use the transition in question.

Hence, if we can remove all cycles by deleting transitions that are not in $\mathcal{C} \cup \mathcal{R}$, then we are done.

Assume that there is a cycle remaining that only has transitions in $\mathcal{C} \cup \mathcal{R}$. Let (z_1, σ, z_2) be a transition on this cycle. From the construction of the subgraph (V_0, T_0), we have that there is a string $s \in ppr(L)$ such that $\delta_{\|}(z_0, s) = z_1$. The following lemma provides the crucial result.

LEMMA 4.10. *Let L and $\mathcal{C} \cup \mathcal{R}$ be as above. If there is a cycle that has transitions only in $\mathcal{C} \cup \mathcal{R}$, then there are strings of arbitrary length in L.*

Proof. Let $x \in A^*$ be the labels of the transitions on this cycle, i.e., $\delta_{\|}(z_1, x) = z_1$. Let $m = |x|$. Let z_i denote the state immediately preceding label $x(i)$, i.e., $(z_i, x(i), z_{i+1})$ is a transition on this cycle. Note that $z_{m+1} = z_1$.

From the controllability of L with respect to $L_1 \cap L_2$, Lemma 4.8, and the fact that every transition on the cycle is in $\mathcal{C} \cup \mathcal{R}$, it is clear that $sx(1) \in ppr(L)$. The same result holds for each i, i.e. for all $i : 1 \leq i \leq m : sx(1) \ldots x(i) \in ppr(L)$. Hence, $sx \in ppr(L)$, implying that $sx^* \subseteq ppr(L)$. This last statement contradicts the finiteness of L; consequently, there can be no cycle consisting only of transitions from $\mathcal{C} \cup \mathcal{R}$. □

As a result of Lemma 4.10, any cycle in the subgraph can be removed while retaining G-connectivity and realizability. Hence, the conditions of Proposition 2.2 are satisfied and we get that $z_0 \in \Omega_{M_1 \| M_2}(G)$. □

Based on Proposition 4.7, we define a predicate that is true if two machines can be distinguished as described in the proposition.

DEFINITION 4.11. *Given machines M_1 and M_2, where $M_1 = (Q_1, A, \delta_1, q_{1,0})$ and $M_2 = (Q_2, A, \delta_2, q_{2,0})$, the predicate* $\text{disting}(M_1, M_2)$ *holds if* $(q_{1,0}, q_{2,0}) \in \Omega_{M_1 \| M_2}(G_{1,2})$, *where*

$$G_{1,2} = \left\{ (q_1, q_2) \in Q_1 \times Q_2 \middle| \begin{array}{l} (\exists \sigma \in A : \ \text{different}((q_1, q_2), \sigma)) \\ \wedge \\ (\neg \exists \sigma_u \in A_u : \delta_{\|}((q_1, q_2), \sigma_u)!) \end{array} \right\}.$$

If two machines satisfy this predicate, then the machines, or their languages, are said to be controllably distinguishable.

The following corollary provides the application of Proposition 4.7 to resolving an uncertainty that gives two potentially correct models. If $z_0 \in \Omega_{M_{\|}}(G)$, then a supervisor corresponding to the machine represented by the graph created by the algorithm that generates the region of weak attraction can be built that will constrain the behavior of the unknown system such that an incorrect model can be identified in a finite number of transitions. (See [13] for details on how a machine representation of a supervisor is used to control a plant.)

COROLLARY 4.12. *Let M_1 and M_2 be two models, such that one of them correctly models the plant M. Let $M_{\|} = M_1 \| M_2$, $z_0 = (q_{1,0}, q_{2,0})$, and G be the controllably distinguishing set of states for M_1 and M_2.*

$z_0 \in \Omega_{M_{\|}}(G)$ if and only if the correct model can be chosen in a finite number of transitions.

Proof. $\Longrightarrow$: By Proposition 4.7, if $z_0 \in \Omega_{M_{\|}}(G)$, then there exists a finite nonempty controllable language L such that any string in the language can distinguish $L(M_1)$ and $L(M_2)$. By Theorem 2.1, a supervisor f can be constructed such that $L_f = \overline{L}$; hence, the plant can be controlled so as to execute strings from L. Since one of the models correctly models the plant, $ppr(L) \subseteq L(M)$, and any proper prefix of a string t in L can be executed by the plant. Since $L \subseteq L(M_1) \triangle L(M_2)$, the last symbol in the string will either be executed or not depending on whether $M = M_1$ or $M = M_2$ and which model has t defined in its language.

$\Longleftarrow$: By Proposition 4.7, if $z_0 \notin \Omega_{M_{\|}}(G)$, then we can construct a string of arbitrary length that the plant can execute such that no supervisor may disable events such that an inconsistency is observed. □

The complexity of this approach is governed by the necessity of considering the product machine for M_1 and M_2 to determine the distinguishing language. This operation requires $O(|Q_1||Q_2|)$ operations. In this paper, the dependency of the complexity on the size of the event set A is assumed to be a constant factor; hence it is not included in the expression for the order of complexity. As shown in §5.1, this is a sharp bound on the complexity.

4.2. Distinguishing multiple models. The technique for distinguishing between multiple models with a reset capability available is an extension of the technique used to distinguish between two models. The strategy is to construct a product machine from two models in the set of models that results from considering all possible permutations of the uncertainties. Then, from this product machine, calculate the region of weak attraction for the set of controllably distinguishing states for these two models as described in Corollary 4.12. By using the machine generated by the region of weak attraction as a supervisor for the plant, at least one of these models can be removed from the set of possibly correct models by controlling the plant to enter a state that is a component of one of the product states in the set of controllably distinguishing states. Then, after at least one of these models has been eliminated as a possibly correct model, reset the plant and start the procedure over with another pair of models. Note that it is possible that neither of the models that are chosen is the correct model for the system; hence, the plant might generate a string that is not defined in either of the models used to generate the supervisor. In this case, both models are removed and the procedure continues by choosing another two models. This procedure continues until all uncertainties have been resolved or until no pair of models can be found that satisfy the conditions of Corollary 4.12.

We denote each possible model as $M_i = (Q_i, A, \delta_i, q_{i,0}, Q_{i,m})$, where $i = 1, \ldots, k$, and k is the initial number of models from which the correct model is to be chosen. We denote the initial set of all possible models by S_0. Using the notation given,

$$S_0 = \{M_i | i = 1, \ldots, k\},$$

where each model is in minimal canonical form [5], [9].

The following algorithm specifies the procedure given above. P denotes the actual machine or the plant which is to be correctly modeled.

ALGORITHM 4.1.

```
Input:
    $S_0$ as given above.
Output:
    $S_n = \{M_i|$ no additional uncertainties can be resolved$\}$.
Algorithm:
    $p = 0$.
    While $(|S_p| > 1) \wedge (\exists M_i \in S_p \wedge \exists M_j \in S_p : disting(M_i, M_j))$:
        Calculate $\Omega_{M_i \| M_j}(G_{i,j})$.
        Use $M_\Omega(G_{i,j})$ as supervisor for $P$.
        Determine which model is still a possibly correct model and which
        model is not consistent with the plant.
        $S_{p+1} = S_p \backslash \{$ models which have been determined to be inconsistent
        with plant $\}$.
        Reset the plant.
        $p = p + 1$.
    End while.
End of algorithm.
```

In the worst case, the product for every pair of models would need to be calculated to check for pairs that satisfy $z_0 \in \Omega_{M_\|}(G_{i,j})$. Since there are k models in S, this calculation of products results in an algorithm with $O(k^2|Q|^2)$ complexity.

A slight modification of the proposed algorithm is to simulate, on all models in S_p, the strings that result from using $M_\Omega(G_{i,j})$ as a supervisor for the plant. Using this technique, any model that cannot successfully simulate the activity of the plant can be eliminated from consideration and need not be considered in any future pairing.

This modified approach also has worst case complexity of $O(k^2|Q|^2)$ since there is no guarantee that more than one model will be eliminated on each iteration. Also the actual complexity to accomplish the simulation results in an additional $O(k|Q|^2)$ term in the operation count. These counts are a result of the following reasoning. Each calculation of $\Omega_{M_\|}(G)$ adds a $|Q^2|$ term to the count. There are at most k^2 pairs that must be calculated, hence, the $O(k^2|Q^2|)$ term in the count. To simulate any test string on all remaining potential models, $O(k|Q^2|)$ operations are necessary; hence, this term is added to the count retaining an overall complexity of $O(k^2|Q^2|)$.

To demonstrate the correctness of the algorithm, several points must be addressed: insuring that only bad models are removed from S_i, that the order of choosing models for the test $disting(M_i, M_j)$ does not affect the output, and that there is no other technique which might produce a smaller set of potential models.

4.2.1. Only bad models removed. The first point is easily addressed. A model is removed if it is inconsistent with the plant. An inconsistency arises from either the plant executing a transition that is not in the model, such as an uncontrolled transition, or the plant not executing a transition that is defined in the model, such as from a state at which a single controlled or uncontrolled transition is defined in the model but is not executed by the plant. Hence, only "bad" models are removed from the set used to keep potentially correct models. Note that a consistent model will not be removed from this set.

4.2.2. Order does not affect the result. The second point is more subtle. A priori it appears that the order of testing models might be significant. That is, there might be some incorrect model A that, when combined with another incorrect model B, generates a test string that provides that model B is removed, but that when model B is combined with the correct model a test string cannot be generated.

That the order does not matter follows from the following proposition. Proposition 4.13 states that if a pair of models, M_i and M_j, satisfy $z_0 \in \Omega_{M_i \| M_j}(G_{i,j})$ and if M_i is removed from the set of possibly correct models, then the correct model M_c and M_i satisfy $z_0 \in \Omega_{M_i \| M_c}(G_{i,c})$. Hence, if model M_j can be used to remove M_i, then model M_c can be used instead.

In the statements of the following propositions, the language generated by model M_i, which is usually denoted by $L(M_i)$, is denoted by L_i. In the proof of Proposition 4.13, a language L is used to link the fact that the initial state is in the region of attraction of each of the product machines. The conditions on L are very similar to the conditions for a distinguishing language for M_i and M_j; however, the fact that neither M_i nor M_j might be the correct model requires that slightly different characteristics describe how M_j can be used with M_i to generate a supervisor that will cause M_i to be removed from the set of possibly correct models. In the following, M_c denotes the correct model. $z_{0,i,j}$ is the initial state of $M_i \| M_j$. $z_{0,i,c}$ is the initial state of $M_i \| M_c$. $G_{i,j}$ is the set of controllably distinguishing states for $M_i \| M_j$. $G_{i,c}$ is the set of controllably distinguishing states for $M_i \| M_c$.

PROPOSITION 4.13. *If $z_{0,i,j} \in \Omega_{M_i \| M_j}(G_{i,j})$ and M_i is removed from the set of possible machines, then $z_{0,i,c} \in \Omega_{M_i \| M_c}(G_{i,c})$.*

Proof. Consider the language L marked by the supervisor generated by $\Omega_{M_i \| M_j}(G_{i,j})$ and used to remove M_i. By the assumptions in the proposition statement, this language generates tests that are used to remove M_i from the set of potentially correct models.

We first describe some characteristics that this language satisfies.

LEMMA 4.14.

(a) *L controllable with respect to $L_i \cap L_j \cap L_c$,*

(b) *$L \subseteq (L_i \triangle L_c) \cap (L_j \cup L_c)$,*

(c) *$ppr(L) \subseteq L_i \cap L_j \cap L_c$, and*

(d) *L is finite and nonempty.*

Proof. The controllability of L with respect to $L_i \cap L_j$ follows from the construction of $\Omega_{M_i \| M_j}(G_{i,j})$ and the fact that $\Omega_M(G)$ is realizable. That L is controllable with respect to $L_i \cap L_j \cap L_c$, follows from the fact that all prefixes of the closed-loop behavior are necessarily constrained to L_c.

The strings that occur in the distinguishing language generated by $\Omega_{M_i \| M_j}(G_{i,j})$ consist of the following types:

(1) strings that occur in the plant, M_i, and M_j, i.e., $L_i \cap L_j \cap L_c$,

(2) strings that occur in the plant and M_j but not in M_i, i.e., $L_c \cap L_j \cap L_i^c$,

(3) strings that occur in M_i and M_j but not in the plant, i.e., $L_c^c \cap L_i \cap L_j$, and

(4) strings that occur in the plant but not in M_i and M_j, i.e., $L_c \cap (L_i \cup L_j)^c$.

For this supervisor to cause M_i to be removed from the set of possibly correct machines, the strings that occur in all three must be in the prefix of strings that will cause M_i to be removed. Hence, $ppr(L) \subseteq L_i \cap L_j \cap L_c$.

For this supervisor to cause M_i to be removed from the set of possibly correct machines in a controllable fashion, we claim that only the strings in (2), (3), and (4) above can occur as strings in the language. (See shaded areas in Fig. 3.) No other string can occur and still allow M_i to be removed from the set of possibly correct machines. Any other string would not allow M_i to be removed.

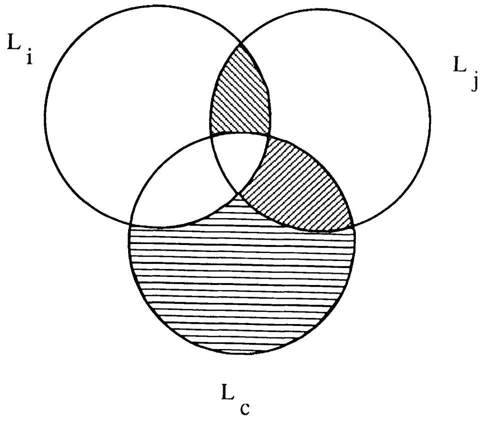


FIG. 3. *Parts of languages that allow removal of M_i.*

From this observation we have that

$$\begin{aligned} & L \subseteq (L_c \cap L_j \cap L_i^c) \cup (L_c^c \cap L_j \cap L_i) \cup (L_c \cap (L_j \cup L_i)^c) \\ \Rightarrow \quad & L \subseteq ((L_j - L_i) \cap L_c) \cup ((L_j \cap L_i) - L_c) \cup (L_c - (L_j \cup L_i)), \end{aligned}$$

which gives Lemma 4.14(b).

The finiteness of L is a result of the fact that $\Omega_{M_i \| M_j}(G_{i,j})$ is acyclic. That L is nonempty is a result of the fact that M_i is removed, i.e. at least one event must be used to determine that M_i is not correct. □

We now use this language to demonstrate that $z_{0,i,c} \in \Omega_{M_i \| M_c}(G_{i,c})$. We demonstrate this fact by verifying that L satisfies the requirements of Proposition 4.7 for M_c and M_i.

1. L finite and nonempty:

L is finite and non-empty by hypothesis.

2. L controllable with respect to $L_c \cap L_i$: From the definition of $\overline{L}$, we have that $(t \in \overline{L}) \Rightarrow t \in L \cup ppr(L)$.

$$\begin{array}{ll} & (t\sigma_u \in L_c \cap L_i) \\ \Rightarrow & \\ & t \in (L_i \cap L_c) \\ \Rightarrow & \qquad \{(b) \text{ and } (c)\} \\ & t \in ppr(L) \\ \Rightarrow & \qquad \{(a) \text{ and } (c)\} \\ & t\sigma_u \in \overline{L} \\ \Rightarrow & \\ & L \text{ controllable w.r.t. } L_i \cap L_c. \end{array}$$

3. $L \subseteq L_c \triangle L_i$:

$$
\begin{aligned}
& L \subseteq (L_i \triangle L_c) \cap (L_j \cup L_c) \\
\Leftrightarrow \\
& L \subseteq ((L_j - L_i) \cap L_c) \cup (L_c - (L_j \cup L_i)) \cup ((L_j \cap L_i) - L_c) \\
\Rightarrow \\
& L \subseteq (L_i^c \cap L_c) \cup (L_c \cap L_i^c) \cup (L_i \cap L_c^c) \\
\Rightarrow \\
& L \subseteq (L_c \cap L_i^c) \cup (L_i \cap L_c^c) \\
\Rightarrow \\
& L \subseteq L_c \triangle L_i.
\end{aligned}
$$

4. $ppr(L) \subseteq L_c \cap L_i$:

$$
\begin{aligned}
& ppr(L) \subseteq L_i \cap L_j \cap L_c \\
\Rightarrow \\
& ppr(L) \subseteq L_c \cap L_i
\end{aligned}
$$

Since L satisfies the requirements for Proposition 4.7 with respect to M_i and M_c, we immediately have that $z_{0,i,c} \in \Omega_{M_i \| M_c}(G_{i,c})$. □

Proposition 4.13 provides that the order does not matter when choosing which pair of models to use to generate the next test. When combined with the first point, that only "bad" models are removed, we have that it is sufficient to test bad models with the correct model, which will never be removed from the set of potentially correct models.

4.2.3. Optimal complexity of Algorithm 4.1. Now we address the question of whether some other procedure might be used to generate a smaller set of potentially correct models.

PROPOSITION 4.15. *Algorithm 4.1 provides a minimal set of potentially correct models. (Minimal in the sense that there does not exist another technique to controllably remove more models than are removed by Algorithm* 4.1 *from the set of potentially correct models in a finite number of transitions.)*

Proof. Let $\overline{S}$ be the set of potentially correct models output by Algorithm 4.1.

If $|\overline{S}| = 1$, then we are done, i.e., $\overline{S} = \{M_c\}$.

If $|\overline{S}| > 1$, then we know that since the test disting(M_i, M_j) is not satisfied for any pair of models in $\overline{S}$ and consequently

$$\forall i, j : M_i, M_j \in \overline{S} : z_{o,i,j} \notin \Omega_{M_i \| M_j}(G_{i,j}).$$

In particular, we know that

$$z_{o,i,c} \notin \Omega_{M_i \| M_c}(G_{i,c}). \tag{1}$$

Assume that there is some other technique that controllably removes $M_i \neq M_c$ from $\overline{S}$. To remove M_i, a test must cause M_i to reach a state at that an inconsistency with the plant arises. This state in M_i is a component to a state in $G_{i,c}$, otherwise by controllability constraints M_i cannot be removed by this test. By (1), we know that we can construct a behavior for M_i that never enters G. Hence, we have a contradiction and there cannot be any technique that can controllably generate a smaller set of potentially correct models. □

4.2.4. Resolving uncertainty without reset. To resolve uncertainty without a reset capability, a slight modification must be made to the algorithms given previously. The modification consists of updating the models still under consideration to reflect any actions that the actual plant has taken. This update is manifested by modifying the model descriptions so that the initial state has a dependence on events that have already occurred.

Hence, the old model $M = \{Q, A, \delta, q_0, Q_m\}$ is modified to incorporate the string s, which the plant has executed to this point and denoted by $M(s) = \{Q, A, \delta, q_{\hat{0}}(s), Q_{\hat{m}}\}$. Note that only the initial state needs to have this dependence. The other components of the model do not need to be modified.

Note that for this modification, all models must be updated to determine if the new initial state after a test string has been executed is in the region of attraction for the set of distinguishing states. However, the actual region of attraction does not need to be recalculated because the states that can be attracted to the distinguishing states do not change with each test string; only the initial state changes.

The need to simulate the test strings does not increase the complexity of the algorithm. In the worst case, this algorithm could require that $O(k^2)$ regions of weak attraction be calculated to find enough test strings. Hence, this algorithm also has $O(k^2|Q|^2)$ complexity.

5. Examples.

5.1. Optimality for single transition uncertainty. This example demonstrates that resolving a single uncertainty has complexity at least as great as that of creating the product machine for the two potentially correct models. This complexity arises from the fact that the product machine is used to generate the set of controllably distinguishing states and hence the minimally distinguishing language. For this example, z is the event for the uncertain arcs and $A_u = \{a, c, d\}$. Figure 4 illustrates the two possible transition functions for the machine.

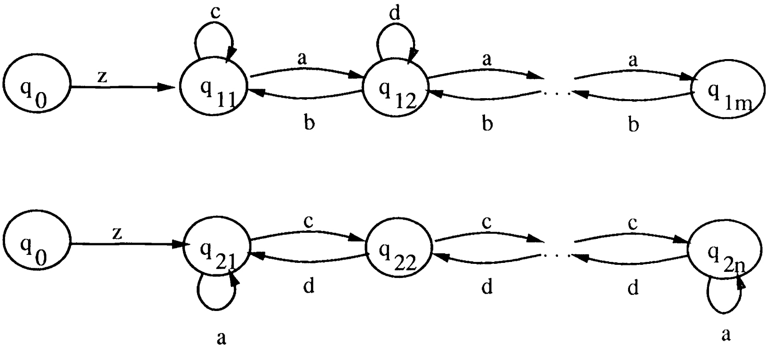


FIG. 4. *Set of models for which product method is optimal.*

Following the procedure specified in Proposition 4.7, we create the product machine (Fig. 5) and calculate the states G that can be used to distinguish q_{11} and q_{21} and the region of weak attraction for G.

From the graph representation of the transition function for the product machine, we can determine that

$$G = \{q_{1m,2n}, q_{0,11}, \dots, q_{0,1m}, q_{0,21}, \dots, q_{0,2n}, \}.$$

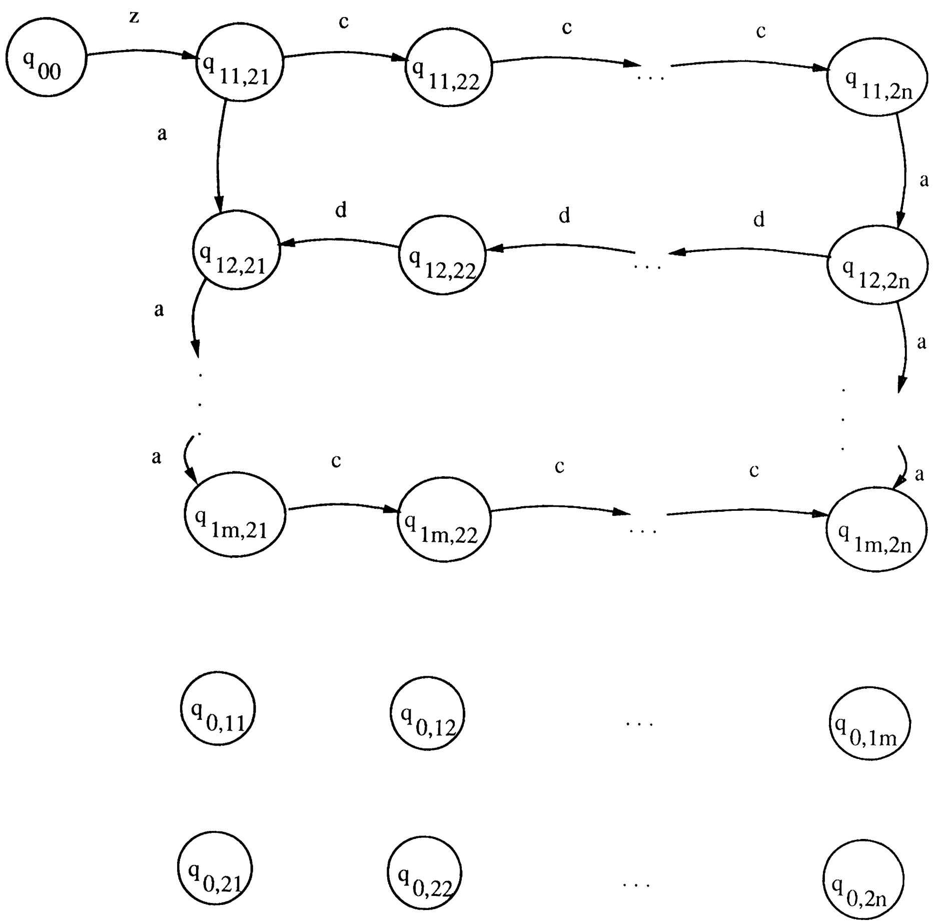


FIG. 5. *Product machine for system for which product method is optimal.*

From Fig. 5, we observe that the only state in the current G that can have $q_{0,0}$ in the region of weak attraction is $q_{1m,2n}$; hence, we will limit our calculations for a new set $G' = \{q_{1m,2n}\}$. Some of the iterations of the algorithm to calculate the region of weak attraction are given:

$$\begin{aligned} V_0 &= \{q_{1m,2n}\} \\ V_1 &= V_0 \cup \{q_{1m,2(n-1)}\} \\ &\;\;\vdots \\ V_n &= V_{n-1} \cup \{q_{1(m-1),21}\} \\ &\;\;\vdots \\ V_{m*n} &= V_{m*n-1} \cup \{q_{0,0}\}. \end{aligned}$$

Hence, by Proposition 4.7, since $z_0 \in \Omega_{M_{\|}}(G')$, the two states q_{11}, q_{21} are controllably distinguishable, and the uncertain arc can be resolved. Observe that $q_{1m,2n} \in G'$ and that there is a string $z(c^n ad^n a)^m$ that can occur uncontrollably before reaching

$q_{1m,2n}$; hence, to resolve the uncertainty, every state that can be reached in the product machine from the initial state might be visited. This fact demonstrates that resolving this uncertain transition requires $O(mn)$ operations.

To resolve the uncertainty, construct a supervisor with finite state machine representation as shown in Fig. 5 and run the unknown plant and supervisor as a closed-loop system. (See [13] for more detail on this procedure.) A distinguishing language for this example is $L = z(c^n ad^n a)^m a$.

5.2. Finiteness of languages in Proposition 4.7. This example demonstrates the requirement for finiteness in Proposition 4.7.

Let $L_1 = (av)^*$ and $L_2 = (a(u+v))^*$ be the languages for two possible models where $a \in A$ and $u, v \in A_u$. See Fig. 6 for the machine representation for these languages.

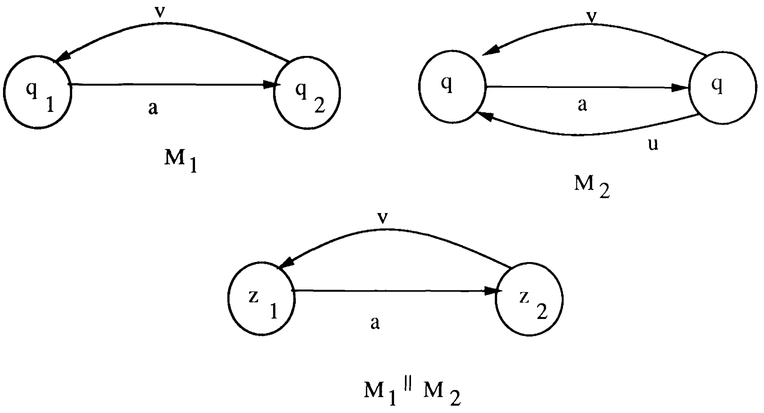


FIG. 6. *System to demonstrate requirement for finiteness.*

For this example, if $L = (av)^*au$, then L is controllable with respect to $L_1 \cap L_2$, $L \subseteq L_1 \triangle L_2$, and $ppr(L) \subseteq L_1 \cap L_2$ as required for a language that can be used to distinguish L_1 and L_2 as described in Proposition 4.7; however, the initial state of the product machine is not in the region of weak attraction of the set of controllable distinguishing states, which is empty in this example.

6. Conclusions. In this paper we have presented a model of uncertainty related to the transitions of systems modeled with finite state machines. We developed a test for determining whether or not such uncertainty can be controllably resolved. The test using a region of weak attraction calculation also provides an algorithm for constructing a supervisor which can resolve the uncertainties. An example demonstrating the optimality of the deterministic approach for a single uncertainty is provided. Also, an example is given which demonstrates how the controllability and finiteness requirements are both necessary for Proposition 4.7. This approach to choosing the correct model can be applied in any situation that has a set of models from which the correct one should be chosen.

Several possibilities exist for extensions to this work. One possibility is to expand the model used to describe a discrete event system to one that can describe a broader category of systems, such as a Petri net [8] or algebraic [10] models. Another direction of current interest is the influence that different uncertainty models have on the control and stabilization of systems modeled with discrete event system formalisms. This influence incorporates the effect that limiting the behavior of a system to a desired constraint language would have on correctly controlling the system and resolving any uncertainty in the model. A further extension is to consider how the addition of unobserved events affects the problem described in this work.

Acknowledgments. The authors thank Ratnesh Kumar, Steve Marcus, Alex Tomlinson, and the reviewers for critically reading and assisting in the preparation of this paper.

REFERENCES

[1] A. Aho, J. Hopcroft, and J. Ullman, *The Design and Analysis of Computer Algorithms*, Addison-Wesley, Reading, MA, 1974.

[2] D. Angluin and C. H. Smith, *Inductive Inference: Theory and Methods*, Comput. Surveys, 15 (1983), pp. 237–269.

[3] Y. Brave and M. Heymann, *On Stabilization of Discrete Event Processes*, Internat. J. Control, 51 (1990), pp. 1101–1117.

[4] N. Deo, *Graph Theory with Applications to Engineering and Computer Sciences*, Prentice–Hall, Englewood Cliffs, NJ, 1974.

[5] S. Eilenberg, *Automata, Languages, and Machines Volume A*, Academic Press, New York, NY, 1974.

[6] E. Gold, *System Identification via State Characterization*, Automatica, 8 (1972), pp. 621–636.

[7] Y. Ho, *Scanning the Issue*, Proceedings of IEEE, 77 (1989), pp. 3–6.

[8] L. Holloway and B. Krogh, *Synthesis of feedback control logic for a class of controlled Petri nets*, IEEE Trans. Automat. Control, 35 (1990), pp. 514–523.

[9] J. Hopcroft and J. Ullman, *Introduction to Automata Theory, Languages and Computation*, Addison-Wesley, Reading, MA, 1979.

[10] K. Inan and P. Varaiya, *Algebras of Discrete Event Models*, Proceedings of IEEE, 77 (1989), pp. 24–38.

[11] R. Kumar, V. Garg, and S. Marcus, *Language stability and stabilizability of discrete event systems*, SIAM J. Control Optim., 31 (1993), pp. 1294–1320.

[12] P. Ramadge and W. Wonham, *Supervisory control of a class of discrete event processes*, SIAM J. Control Optim., 25 (1987), pp. 206–230.

[13] ———, *The control of discrete event systems*, Proceedings of IEEE, 77 (1989), pp. 81–98.

[14] R. Rivest and R. Schapire, *Diversity-Based Inference of Finite Automata*, in Proceedings 28th Annual Symposium on Foundations of Computer Science, Los Angeles, CA, 1987, pp. 78–87.

[15] ———, *Inference of Finite Automata Using Homing Sequences*, in Proceedings 21st Symposium on Theory of Computing, Seattle, WA, 1989, pp. 411–420.

[16] L. Valiant, *A Theory of the Learnable*, Comm. ACM, 27 (1984), pp. 1134–1142.

[17] J. Willems, *Paradigms and Puzzles in the Theory of Dynamical Systems*, IEEE Trans. Automat. Control, 36 (1991), pp. 259–294.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 33, No. 1, pp. 227–254, January 1995

© 1995 Society for Industrial and Applied Mathematics

011

ANALYSIS OF COSTATE DISCRETIZATIONS IN PARAMETER ESTIMATION FOR LINEAR EVOLUTION EQUATIONS*

C. R. VOGEL[†] AND J. G. WADE[‡]

Abstract. A widely used approach to parameter identification is the output least-squares formulation. Numerical methods for solving the resulting minimization problem almost invariably require the computation of the gradient of the output least-squares functional. When the identification problem involves time-dependent distributed parameter systems (or approximations thereof), numerical evaluation of the gradient can be extremely time consuming. The costate method can greatly reduce the cost of computing these gradients. However, questions have been raised concerning the accuracy and convergence of costate approximations, even when the numerical methods being used are known to converge rapidly on the forward problem.

In this paper it is shown that the use of time-marching schemes that yield high-order accuracy on the forward problem does not necessarily lead to high-order accurate costate approximations. In fact, in some cases these approximations do not converge at all. However, under certain circumstances, rapidly converging gradient approximations do result because of rapid weak-star-type convergence of the costate approximations. These issues are treated both theoretically and numerically.

Key words. parameter estimation, evolution equations, costate method

AMS subject classifications. 35R30, 49D07

1. Introduction.

In this paper we analyze temporal discretizations of the costate method for computing gradients in the output least-squares approach to parameter estimation. This analysis applies to initial value problems of the form

$$\begin{aligned} \dot{u}(t) &= A(q)\,u(t) + f(t), \qquad 0 < t < t_F, \\ u(0) &= 0. \end{aligned} \tag{1.1}$$

Here $A(q)$ is a bounded linear operator on a Hilbert space H, and the inner product on H is denoted $\langle \cdot, \cdot \rangle_H$. The dot over u indicates differentiation with respect to t. We will refer to u as the state variable and to (1.1) as the state equation. We assume q lies in set $\mathcal{Q}_{AD}$ of admissible parameters contained in a normed linear "parameter" space $\mathcal{Q}$. Throughout the paper we assume that the map $q \mapsto A(q)$ is Gateaux differentiable in the operator norm.

In applications of interest, (1.1) is a finite-dimensional (i.e., $\dim(H) < \infty$) approximation of a time-dependent partial differential equation (PDE). An important example is the diffusion equation

$$\begin{aligned} \frac{\partial u}{\partial t}(t,x) &= \nabla \cdot (q(x)\nabla u(t,x)) + f(t,x),\ x \in \Omega,\ 0 < t < t_F, \\ u(t,x) &= 0, x \in \partial\Omega,\ 0 < t < t_F, \\ u(t,x) &= 0, \qquad\qquad x \in \Omega. \end{aligned} \tag{1.2}$$

* Received by the editors June 15, 1992; accepted for publication (in revised form) August 24, 1993.

[†] Department of Mathematical Sciences, Montana State University, Bozeman Montana 59717. The research of this author was supported in part by National Science Foundation grant DMS-9106609 and by the Center for Interfacial Microbial Process Engineering at Montana State University, a National Science Foundation sponsored Engineering Research Center, and by the Center's Industrial Associates.

[‡] Center for Applied Mathematical Sciences, University of Southern California, Los Angeles, California 90089-1113. Present Address, Department of Mathematics, Texas A & M University, College Station, Texas 77843–3368. The research of this author was supported in part by Air Force Office of Scientific Research grants AFOSR-90-0091 and AFOSR-86-0126, and by National Science Foundation grants DMS-8818530 and DMS-8704169.

In these situations, $\dim(H)$ can be arbitrarily large.

In general we assume that the solution u lies in the "state space"

$$\mathcal{H} = L^2(0, t_F; H),$$

which is a Hilbert space with inner product

$$\langle\langle f, g \rangle\rangle_{\mathcal{H}} = \int_0^{t_F} \langle f(t), g(t) \rangle_H \, dt. \tag{1.3}$$

As in Banks and Kunish [3], we assume the existence of an "observation space" $\mathcal{Z}$, which is a Hilbert space with inner product $\langle\langle \cdot, \cdot \rangle\rangle_{\mathcal{Z}}$, and an "observation operator"

$$\mathcal{C} : \mathcal{H} \to \mathcal{Z}.$$

Given an observation z of u, we wish to estimate the parameter q. In the output least-squares approach to parameter estimation, $q \in \mathcal{Q}_{AD}$ is selected to minimize the functional

$$T_\alpha(q) \overset{\text{def}}{=} \frac{1}{2} ||\mathcal{C}u(q) - z||_{\mathcal{Z}}^2 + \frac{\alpha}{2} \mathcal{N}^2(q), \tag{1.4}$$

where $u(q)$ is the solution to (1.1). Here the scalar α is a positive regularization parameter and $\mathcal{N}^2(q)$ is a regularization functional whose purpose is to stabilize the minimization. Usually, $\mathcal{N}(q)$ is a norm or seminorm on $\mathcal{Q}$.

Computational methods to minimize $T_\alpha(q)$ typically require gradients or gradient approximations. For example, assuming a discretization of the parameter of the form

$$q = \sum_{i=1}^{n_q} \hat{q}_i \psi_i,$$

we often (see, for example, [3, §V.6]) approximate the components of the gradient using finite differences, e.g.,

$$[\nabla T_\alpha(q)]_i \approx \frac{T_\alpha(q + \tau\psi_i) - T_\alpha(q)}{\tau}, \tag{1.5}$$

for τ a small scalar. Note that in the limit as $\tau \to 0$, the ith component of $\nabla T_\alpha(q)$ is the directional (i.e., Gateaux) derivative of $T_\alpha(q)$ in the direction ψ_i. When $\dim(H)$ and n_q are large, gradient approximations based directly on (1.5) are extremely expensive, requiring $n_q + 1$ evaluations of the functional T_α, and hence, $n_q + 1$ solutions of the evolution equation (1.1).

An attractive alternative is the costate approach [3, §V.5], [7], which is described in its continuous form in §2. In sharp contrast to the $(n_q + 1)$ state equation solutions needed for the finite difference method (1.5), the costate method requires the solution of only two evolution equations, the state equation and the "costate" or "adjoint" equation, followed by n_q $\mathcal{H}$-inner products.

Unfortunately, computational experience with adjoint approximations in parameter estimation for certain evolution equations has been puzzlingly disappointing. For example, the authors of [4] concluded that this approach was unsatisfactory because "... it was extremely difficult to obtain accurate search directions with gradients computed in this manner." Also, in [2] the authors suggest the costate method for a certain class of damped elastic systems and explain how it may facilitate the efficient use of gradient-based methods such as conjugate gradients or BFGS. They also report having used it in some of their numerical experiments. However, in *none* of the specific examples upon which they report did they use this method. Subsequent private communication with two of the authors revealed that they encountered unexpected difficulty with the costate approach. Later commenting on this, in [1] Banks states that "...we experienced so much difficulty with the costate based gradient methods that we abandoned them ...."

Similar difficulties have been encountered in optimal control settings. In [5], which dealt with computational methods for control systems governed by delay differential equations, the authors observed that certain spline approximations to the solution of operator Ricatti equations failed to converge strongly. In a subsequent work [6] it was carefully shown that the underlying difficulty was lack of strong convergence of the adjoint approximations.

In this paper we focus on the temporal discretizations of the costate system in least-squares parameter estimation. Since the costate method of computing $\nabla T_\alpha(q)$ requires the solution of two evolution equations plus n_q inner products, two important components of any numerical scheme for approximating the gradient are (a) the scheme for approximating the solution of the evolution equations (which we assume is a time-marching scheme), and (b) the numerical quadrature scheme. If both of these are, say, νth-order accurate, then one would reasonably hope to attain νth-order convergence of the gradient approximations. However, if the observation operator $\mathcal{C}$ involves pointwise evaluation in time, then certain subtleties arise and we may observe unexpectedly poor convergence of the gradient approximations. Perhaps the most striking examples we present involve the fourth-order Runge–Kutta (RK4) time-marching scheme. We show that if RK4 is used in conjunction with Simpon's rule (which is fourth order) for numerical quadrature, then the gradient approximations fail altogether to converge. We show that in fact these RK4/Simpson's approximations converge with *second-order* accuracy to the (3/2) times the true gradient! More positively but perhaps equally surprisingly, we prove that RK4 together with the *second-order accurate* "trapezoidal" quadrature rule yields *fourth-order* convergence!

The underlying reason for these strange phenomena is that when $\mathcal{C}$ involves pointwise evaluation in time, the costate equation is an evolution equation with Dirac-delta functions in the forcing term. For this reason, the costate approximate solutions do not converge strongly. However, in some cases they do exhibit high-order convergence in the weak* topology of the dual space of C^ν.

These considerations are similar in spirit to the work reported in [6]. In the introduction of that paper, the authors state "We feel that many distributed parameter control systems are such that 'standard' ...schemes might lead to numerical difficulties" when used in optimization schemes, and "we hope that the reader will be motivated to think about similar problems for more complex distributed parameter systems." While that paper addresses the optimal control problem, it is nonetheless relevant to us to the extent that output least-squares parameter estimation is mathematically similar to optimal control.

Banks [1] has suggested that nonconvergence of costate approximations can also occur in parameter estimation because of a lack of convergence of the *spatial* discretization. For example, in systems such as (1.2), we must discretize with respect to each component of x as well as with respect to t. Then a result needed for the costate approximations is essentially weak* convergence to the dual semigroups. These considerations suggest directions for future work, but we do not pursue them here.

In §2 we discuss the costate method for the continuous problem, apart from any temporal approximations. In §3 we introduce a fairly wide class of time-marching schemes for (1.1) and a corresponding costate approximation scheme. We then analyze these approximations in some detail, making certain assumptions about the properties of the time-marching scheme, the observation operator, and the various operator discretizations. Section 4 contains some numerical examples that illustrate the results of the analysis. In §5, several alternative approaches that avoid the difficulties mentioned above are presented. Our conclusions are discussed in §6.

2. The continuous costate approach. As stated in the introduction, typical implementations of the output least-squares approach to parameter estimation require gradients of the functional $T_\alpha(q)$ given in (1.4). Most of the effort in obtaining this gradient lies in computing the gradient of the least-squares functional

$$T(q) \overset{\text{def}}{=} \tfrac{1}{2}||\mathcal{C}u - z||^2_{\mathcal{Z}}, \tag{2.1}$$

where $u = u(t;q)$ solves the state equation

$$\begin{aligned} \dot{u}(t) &= A(q)u(t) + f(t), \qquad 0 < t < t_F, \\ u(0) &= 0. \end{aligned} \tag{2.2}$$

Since the gradient of $T(q)$ can be obtained by computing directional derivatives, we focus throughout on the computation of the directional derivative of T at q in the direction p. This is given by

$$\delta_p T(q) \overset{\text{def}}{=} \lim_{\tau \to 0} \frac{T(q + \tau p) - T(q)}{\tau}. \tag{2.3}$$

The residual is defined by

$$r = \mathcal{C}u - z, \tag{2.4}$$

and hence,

$$\delta_p T(q) = \langle\langle\, r,\, \mathcal{C}\delta_p u \,\rangle\rangle_{\mathcal{Z}}. \tag{2.5}$$

The solution $u = u(q)$ to (2.2) is given in terms of the "solution operator" $\mathcal{S}(q)$ by

$$u(t) = [\mathcal{S}(q)f](t) \overset{\text{def}}{=} \int_0^t e^{A(q)s} f(t-s)\, ds. \tag{2.6}$$

In terms of $\mathcal{S}(q)$, we have

$$\delta_p u = \mathcal{S}(q)\left(\delta_p A(q)\right) u. \tag{2.7}$$

Thus

$$\delta_p T(q) = \langle\langle\, r,\, \mathcal{C}\mathcal{S}(q)\, \delta_p A(q)\ u \,\rangle\rangle_{\mathcal{Z}}.$$

Taking adjoints of $\mathcal{C}$ and $\mathcal{S}(q)$, we obtain

$$\delta_p T(q) = \langle\langle\, y,\, \delta_p A(q)\ u \,\rangle\rangle_{\mathcal{H}},$$

where y is the solution to the "costate equation"

$$y = \mathcal{S}^*(q)\mathcal{C}^* r. \tag{2.8}$$

The operator $\mathcal{S}^*(q)$ is given, for $g \in \mathcal{H}$, by

$$[\mathcal{S}^*(q)g](t) = \int_t^{t_F} e^{-A^*(q)s} g(t-s)\, ds.$$

The differential equation that y satisfies is

$$\begin{aligned} -\dot{y}(t) &= A^*(q)y(t) + (\mathcal{C}^* r)(t), \qquad 0 < t < t_F, \\ y(t_F) &= 0. \end{aligned}$$

Since this equation has a *final* condition instead of an initial condition, it is useful to note that it is equivalent to

$$\begin{aligned} \dot{\hat{y}} &= A^*(q)\hat{y} + J(\mathcal{C}^* r), \\ \hat{y}(0) &= 0, \\ y &= J\hat{y}. \end{aligned} \tag{2.9}$$

Here J is the "time reversal" operator on $\mathcal{H}$ defined by

$$(Jf)(t) = f(t_F - t). \tag{2.10}$$

The costate approach to computing $\nabla T(q)$ then consists of the following four steps.

The continuous costate method.

(i) Solve the state equation (2.2) for the state variable $u = u(q)$.

(ii) Compute the residual $r = \mathcal{C}u(q) - z$.

(iii) Solve the costate equation (2.9) for the costate variable y.

(iv) Compute the directional derivatives of T in directions p (cf. (1.5), where $p = \phi_i$) by

$$\delta_p T(q) = \langle\!\langle\, y,\, \delta_p A(q)\; u \,\rangle\!\rangle_{\mathcal{H}}, \tag{2.11}$$

Note that this approach requires the solution of only two evolution equations, the state equation (2.2) and costate equation (2.9), as opposed to the $(n_q + 1)$ evolution equations of (1.5). Moreover, if A is linear in q (as is the case with the approximations of (1.2), for example), then $\delta_p A(q) = A(p)$, so that the implementation of step (iv) is straightforward. Finally, we note that to complete the computation of $\nabla T_\alpha(q)$ it is necessary to compute the gradient $\nabla \mathcal{N}^2(q)$ of the regularization functional, but this is usually trivial.

3. Analysis of discrete approximations.

3.1. The discrete approximations.

Usually in solving (2.2) numerically, we use some sort of finite-difference or time-marching scheme (TMS). In this section, we pose and analyze a rather natural algorithm for computing costate approximations based on a given TMS and the continuous costate method described above.

For a given n, let

$$\mathbf{t}^n = \{t_k^n\}_{k=0}^n, \qquad 0 = t_0^n < t_1^n < \ldots < t_n^n = t_F$$

denote a specified mesh on $[0, t_F]$, and define $h_k \stackrel{\text{def}}{=} t^n_{k+1} - t^n_k$. For later use in the discussions of asymptotic rates of convergence, we define

$$|\mathbf{h}| \stackrel{\text{def}}{=} \max_{0 \leq k \leq n-1} h_k.$$

Also, let $u^n = \{u^n_k\}^n_{k=0}$, where u^n_k is the approximation to the state variable u at t^n_k, which is obtained using a particular TMS with this mesh. If we denote $\{f(t^n_k)\}$ by f^n, then we may express this approximation by

$$u^n(q) \leftarrow \mathrm{TMS}(\mathbf{t}^n, A(q), f^n).$$

This approximation is then used in a minimization scheme for $T_\alpha(q)$ in the parameter estimation problem. Corresponding to the continuous least-squares functional in (2.1), define

$$T^n(q) = \tfrac{1}{2} \| \mathcal{C}^n u^n - z^n \|^2_{\mathcal{Z}^n}. \tag{3.1}$$

Here z^n, $\mathcal{Z}^n$, and $\mathcal{C}^n$ are discretizations of z, $\mathcal{Z}$, and $\mathcal{C}$, which are discussed in detail below. If the minimization scheme requires directional derivatives (i.e., gradients), then a seemingly natural approach ([3, §V.5]), which we call the "discretized costate approximation," based on an obvious discretization of the continuous costate method, suggests itself.

The discretized costate approximation.

(i) Compute $u^n(q)$ by

$$u^n(q) \leftarrow \mathrm{TMS}(\mathbf{t}^n, A(q), f^n).$$

(ii) Compute the residual $r^n = \mathcal{C}^n u^n(q) - z^n$.

(iii) Compute y^n by

$$\begin{aligned} \tilde{\mathbf{t}}^n &\leftarrow J^n(t_F - \mathbf{t}^n), \\ \tilde{y}^n &\leftarrow \mathrm{TMS}(\tilde{\mathbf{t}}^n, A^*(q), J^n(\mathcal{C}^*)^n r^n), \\ y^n &\leftarrow J^n \tilde{y}^n. \end{aligned} \tag{3.2}$$

Here, J^n is the approximation of J (cf. (2.10)) given by

$$[J^n f^n]_k = f^n_{n-k} \tag{3.3}$$

and $(\mathcal{C}^*)^n$ is an approximation to the adjoint of $\mathcal{C} : \mathcal{H} \rightarrow \mathcal{Z}$.

(iv) Approximate directional derivatives in directions p by

$$\delta_p T(q) \approx \widetilde{\delta_p T^n} \stackrel{\text{def}}{=} \langle\!\langle\, y^n, \, (\delta_p A(q))\, u^n(q) \,\rangle\!\rangle_{\mathcal{H}^n}. \tag{3.4}$$

Here, $\langle\!\langle \cdot, \cdot \rangle\!\rangle_{\mathcal{H}^n}$ is a discretization of the continuous inner product (1.3).

The convergence of $\widetilde{\delta_p T^n}$ to $\delta_p T^n$ depends on factors such as the convergence of the time-marching scheme TMS on the forward problem, the convergence of approximations $\mathcal{C}^n$ to the observation operator $\mathcal{C}$, and convergence of the discrete inner products $\langle\!\langle \cdot, \cdot \rangle\!\rangle_{\mathcal{H}^n}$ and $\langle\!\langle \cdot, \cdot \rangle\!\rangle_{\mathcal{Z}^n}$ to the continuous inner products $\langle\!\langle \cdot, \cdot \rangle\!\rangle_{\mathcal{H}}$ and $\langle\!\langle \cdot, \cdot \rangle\!\rangle_{\mathcal{Z}}$, respectively. To carry out a convergence analysis, we first carefully state the discrete version of the problem.

Define $\mathcal{H}^n$ by

$$\mathcal{H}^n = \bigotimes_{k=0}^{n} H.$$

For $f \in C([0,t_F];H)$, we define $P_n f \in \mathcal{H}^n$ by

$$[P_n f]_k = f(t_k^n), \tag{3.5}$$

and for $f^n = \{f_k^n\}_{k=0}^n$ and $g^n = \{g_k^n\}_{k=0}^n$ in $\mathcal{H}^n$, the inner product is

$$\langle\!\langle f^n, g^n \rangle\!\rangle_{\mathcal{H}^n} \overset{\text{def}}{=} \sum_{k=0}^{n} w_k^n \langle f_k^n, g_k^n \rangle_H, \tag{3.6}$$

where $\{w_k^n\}_{k=0}^n$ is a given sequence of weights. The purpose of these weights is to facilitate the definition of $\langle\!\langle \cdot, \cdot \rangle\!\rangle_{\mathcal{H}^n}$ in such a way that it well approximates $\langle\!\langle \cdot, \cdot \rangle\!\rangle_{\mathcal{H}}$. Accordingly, we assume, essentially without loss of generality, that there exists an $M_1 < \infty$ independent of n, for which

$$\sum_{k=0}^{n} w_k^n \leq M_1. \tag{3.7}$$

This forms the basis for the following lemma.

LEMMA 1. *If a sequence $\{f^n\} \in \mathcal{H}^n$ has the property that for some constant $\nu > 0$,*

$$\max_{0 \leq k \leq n} \|f_k^n\|_H = \mathcal{O}(|\mathbf{h}|^\nu),$$

then

$$\|f^n\|_{\mathcal{H}^n} = \mathcal{O}(|\mathbf{h}|^\nu).$$

We also denote by $\langle\!\langle \cdot, \cdot \rangle\!\rangle_{E^n}$ the inner product that, loosely speaking, uses the standard Euclidean inner product for the time component. Specifically,

$$\langle\!\langle f^n, g^n \rangle\!\rangle_{E^n} \overset{\text{def}}{=} \sum_{k=0}^{n-1} \langle f_k^n, g_k^n \rangle_H. \tag{3.8}$$

This inner product will be used below in adjoint computations. The two inner products are related by

$$\begin{aligned} \langle\!\langle f^n, g^n \rangle\!\rangle_{\mathcal{H}^n} &= \langle\!\langle \mathcal{W}^n f^n, g^n \rangle\!\rangle_{E^n} \\ &= \langle\!\langle f^n, \mathcal{W}^n g^n \rangle\!\rangle_{E^n}, \end{aligned} \tag{3.9}$$

where $\mathcal{W}^n$ is the diagonal operator on $\mathcal{H}^n$ defined by

$$[\mathcal{W}^n f]_k = w_k^n f_k^n.$$

We assume that the time-marching scheme TMS is of the form

$$u^n = \mathcal{S}^n(q)\mathcal{R}^n(q) P_n f, \tag{3.10}$$

where $\mathcal{S}^n(q)$ is defined by recursion, for $g \in \mathcal{H}^n$, by

$$\begin{aligned}[][\mathcal{S}^n(q)g]_0 &= 0, \\ [\mathcal{S}^n(q)g]_{k+1} &= B_k(q)[\mathcal{S}^n(q)g]_k + h_k g_k \end{aligned} \tag{3.11}$$

for some bounded operator $B_k(q)$. Expressed in terms of components,

$$u_{k+1}^n = B_k(q)u_k^n + h_k[\mathcal{R}^n(q)P_n f]_k, \qquad k = 0, 1, \ldots, n-1. \tag{3.12}$$

See §4 for specific examples.

We make the following assumptions on the TMS.

(A1) The TMS is *stable* [10], i.e., there exists $M_2 > 0$ such that for $0 \le j \le n$,

$$\|\prod_{k=0}^{j} B_k(q)\| \le M_2.$$

The terms in this product are ordered from left to right with decreasing indices.

(A2) For some constant $\nu > 0$, $B_k(q) = e^{A(q)h_k} + \mathcal{O}(|\mathbf{h}|^{\nu+1})$.

(A3) With ν as in (A2), for any $f \in C^\nu([0, t_F]; H)$ we have

$$[\mathcal{R}^n(q)P_n f]_k = \frac{1}{h_k}\int_0^{h_k} e^{As} f(t_{k+1} - s)\, ds + \mathcal{O}(|\mathbf{h}|^\nu).$$

(A4) In the state equation (2.2), the forcing function f lies in $C^\nu([0, t_F]; H)$.

The following theorem is then readily established.

THEOREM 1. *If* (A1)–(A4) *hold, then*

$$\|u^n(q) - P_n u(q)\|_{\mathcal{H}^n} = \mathcal{O}(|\mathbf{h}|^\nu).$$

Proof. Defining $\varepsilon_k^n \overset{\text{def}}{=} u_k^n(q) - u(t_k^n; q) = u_k^n(q) - [P_n u(q)]_k$ and using (A2)–(A4), (2.2), and (3.12) we obtain

$$\varepsilon_{k+1}^n = B_k(q)\,\varepsilon_k^n + \mathcal{O}(|\mathbf{h}|^{\nu+1}). \tag{3.13}$$

From this we can use (A1) and induction to show that

$$\max_{0 \le k \le n} \|\varepsilon_k^n\|_H \le \mathcal{O}(|\mathbf{h}|^\nu).$$

The theorem now follows from Lemma 1. □

3.2. The rate of convergence of the gradients. We now address the convergence of the (state) directional derivative approximations $\delta_p T^n$ to $\delta_p T$. The results obtained in this subsection will be used later in proving convergence of the discrete costate approximations $\widetilde{\delta_p T^n}$ to $\delta_p T$.

We assume the existence of operators $Q_n : \mathcal{Z} \to \mathcal{Z}^n$, for which

$$z^n = Q_n z. \tag{3.14}$$

Also, note that $\mathcal{C}^n : \mathcal{H}^n \to \mathcal{Z}^n$. We make the following assumptions on the relationships between Q_n, $\mathcal{C}^n$, and the inner product approximations.

(A5) For $\eta, \zeta \in C^\nu([0, t_F]; Z)$, $\langle\!\langle Q_n\eta, Q_n\zeta \rangle\!\rangle_{\mathcal{Z}^n} - \langle\!\langle \eta, \zeta \rangle\!\rangle_{\mathcal{Z}} = \mathcal{O}(|\mathbf{h}|^\nu)$.

(A6) For $v \in C^\nu([0, t_F]; H)$, $\|(\mathcal{C}^n P_n - Q_n \mathcal{C})v\|_{\mathcal{Z}^n} = \mathcal{O}(|\mathbf{h}|^\nu)$.

We now consider the convergence of the directional derivatives. From (2.5) and (3.1),

$$\delta_p T^n(q) - \delta_p T(q) = \langle\!\langle r^n, \mathcal{C}^n \delta_p u^n(q) \rangle\!\rangle_{\mathcal{Z}^n} - \langle\!\langle r, \mathcal{C}\delta_p u(q) \rangle\!\rangle_{\mathcal{Z}} \\ = e_1 + e_2 + e_3, \tag{3.15}$$

where

$$e_1 = \langle\!\langle r^n - Q_n r, \mathcal{C}^n \delta_p u^n(q) \rangle\!\rangle_{\mathcal{Z}^n}, \tag{3.16}$$

$$e_2 = \langle\!\langle Q_n r, \mathcal{C}^n \delta_p u^n(q) - Q_n \mathcal{C}\delta_p u(q) \rangle\!\rangle_{\mathcal{Z}^n}, \tag{3.17}$$

$$e_3 = \langle\!\langle Q_n r, Q_n \mathcal{C}\delta_p u(q) \rangle\!\rangle_{\mathcal{Z}^n} - \langle\!\langle r, \mathcal{C}\delta_p u(q) \rangle\!\rangle_{\mathcal{Z}}. \tag{3.18}$$

If (A6) holds, the rate of convergence of e_1 is directly determined by the rate of convergence of solutions to the forward problem, while that of e_3 is assured under (A5). The rate convergence of e_2 is determined by how fast

$$\|\delta_p u_k^n(q) - P_n \delta_p u(q)\|_{\mathcal{H}^n} \longrightarrow 0.$$

Taking directional derivatives in the component form of the recursion (3.12),

$$\delta_p u_{k+1}^n = B_k(q)\delta_p u_k^n + \big(\delta_p B_k(q)\big) u_k^n + h_k [\delta_p \mathcal{R}^n(q) P_n f]_k. \tag{3.19}$$

On the other hand,

$$u(t_{k+1}^n) = e^{Ah_k} u(t_k^n) + \int_0^{h_k} e^{As} f(t_{k+1}^n - s)\, ds,$$

so taking directional derivatives gives

$$\delta_p u(t_{k+1}^n) = e^{A(q)h_k}\delta_p u(t_k^n) + \left(\delta_p e^{A(q)h_k}\right) u(t_k^n) + \delta_p \int_0^{h_k} e^{A(q)s} f(t_{k+1}^n - s)\, ds. \tag{3.20}$$

Comparing (3.19) and (3.20), one can show the desired convergence provided the directional derivative operator δ_p preserves convergence rates and smoothness in t. More precisely, we assume

(A2′) $\delta_p B_k(q) = \delta_p e^{A(q)h_k} + \mathcal{O}(|\mathbf{h}|^{\nu+1})$.

(A3′) $\delta_p[\mathcal{R}^n(q)P_n f]_k = \frac{1}{h_k}\delta_p \int_0^{h_k} e^{A(q)s} f(t_{k+1} - s)\, ds + \mathcal{O}(|\mathbf{h}|^{\nu})$.

(A4′) If $v(\cdot;q) \in C([0,t_F];H)$, then $\delta_p v(\cdot;q) \in C([0,t_F];H)$.

We then obtain the following analogue of Theorem 1.

THEOREM 2. *If assumptions* (A1)–(A4) *and* (A2′)–(A4′) *hold, then*

$$\|\delta_p u^n(q) - P_n \delta_p u(q)\|_{\mathcal{H}^n} = \mathcal{O}(|\mathbf{h}|^{\nu}).$$

Proof. For notational convenience, we suppress dependence on q. First we show that each $\|\delta_p u_k^n\|_H$ is bounded independently of k or n. Applying (3.19) recursively,

$$\delta_p u_{k+1}^n = \left(\Pi_{j=0}^k B_j\right) \delta_p u_0^n + \sum_{j=0}^{k} \left(\Pi_{i=j+1}^k B_i\right) g_j,$$

where

$$g_j = \delta_p B_j u_j^n + h_j [\delta_p \mathcal{R}^n P_n f]_j.$$

From assumption (A1),

$$\|\delta_p u_{k+1}^n\|_H \le M_2 \left(\|\delta_p u_0^n\|_H + (n+1) \max_{0\le j\le n-1} \|g_j\|_H \right).$$

But for each j, by assumptions (A2$'$)–(A3$'$) and Theorem 1,

$$\begin{aligned}\|g_j\|_H &\le \|\delta_p B_j\| \|u_j^n\|_H + h_j \|[\delta_p \mathcal{R}^n P_n f]_j\|_H \\ &\le h_j \left(\|\delta_p A\| \sup_{0\le t\le t_F} \|u(t)\|_H + \|\frac{1}{h_j}\delta_p \int_0^{h_j} e^{As} f(t_{k+1}-s)\,ds\|_H + \mathcal{O}(|\mathbf{h}|) \right).\end{aligned}$$

The integral term in the above inequality is bounded independently of j as a consequence of assumption (A4$'$).

Next, define $\varepsilon_k^n \overset{\text{def}}{=} \delta_p u_k^n - \delta_p u(t_k^n) = \delta_p u_k^n - [P_n \delta_p u]_k$. Subtracting (3.20) from (3.19);

$$\begin{aligned}\varepsilon_{k+1}^n &= B_k \varepsilon_k^n + (B_k - e^{Ah_k})\delta_p u_k^n + \delta_p B_k(u_k^n - u(t_k^n)) + (\delta_p B_k - \delta_p e^{Ah_k}) u(t_k^n) \\ &\quad + h_k \left(\delta_p [\mathcal{R}^n(q) P_n f]_k - \frac{1}{h_k}\delta_p \int_0^{h_k} e^{As} f(t_{k+1}^n - s)\,ds \right).\end{aligned} \tag{3.21}$$

From the boundedness of the $\delta_p u_k^n$'s, the result of Theorem 1, and the assumptions, all the terms on the right-hand side of this equation are $\mathcal{O}(|\mathbf{h}|^{\nu+1})$, except the first term. Thus this equation has the form (3.13), and the same argument used in the proof of Theorem 1 applies. □

From (3.15) we obtain the following result.

THEOREM 3. *Under assumptions* (A1)–(A6) *and* (A2$'$)–(A4$'$),

$$|\ \delta_p T^n(q) - \delta_p T(q)\ | = \mathcal{O}(|\mathbf{h}|^\nu).$$

Proof. Referring to (3.16), from Schwartz's inequality and the definitions of r, r^n, and z^n,

$$\begin{aligned}|e_1| &\le \|\mathcal{C}^n \delta_p u^n\|_{\mathcal{Z}^n} \|\mathcal{C}^n u^n - Q_n \mathcal{C} u\|_{\mathcal{Z}^n} \\ &\le \|\mathcal{C}^n\| \|\delta_p u^n\|_{\mathcal{H}^n} \left(\|\mathcal{C}^n\| \|u^n - P_n u\|_{\mathcal{H}^n} + \|(\mathcal{C}^n P_n - Q_n \mathcal{C})u\|_{\mathcal{Z}^n} \right).\end{aligned}$$

The first term within the above parentheses is $\mathcal{O}(|\mathbf{h}|^\nu)$ by Theorem 1, while the second is $\mathcal{O}(|\mathbf{h}|^\nu)$ by (A6). The $\|\delta_p u^n\|_{\mathcal{H}^n}$ are bounded by Theorem 2. Consequently, $|e_1| = \mathcal{O}(|\mathbf{h}|^\nu)$.

Similarly, from (3.17), Theorem 2, and assumption (A6),

$$\begin{aligned}|e_2| &\le \left(\|Q_n \mathcal{C} u\|_{\mathcal{Z}^n} + \|z^n\|_{\mathcal{Z}^n} \right) \left(\|\mathcal{C}^n\| \|\delta_p u^n - P_n \delta_p u\|_{\mathcal{H}^n} + \|(\mathcal{C}^n P_n - Q_n \mathcal{C})\delta_p u\|_{\mathcal{Z}^n} \right) \\ &= \mathcal{O}(|\mathbf{h}|^\nu).\end{aligned}$$

Finally, from (3.18) and (A5), $|e_3| = \mathcal{O}(|\mathbf{h}|^\nu)$. □

3.3. Continuous-time versus discrete-time observations.

In the following subsection we analyze the convergence of the costate approximations. In this subsection, we make some preliminary considerations toward that end, leading to a specification of two distinct classes of observation operators $\mathcal{C}$.

For brevity of notation, we set

$$\begin{aligned} \phi &\overset{\text{def}}{=} (\delta_p A)u \in \mathcal{H}, \\ \phi^n &\overset{\text{def}}{=} (\delta_p A)u^n \in \mathcal{H}^n. \end{aligned} \tag{3.22}$$

From (2.11) and (3.4) we obtain

$$\begin{aligned} \widetilde{\delta_p T^n} - \delta_p T &= \langle\!\langle y^n, \phi^n - P_n\phi \rangle\!\rangle_{\mathcal{H}^n} + \Big(\langle\!\langle y^n, P_n\phi \rangle\!\rangle_{\mathcal{H}^n} - \langle\!\langle y, \phi \rangle\!\rangle_{\mathcal{H}} \Big) \\ &\overset{\text{def}}{=} E_1 + E_2. \end{aligned} \tag{3.23}$$

As a consequence of Theorem 1, the boundedness of $\delta_p A$, and the fact that operators $\delta_p A$ and P_n commute, we have

$$\|\phi^n - P_n\phi\|_{\mathcal{H}^n} = \mathcal{O}(|\mathbf{h}|^\nu). \tag{3.24}$$

If the sequence $\{\|y^n\|_{\mathcal{H}^n}\}$ is bounded, then Schwartz's inequality and (3.24) imply that $|E_1| = \mathcal{O}(|\mathbf{h}|^\nu)$. But $\{\|y^n\|_{\mathcal{H}^n}\}$ is bounded if $|E_2| \to 0$. Hence, the rate of convergence of the costate approximations to the directional derivatives depends on the rate at which $E_2 \to 0$.

From (2.9) we see that

$$y = J\mathcal{S}_{(*)}J\mathcal{C}^* r, \tag{3.25}$$

where the operator $\mathcal{S}_{(*)} : \mathcal{H} \to \mathcal{H}$ has a representation

$$(\mathcal{S}_{(*)}g)(t) = \int_0^t e^{A(q)^* s} g(t-s)\, ds.$$

Similarly, from (3.2)

$$y^n = J^n \mathcal{S}^n_{(*)} \mathcal{R}^n_{(*)} J^n (\mathcal{C}^*)^n r^n. \tag{3.26}$$

Here, $\mathcal{S}^n_{(*)}$ and $\mathcal{R}^n_{(*)}$ are the operators that are obtained if A is replaced by $A^*(q)$ in the formation of $\mathcal{S}^n$ and $\mathcal{R}^n$; cf. (3.10), (3.11). The operator $(\mathcal{C}^*)^n : \mathcal{Z}^n \to \mathcal{H}^n$ is an approximation to the adjoint of $\mathcal{C} : \mathcal{H} \to \mathcal{Z}$. For now we select $(\mathcal{C}^*)^n = (\mathcal{C}^n)^*$, the adjoint of the operator $\mathcal{C}^n : \mathcal{H}^n \to \mathcal{Z}^n$. A different choice might be made on the basis of the discussion in §5.

Since the costate vector y is given in terms of an evolution equation with $\mathcal{C}^* r$ as the source term (and a similar statement is true for the approximations), we find it convenient to specify $\mathcal{C}$ further. This facilitates analysis of the convergence of $|E_2|$. It commonly happens in applications that $\mathcal{C}$ inherits a type of tensor-product structure from the state space $\mathcal{H} = L^2(0, t_F; H)$. In particular, we assume the existence of a bounded "spatial observation operator"

$$C : H \mapsto Z, \tag{3.27}$$

where Z is a Hilbert space related to $\mathcal{Z}$ as described below. We distinguish two cases that are of practical importance.

3.3.1. Observations continuous in time. In this first case, the observation operator $\mathcal{C}$ is "continuous in time." It is defined in terms of a space Z and a bounded operator $C : H \mapsto Z$ that acts at each t by

$$(\mathcal{C}u)(t) = Cu(\cdot, t) \in Z, \qquad 0 \le t \le t_F.$$

In this case,

$$\mathcal{Z} = C\big((0,t_F);Z\big), \qquad \mathcal{Z}^n = \bigotimes_{k=0}^{n} Z,$$

and $\mathcal{C}^n : \mathcal{H}^n \to \mathcal{Z}^n$ is defined by

$$[\mathcal{C}^n f^n]_k = C f_k^n, \qquad k = 0,1,\dots,n. \tag{3.28}$$

In addition, we define Q_n and the inner product approximation in a manner analogous to (3.5) and (3.6):

$$[Q_n \zeta]_k = \zeta(t_k^n), \qquad k = 0,1,\dots,n,$$

and

$$\langle\!\langle \eta^n, \zeta^n \rangle\!\rangle_{\mathcal{Z}^n} = \sum_{k=0}^{n} w_k^n \langle \eta_k^n, \zeta_k^n \rangle_Z.$$

Note that in this case, the validity of assumption (A5) is determined by the quadrature weights w_k^n, which also determine the accuracy by which $\langle\!\langle \cdot,\cdot \rangle\!\rangle_{\mathcal{H}^n}$ approximates $\langle\!\langle \cdot,\cdot \rangle\!\rangle_{\mathcal{H}}$. Assumption (A6) is trivially satisfied in this case since $\mathcal{C}^n P_n = Q_n \mathcal{C}$.

From (3.28) and the definition of the adjoint,

$$\begin{aligned} [\mathcal{C}^* r](t) &= C^* r(t), \qquad 0 \le t \le t_F, \\ [(\mathcal{C}^n)^* r^n]_k &= C^* r_k^n, \qquad 0 \le k \le n. \end{aligned}$$

The residuals r and r^n appearing here are given by

$$\begin{aligned} r(t) &= Cu(t) - z(t), \\ r_k^n &= Cu_k^n - z(t_k^n). \end{aligned}$$

3.3.2. Observations discrete in time. In the second case we assume that the observation operator $\mathcal{C}$ consists of some spatial observations taken at m discrete points τ_i, $i = 1,\dots,m$ in time. Accordingly, we assume a discrete observation space $\mathcal{Z}$ of the form

$$\mathcal{Z} = \bigotimes_{i=1}^{m} Z.$$

For simplicity we assume that, given the set of temporal observation points $\{\tau_i\}$, the time-marching grid $\mathbf{t}^n$ is always chosen so that $\{\tau_i\} \subset \mathbf{t}^n$, so that there is always an injective map

$$\kappa : \{1,2,\dots,m\} \mapsto \{0,1,2,\dots,n\}$$

such that for $f \in C((0,t_F];H) \subset \mathcal{H}$ and $f^n \in \mathcal{H}^n$,

$$[\mathcal{C}f]_i \overset{\text{def}}{=} Cf(t_{\kappa(i)}^n), \tag{3.29}$$

$$[\mathcal{C}^n f^n]_i \overset{\text{def}}{=} Cf_{\kappa(i)}^n. \tag{3.30}$$

In this case, since the observation space is already discrete, we define

$$\mathcal{Z}^n = \mathcal{Z}, \quad Q_n = I = \text{ the identity on } \mathcal{Z}, \tag{3.31}$$

and

$$\langle\!\langle \eta, \zeta \rangle\!\rangle_{\mathcal{Z}^n} = \langle\!\langle \eta, \zeta \rangle\!\rangle_{\mathcal{Z}} = \sum_{i=1}^{m} \langle \eta_i, \zeta_i \rangle_Z.$$

In this case, assumption (A5) is always true since $\mathcal{Z}^n = \mathcal{Z}$. Also, (A6) is always true since, as in the continuous-time observation case, $\mathcal{C}^n P_n = Q_n \mathcal{C}$. (This is true since we have assumed that $\{\tau_i\} \subset \mathbf{t}^n$.)

From (3.29) and the definition of the adjoint, we find that $\mathcal{C}^* : \mathcal{Z} \mapsto \mathcal{H}$ is given, for $r = \{r_i\}_{i=1}^m \in \mathcal{Z}$, by

$$[\mathcal{C}^* r](t) = \sum_{i=1}^{m} C^* r_i \, \delta(t - \tau_i), \qquad 0 \leq t \leq t_F, \tag{3.32}$$

where $\delta(\cdot)$ denotes the Dirac delta function. The discrete analogue of (3.32) is

$$[(\mathcal{C}^*)^n r]_k = [(\mathcal{C}^n)^* r]_k = \sum_{i=1}^{m} \frac{1}{w_k} C^* r_i \delta_{\kappa(i),k}, \qquad 0 \leq k \leq n,$$

where $\delta_{\kappa,k}$ denotes the Kronecker delta function for integer pairs.

3.4. The rate of convergence of the costate approximations. As discussed in §3.3, immediately following (3.23), it suffices to consider the convergence of the term $|E_2|$ to zero. We address the simple case, that of continuous-time observations, first. In this case, νth-order convergence of E_2 results if the observed data z is smooth enough.

THEOREM 4. *If* (A1)–(A6) *and* (A2′)–(A4′) *hold and if the data z lies in* $C^\nu([0, t_F]; Z)$, *then*

$$\|y^n - P_n y\|_{\mathcal{H}^n} = \mathcal{O}(|\mathbf{h}|^\nu). \tag{3.33}$$

Proof. From (3.25) and (3.25),

$$\begin{aligned} \|y^n - P_n y\|_{\mathcal{H}^n} \leq{}& \|J^n\| \|\mathcal{S}^n_{(*)} \mathcal{R}^n_{(*)}\| \|J^n (\mathcal{C}^n)^* r^n - P_n J \mathcal{C}^* r\|_{\mathcal{H}^n} \\ &+ \|(J^n \mathcal{S}^n_{(*)} \mathcal{R}^n_{(*)} P_n - P_n J \mathcal{S}_{(*)}) J \mathcal{C}^* r\|_{\mathcal{H}^n}. \end{aligned} \tag{3.34}$$

Note that J, $\mathcal{C}$, and $\mathcal{C}^*$ each preserve smoothness with respect to t. From the smoothness of z and the smoothness of u (which follows from (A4)), $J\mathcal{C}^* r \in C^{\nu+1}([0, t_F]; H)$. Applying Theorem 1 with $J\mathcal{C}^* r$ in place of the forcing function f and A^* in place of A, and noting that

$$P_n J = J^n P_n, \tag{3.35}$$

$$\|J^n\| = 1, \tag{3.36}$$

the last term in (3.34) is $\mathcal{O}(|\mathbf{h}|^\nu)$. Also, by (A1) and (A3), $\|\mathcal{S}^n_{(*)} \mathcal{R}^n_{(*)}\|$ is bounded independently of n. From (3.35) and (3.36), it suffices to show

$$\|(\mathcal{C}^n)^* r^n - P_n \mathcal{C}^* r\|_{\mathcal{H}^n} = \mathcal{O}(|\mathbf{h}|^\nu).$$

However, $[(C^n)^* r^n]_k - [P_n C^* r]_k = C^*C(u_k^n - u(t_k^n))$, so this follows from Theorem 1 and the boundedness of C and C^*. □

With the result of this theorem and an assumption that the quadrature weights are chosen correctly (cf. (3.6)), we obtain costate convergence.

COROLLARY 1. *If the hypotheses of Theorem* 4 *hold and if*

$$|\langle\!\langle P_n g_1, P_n g_2 \rangle\!\rangle_{\mathcal{H}^n} - \langle\!\langle g_1, g_2 \rangle\!\rangle_{\mathcal{H}}| = \mathcal{O}(|\mathbf{h}|^\nu), \tag{3.37}$$

for any g_1 *and* g_2 *in* $C^\nu([0, t_f]; H)$, *then*

$$|\widetilde{\delta_p T^n} - \delta_p T| = \mathcal{O}(|\mathbf{h}|^\nu).$$

Proof. Theorem 4 implies that the sequence $\{\|y^n\|_{\mathcal{H}^n}\}$ is bounded. From the discussion immediately following (3.23), it suffices to show that $E_2 = \mathcal{O}(|\mathbf{h}|^\nu)$. However,

$$\begin{aligned} |E_2| &= |\langle\!\langle y^n, P_n\phi \rangle\!\rangle_{\mathcal{H}^n} - \langle\!\langle y, \phi \rangle\!\rangle_{\mathcal{H}}| \\ &\leq \|y^n - P_n y\|_{\mathcal{H}^n} \|P_n \phi\|_{\mathcal{H}^n} \\ &\quad + |\langle\!\langle P_n y, P_n \phi \rangle\!\rangle_{\mathcal{H}^n} - \langle\!\langle y, \phi \rangle\!\rangle_{\mathcal{H}}|. \end{aligned}$$

The first term is $\mathcal{O}(|\mathbf{h}|^\nu)$ by Theorem 4. Since y and ϕ are sufficiently smooth, the second term is $\mathcal{O}(|\mathbf{h}|^\nu)$ by (3.37). □

We next address the case of discrete-time observations. In this case, $\mathcal{C}^* r$ is given by a linear combination of Dirac delta functions, i.e., (3.32). Since this is the forcing term in the costate equation, the question of how fast (if at all) $\widetilde{\delta_p T^n} \to \delta_p T$ is rather delicate. In particular, the smoothness of $\mathcal{C}^* r$ played a crucial role in the proof of Theorem 4. However, in the present case, $\mathcal{C}^* r$ is not smooth at all; it only exists as a distribution. Thus the proof of Theorem 4 (and hence, Corollary 1) is not valid in this case.

On the other hand, the convergence in (3.33) is stronger than what is actually needed. It is essentially a statement of *strong*, or pointwise, convergence in $\mathcal{H}$. From (3.23), we see that a sort of weak* convergence would be sufficient. Rewriting E_2 as

$$E_2 = \langle\!\langle P_n^* y^n - y, \phi \rangle\!\rangle_{\mathcal{H}} \tag{3.38}$$

and noting that $\phi \in C^{\nu+1}([0, t_F]; H)$, we need only establish νth-order weak* convergence of $P_n^* y$ to y in the dual space of $C^{\nu+1}([0, t_F]; H)$.

There are seemingly natural situations in which this weak* convergence does not apply, and a number of subtleties arise in our exploration of these matters. These will be discussed shortly. There is, however, one class of problems that we fully analyze here. It covers a variety of second-order methods with uniform meshes $\mathbf{t}^n$.

THEOREM 5. *Assume that*

(a) *Assumptions* (A1)–(A4) *hold*;

(b) *In* (A2)–(A4) *we have* $\nu = 2$;

(c) *The observation points* τ_i *all lie in the interior of* $(0, t_F)$.

(d) *The partition* $\mathbf{t}^n$ *is uniform, in the sense that* $h_k = h \overset{\text{def}}{=} t_F/n$ *for* $1 \leq k \leq n$;

(e) *The quadrature weights* w_k *in equation* (3.6) *are given by* $w_0 = h/2 = w_n$ *and* $w_k = h$ *for* $1 \leq k \leq n-1$ (*note that these weights correspond to the "trapezoid rule" for numerical quadrature*);

(f) *The operators* B_k *in terms of which* $\mathcal{S}^n$ *is defined* (*cf.* (3.11)) *are the same for all* k; *there is an operator* B *for which*

$$B_k = B, \qquad 0 \le k \le n-1; \tag{3.39}$$

(g) *The operator* $\mathcal{R}^n$ *may be written, for* $g \in \mathcal{H}^n$, *as*

$$[\mathcal{R}^n g]_k = R_0 g_k + R_1 g_{k+1}, \qquad 0 \le k \le n-1, \tag{3.40}$$

$$[\mathcal{R}^n g]_n = R_0 g_n. \tag{3.41}$$

(*Note from* (3.11) *that the last component,* $[\mathcal{R}^n f]_n$, *plays no role in the* TMS. *Our purpose in defining it this way here will become clear below.*)

(h) *The operators* B, R_0, *and* R_1 *are all rational functions of* $A(q)$.

Then

$$|\widetilde{\delta_p T^n} - \delta_p T| = \mathcal{O}(h^2).$$

Proof. As in Corollary 1, it suffices to show $|E_2| = \mathcal{O}(h^2)$. This leads us to the study of $\langle\!\langle\, y^n, P_n\phi \,\rangle\!\rangle_{\mathcal{H}^n}$. From (3.26), we see that

$$\langle\!\langle\, y^n, P_n\phi \,\rangle\!\rangle_{\mathcal{H}^n} = \langle\!\langle\, J^n \mathcal{S}^n_{(*)} \mathcal{R}^n_{(*)} J^n (\mathcal{C}^*)^n r^n,\ P_n\phi \,\rangle\!\rangle_{\mathcal{H}^n}. \tag{3.42}$$

We take adjoints in the inner product on the right-hand side of this equation, and the main effort in this proof lies in the subsequent simplification. In particular, our goal is to show that

$$\langle\!\langle\, J^n \mathcal{S}^n_{(*)} \mathcal{R}^n_{(*)} J^n (\mathcal{C}^*)^n r^n,\ P_n\phi \,\rangle\!\rangle_{\mathcal{H}^n} = \langle\!\langle\, r^n, \mathcal{C}^n \mathcal{S}^n \mathcal{R}^n P_n\phi \,\rangle\!\rangle_{\mathcal{Z}}. \tag{3.43}$$

We now proceed with the details. First note from (3.9) that

$$\langle\!\langle\, y^n, P_n\phi \,\rangle\!\rangle_{\mathcal{H}^n} = \langle\!\langle\, \mathcal{W}^n y^n, P_n\phi \,\rangle\!\rangle_{E^n}.$$

However, since $[P_n\phi]_0 = \phi(0) = 0$ and $[y^n]_n = 0$, from Theorem 5(e) we obtain

$$\langle\!\langle\, y^n, P_n\phi \,\rangle\!\rangle_{\mathcal{H}^n} = h\langle\!\langle\, y^n, P_n\phi \,\rangle\!\rangle_{E^n}. \tag{3.44}$$

Referring to (3.42) and (3.44) and noting that $J^n J^n$ equals the identity, we obtain

$$\begin{aligned}
\langle\!\langle\, y^n, P_n\phi \,\rangle\!\rangle_{\mathcal{H}^n} &= h\langle\!\langle\, \left(J^n \mathcal{S}^n_{(*)} J^n\right) \left(J^n \mathcal{R}^n_{(*)} J^n\right) (\mathcal{C}^n)^* r^n,\ P_n\phi \,\rangle\!\rangle_{E^n} \\
&= h\langle\!\langle\, (\mathcal{C}^n)^* r^n,\ \left(J^n \mathcal{R}^n_{(*)} J^n\right)^T \left(J^n \mathcal{S}^n_{(*)} J^n\right)^T P_n\phi \,\rangle\!\rangle_{E^n} \\
&= h\langle\!\langle\, (\mathcal{C}^n)^* r^n,\ \left(J^n (\mathcal{R}^n_{(*)})^T J^n\right) \left(J^n (\mathcal{S}^n_{(*)})^T J^n\right) P_n\phi \,\rangle\!\rangle_{E^n},
\end{aligned} \tag{3.45}$$

where $\cdot^T$ denotes the adjoint with respect to the E^n-inner product defined in (3.9). We have used here the fact that J^n is self-adjoint with respect to $\langle\!\langle\, \cdot, \cdot \,\rangle\!\rangle_{E^n}$.

Next we assert that

$$J^n (\mathcal{S}^n_{(*)})^T J^n = \mathcal{S}^n, \tag{3.46}$$

$$J^n (\mathcal{R}^n_{(*)})^T J^n = \mathcal{R}^n. \tag{3.47}$$

From (3.11), (3.39), and (3.40), the operators $\mathcal{S}^n$ and $\mathcal{R}^n$ have the following block Toeplitz matrix representations:

$$\mathcal{S}^n = h\begin{pmatrix} 0 & 0 & 0 & \cdots & 0 & 0 & 0 & 0 \\ I & 0 & 0 & \cdots & 0 & 0 & 0 & 0 \\ B & I & 0 & \cdots & 0 & 0 & 0 & 0 \\ B^2 & B & I & \ddots & 0 & 0 & 0 & 0 \\ B^3 & B^2 & B & \ddots & 0 & 0 & 0 & 0 \\ \vdots & \vdots & \ddots & \ddots & \ddots & \ddots & \ddots & \vdots \\ B^{n-2} & B^{n-3} & B^{n-4} & \ddots & B & I & 0 & 0 \\ B^{n-1} & B^{n-2} & B^{n-3} & \cdots & B^2 & B & I & 0 \end{pmatrix} \tag{3.48}$$

and

$$\mathcal{R}^n = \begin{pmatrix} R_0 & R_1 & 0 & \cdots & 0 & 0 & 0 \\ 0 & R_0 & R_1 & \ddots & 0 & 0 & 0 \\ \vdots & \ddots & \ddots & \ddots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \ddots & R_0 & R_1 & 0 \\ 0 & 0 & 0 & \ddots & 0 & R_0 & R_1 \\ 0 & 0 & 0 & 0 & 0 & 0 & R_0 \end{pmatrix}. \tag{3.49}$$

The blocks are operators on H. The operators $\mathcal{S}^n_{(*)}$ and $\mathcal{R}^n_{(*)}$ have an identical block Toeplitz form, except B, R_0, and R_1 are replaced by their H-adjoints. From consideration of these block matrix representations, the action of J^n, and the E^n-adjoint, we can verify that (3.46) and (3.47) hold. Thus (3.45) simplifies to

$$\langle\!\langle y^n, P_n\phi \rangle\!\rangle_{\mathcal{H}^n} = h\langle\!\langle (\mathcal{C}^n)^* r^n,\ \mathcal{R}^n\mathcal{S}^n P_n\phi \rangle\!\rangle_{E^n}. \tag{3.50}$$

Thus the question arises as to how much $\mathcal{S}^n\mathcal{R}^n$ and $\mathcal{R}^n\mathcal{S}^n$ differ. Since B, R_0, and R_1 are rational functions of $A(q)$, they commute with each other. This and computations based directly on (3.11) and (3.40) reveal that, for arbitrary $g \in \mathcal{H}^n$,

$$\frac{1}{h}[\mathcal{S}^n\mathcal{R}^n g]_k = (R_0 + BR_1)\Big(\sum_{j=1}^{k-1} B^{k-1-j} g_j\Big) + R_0 B^{k-1} g_0 + R_1 g_k$$
$$\text{for } 0 \le k \le n,$$
$$\frac{1}{h}[\mathcal{R}^n\mathcal{S}^n g]_k = (R_0 + BR_1)\Big(\sum_{j=1}^{k-1} B^{k-1-j} g_j\Big) + (R_0 + BR_1)B^{k-1} g_0 + R_1 g_k$$
$$\text{for } 0 \le k \le n-1,$$
$$\frac{1}{h}[\mathcal{R}^n\mathcal{S}^n g]_n = R_0\Big(\sum_{j=1}^{n-1} B^{n-1-j} g_j\Big) + R_0 B^{n-1} g_0.$$

These results may also be obtained directly from the matrix representations (3.48) and (3.49). Thus the commutator $\mathcal{E}^n \overset{\text{def}}{=} \mathcal{S}^n\mathcal{R}^n - \mathcal{R}^n\mathcal{S}^n$ is given by

$$[\mathcal{E}^n g]_k = hR_1 \begin{cases} -B^k g_0 & \text{for } 0 \le k \le n-1, \\ \sum_{j=1}^{n-1} B^{n-j} g_j & \text{for } k = n. \end{cases} \tag{3.51}$$

Equation (3.51) holds for any $g \in \mathcal{H}^n$. In the case of present interest, namely (3.50), the role of g is played by $P_n\phi$, and $[P_n\phi]_0 = 0$. Thus from (3.50) and (3.51), we have

$$\langle\!\langle y^n, P_n\phi \rangle\!\rangle_{\mathcal{H}^n} = h\langle\!\langle (\mathcal{C}^n)^* r^n, \mathcal{S}^n\mathcal{R}^n P_n\phi \rangle\!\rangle_{E^n} - h^2\Big\langle [(\mathcal{C}^n)^* r^n]_n, R_1 \sum_{j=1}^{n-1} B^{n-j}\phi(t_j) \Big\rangle_H. \tag{3.52}$$

However, the observation points τ_i are all less than t_F and yet occur at mesh points—see the discussion leading up to (3.30). Thus $[(\mathcal{C}^n)^* r^n]_n = 0$, and the last term in this equation drops out. Also, since $[\mathcal{S}^n\mathcal{R}^n P_n\phi]_0 = 0$ and $[(\mathcal{C}^n)^* r^n]_n = 0$, we obtain from Theorem 5(e)

$$\begin{aligned} \langle\!\langle y^n, P_n\phi \rangle\!\rangle_{\mathcal{H}^n} &= \langle\!\langle (\mathcal{C}^n)^* r^n, \mathcal{W}^n\mathcal{S}^n\mathcal{R}^n P_n\phi \rangle\!\rangle_{E^n} \\ &= \langle\!\langle (\mathcal{C}^n)^* r^n, \mathcal{S}^n\mathcal{R}^n P_n\phi \rangle\!\rangle_{\mathcal{H}^n} \\ &= \langle\!\langle r^n, \mathcal{C}^n\mathcal{S}^n\mathcal{R}^n P_n\phi \rangle\!\rangle_{\mathcal{Z}}. \end{aligned} \tag{3.53}$$

Combining this with (3.42) yields (3.43).

Now, $\|r^n - r\|_{\mathcal{Z}} = \mathcal{O}(|\mathbf{h}|^2)$ by Theorem 1. By reasoning similar to that in the proof of Theorem 1, since ϕ is smooth, we also have $\|\mathcal{S}^n\mathcal{R}^n P_n\phi - P_n\mathcal{S}\phi\|_{\mathcal{H}^n} = \mathcal{O}(|\mathbf{h}|^2)$. Consequently, the right-hand side of (3.43) may be replaced by $\langle\!\langle r, \mathcal{C}^n P_n\mathcal{S}\phi \rangle\!\rangle_{\mathcal{Z}}$ to within second order. Using this, the definition of E_2 in (3.23), and the fact that $\langle\!\langle y, \phi \rangle\!\rangle_{\mathcal{H}} = \langle\!\langle r, \mathcal{C}\mathcal{S}\phi \rangle\!\rangle_{\mathcal{H}^n}$, we find that

$$E_2 = \langle\!\langle r, \mathcal{C}^n P_n\mathcal{S}\phi \rangle\!\rangle_{\mathcal{Z}} - \langle\!\langle r, \mathcal{C}\mathcal{S}\phi \rangle\!\rangle_{\mathcal{Z}} + \mathcal{O}(|\mathbf{h}|^2).$$

Now the result follows from assumption (A6). □

A result similar to Theorem 5 can be obtained for certain fourth-order methods (including Runge–Kutta) on a uniform mesh, provided that the forcing function f is zero at $t = 0$ and provided that appropriate quadrature weights are chosen. Surprisingly, these weights correspond to *second-order* quadrature rather than fourth-order quadrature.

THEOREM 6. *Assume that the forcing term f satisfies $f(0) = 0$, and that*

(a) *Same as Theorem* 5(a).

(b) *In* (A2)–(A4) *we have $\nu = 4$;*

(c) *Same as Theorem* 5(c).

(d) *Same as Theorem* 5(d).

(e) *Same as Theorem* 5(e).

(f) *Same as Theorem* 5(f).

(g) *The operator $\mathcal{R}^n$ may be written, for $g \in \mathcal{H}^n$ and $1 \le k \le n-2$, as*

$$[\mathcal{R}^n g]_k = R_{-1}g_{k-1} + R_0 g_k + R_1 g_{k+1} + R_2 g_{k+2}$$

with $R_{-1} = R_2 + \mathcal{O}(h)$.

(h) *Same as Theorem* 5(h).

Then,

$$|\widetilde{\delta_p T^n} - \delta_p T| = \mathcal{O}(h^4).$$

Proof. For given $g \in \mathcal{H}^n$, $[\mathcal{R}^n g]$ is defined for all $k > 0$ but $[\mathcal{R}^n g]_0$ is unspecified. Define $\widetilde{\mathcal{R}}^n$ by

$$[\widetilde{\mathcal{R}}^n g]_k = \begin{cases} R_0 g_0 + R_1 g_1 + R_2 g_2 & \text{for } k = 0, \\ [\mathcal{R}^n g]_k & \text{for } 1 \le k \le n-2, \\ R_{-1} g_{n-2} + R_0 g_{n-1} + R_1 g_n & \text{for } k = n-1, \\ R_{-1} g_{n-1} + R_0 g_n & \text{for } k = n. \end{cases} \tag{3.54}$$

Note that $\widetilde{\mathcal{R}}^n \neq \mathcal{R}^n$ (because (A3) is satisfied by $\mathcal{R}^n$ and not necessarily by $\widetilde{\mathcal{R}}^n$), but that $(\widetilde{\mathcal{R}}^n - \mathcal{R}^n)(\mathcal{C}^n)^* r^n = 0$ in $\mathcal{H}^n$ by (c). Therefore we may replace $\mathcal{R}^n$ by $\widetilde{\mathcal{R}}^n$ in (3.26). Since the block matrix representation of $\widetilde{\mathcal{R}}^n$ is Toeplitz, by following reasoning similar to that leading to (3.50) in the proof of Theorem 5 we obtain

$$\langle\langle y^n, P_n\phi \rangle\rangle_{\mathcal{H}^n} = h\langle\langle (\mathcal{C}^n)^* r^n, \widetilde{\mathcal{R}}^n \mathcal{S}^n P_n \phi \rangle\rangle_{E^n}. \tag{3.55}$$

As in the proof of Theorem 5, we now consider the extent to which $\widetilde{\mathcal{R}}^n$ and $\mathcal{S}^n$ fail to commute. Our goal here is to show that $\mathcal{E}^n P_n \phi = \mathcal{O}(h^4)$, where

$$\mathcal{E}^n \stackrel{\text{def}}{=} \widetilde{\mathcal{R}}^n \mathcal{S}^n - \mathcal{S}^n \mathcal{R}^n.$$

Then we may invoke reasoning similar to that in and following (3.52). With

$$\widetilde{\mathcal{E}^n} \stackrel{\text{def}}{=} \widetilde{\mathcal{R}}^n \mathcal{S}^n - \mathcal{S}^n \widetilde{\mathcal{R}}^n,$$

we have

$$\mathcal{E}^n = \widetilde{\mathcal{E}^n} + \mathcal{S}^n(\widetilde{\mathcal{R}}^n - \mathcal{R}^n). \tag{3.56}$$

From (3.11), (3.54), and the fact that $\phi(0) = 0$, we find after lengthy but straightforward computations that for $0 \le k \le n-2$

$$[\widetilde{\mathcal{E}^n} P_n \phi]_k = h B^k R_2 \phi(t_1^n). \tag{3.57}$$

Also, from the definitions of $\mathcal{S}^n$, $\mathcal{R}^n$ and $\widetilde{\mathcal{R}}^n$ and from (A3), we find that for $1 \le k \le n$,

$$\begin{aligned}\left[\mathcal{S}^n(\widetilde{\mathcal{R}}^n - \mathcal{R}^n)P_n\phi\right]_k = hB^{k-1}\big(&R_1\phi(t_1^n) + R_2\phi(t_2^n) \\ &-\frac{1}{h}\int_0^h e^{As}\phi(t_1^n - s)\,ds\big) + \mathcal{O}(h^4).\end{aligned}$$

Substituting these into (3.56) yields, for $0 \le k \le n-2$,

$$\begin{aligned}[\mathcal{E}^n P_n\phi]_k = hB^{k-1}\Big[&(BR_2 + R_1)\phi(t_1^n) + R_2\phi(t_2^n) \\ &-\frac{1}{h}\int_0^h e^{As}\phi(t_1^n - s)\,ds + \mathcal{O}(h^4)\Big].\end{aligned} \tag{3.58}$$

The integral term in this equation is approximated to within fourth-order accuracy if we replace $\phi(t)$ with its second-order Taylor polynomial $\pi(t)$ about $t = 0$ (this follows by arguments used to show fourth-order convergence of Simpson's rule):

$$\frac{1}{h}\int_0^h e^{As}\phi(t_1^n - s)\,ds = \frac{1}{h}\int_0^h e^{As}\pi(t_1^n - s)\,ds + \mathcal{O}(h^4).$$

Now, $\pi(t)$ can be evaluated at $t_i^n = ih$ for $i = -1, \dots, 2$, so that we can use assumption (A3) to make the substitution

$$\frac{1}{h}\int_0^h e^{As}\phi(t_1^n - s)\,ds = R_{-1}\pi(-h) + R_0\pi(0) + R_1\pi(h) + R_2\pi(2h).$$

However, $f(0) = 0$ by hypothesis, so that $u'(0) = 0 = \phi'(0)$, which means that $\pi(t) = t^2\phi''(0)/2$. Thus

$$\frac{1}{h}\int_0^h e^{As}\phi(t_1^n - s)\,ds = \frac{h^2}{2}(\,R_{-1} + R_1 + 4R_2\,)\phi''(0) + \mathcal{O}(h^4).$$

Also, $\phi(t_1^n)$ and $\phi(t_2^n)$ can be replaced by $\pi(t_1^n)$, and $\pi(t_2^n)$, respectively, to within third-order accuracy in (3.58). This yields

$$[\mathcal{E}^n P_n\phi]_k = \frac{h^3}{2}B^{k-1}(BR_2 - R_{-1})\phi''(0) + \mathcal{O}(h^4).$$

However, $BR_2 = R_{-1} + \mathcal{O}(h)$ by Theorem 6(g) and the fact that $B = I + \mathcal{O}(h)$ (which is true by (A2)). Thus we arrive at

$$[\mathcal{E}^n P_n\phi]_k = \frac{h^4}{2}B^{k-1}\phi''(0) + \mathcal{O}(h^4),$$

and so we have shown that for $1 \le k \le n-2$,

$$[\widetilde{\mathcal{R}}^n\mathcal{S}^n P_n\phi]_k = [\mathcal{S}^n\mathcal{R}^n P_n\phi]_k + \mathcal{O}(h^4). \tag{3.59}$$

Using this, (3.55), and reasoning similar to that in (3.53) we obtain

$$\begin{aligned}\langle\!\langle\, y^n, P_n\phi \,\rangle\!\rangle_{\mathcal{H}^n} &= \langle\!\langle\, r, \mathcal{S}\phi \,\rangle\!\rangle_{\mathcal{Z}} + \mathcal{O}(h^4)\\ &= \langle\!\langle\, y, \phi \,\rangle\!\rangle_{\mathcal{H}^n} + \mathcal{O}(h^4). \qquad \square\end{aligned}$$

It is also natural to ask what happens if we use a fourth-order TMS with a fourth-order quadrature scheme, such as Simpson's rule. We show that the gradient approximations thus computed fail to converge, although their *directions* in the space $\mathcal{Q}$ converge. In particular, the RK4/Simpson costate approximations converge quadratically to (3/2) times the true gradient.

COROLLARY 2. *Assume that all of the hypotheses of Theorem* 6 *hold except* (e), *and assume further that*

(e′) *The quadrature weights are those arising from Simpson's rule, so that*

$$\mathcal{W}^n = \mathcal{W}^n_{\text{Simp}} \overset{\text{def}}{=} \frac{h}{3}\,\text{diag}(1, 4, 2, \dots, 2, 4, 1). \tag{3.60}$$

(i) *The limit $n \to \infty$ is taken by repeated doubling of n. Precisely, there is an integer n_0 such that each n equals $n_0 2^l$ for some positive integer l.*

Then,

$$|\widetilde{\delta_p T^n} - (3/2)\delta_p T| = \mathcal{O}(h^2).$$

By arguments similar to those used to obtain (3.55), we find that

$$\langle\!\langle\, y^n, P_n\phi \,\rangle\!\rangle_{\mathcal{H}^n} = \langle\!\langle\, (\mathcal{C}^n)^* r^n,\ \widetilde{\mathcal{R}}^n\mathcal{S}^n\mathcal{W}^n_{\text{Simp}}P_n\phi \,\rangle\!\rangle_{E^n},$$

where $\widetilde{\mathcal{R}}^n$ is as in (3.54). Define

$$v \overset{\text{def}}{=} \mathcal{S}^n \mathcal{W}^n_{\text{Simp}} P_n \phi,$$

$$\tilde{v} \overset{\text{def}}{=} \mathcal{S}^n \mathcal{W}^n_{\text{Trap}} P_n \phi,$$

where $\mathcal{W}^n_{\text{Trap}}$ is the weight matrix corresponding to the trapezoidal rule:

$$\mathcal{W}^n = h\,\text{diag}(1/2, 1, 1, \dots, 1, 1/2).$$

We show that $\|\tilde{v} - v\|_{\mathcal{H}^n} = \mathcal{O}(h^2)$, which implies that

$$\langle\!\langle y^n, P_n\phi \rangle\!\rangle_{\mathcal{H}^n} = h \langle\!\langle (\mathcal{C}^n)^* r^n, \, \widetilde{\mathcal{R}}^n \mathcal{S}^n P_n \phi \rangle\!\rangle_{E^n} + \mathcal{O}(h^2). \tag{3.61}$$

We show by induction that $\|\tilde{v} - v\|_{\mathcal{H}^n} = \mathcal{O}(h^2)$. We note first that $\tilde{v}_0 = v_0$. So we assume that for some $k \geq 1$ we have $\tilde{v}_{k-1} = v_{k-1} + \mathcal{O}(h^2)$. We note from (3.11) that for any $k = 1, \dots, n-1$,

$$v_{k+1} = B^2 v_{k-1} + h\Big(B\phi(t^n_{k-1}) w^n_{k-1} \; + \; \phi(t^n_k) w^n_k \Big).$$

Here, $\{w^n_k\}$ are the Simpson weights—the diagonal entries of $\mathcal{W}^n_{\text{Simp}}$. Thus,

$$v_{k+1} + v_{k-1} = \big(B^2 + I \big) v_{k-1} + h\Big(B\phi(t^n_{k-1}) w^n_{k-1} \; + \; \phi(t^n_k) w^n_k \Big).$$

From the definition of $\mathcal{W}^n_{\text{Simp}}$, the smoothness of ϕ, and the fact that $B = I + \mathcal{O}(h)$, it follows that the second parenthesized term above is equal to $h\big(2h\phi(t^n_k) + \mathcal{O}(h^2) \big)$. Also, by (A2) we have $B^2 + I = 2B + \mathcal{O}(h^2)$. Using these observations and the definition of $\mathcal{W}^n_{\text{Trap}}$ and dividing by 2, we thus find that

$$\frac{v_{k+1} + v_{k-1}}{2} = \big(B + \mathcal{O}(h^2) \big) v_{k-1} + h\Big([\mathcal{W}^n_{\text{Trap}}\phi]_{k-1} \; + \; \mathcal{O}(h^2) \Big).$$

By the induction hypothesis, $v_{k-1} = \tilde{v}_{k-1} + \mathcal{O}(h^2)$, so we may replace v_{k-1} by $\tilde{v}_{k-1}$ on the right-hand side of this equation. The resulting expression is then, by definition, the formula for $\tilde{v}_k$. Thus we have

$$\big(v_{k+1} + v_{k-1} \big)/2 = \tilde{v}_k + \mathcal{O}(h^2).$$

Since it is also true that

$$\big(\tilde{v}_{k+1} + \tilde{v}_{k-1} \big)/2 = \tilde{v}_k + \mathcal{O}(h^2),$$

we obtain $\tilde{v}_{k+1} = v_{k+1} + \mathcal{O}(h^2)$ and hence also $\tilde{v}_k = v_k + \mathcal{O}(h^2)$. Thus (3.61) holds. We may then use the reasoning leading to (3.59) to obtain

$$\langle\!\langle y^n, P_n\phi \rangle\!\rangle_{\mathcal{H}^n} = h \langle\!\langle (\mathcal{C}^n)^* r^n, \, \mathcal{S}^n \widetilde{\mathcal{R}}^n P_n \phi \rangle\!\rangle_{E^n} + \mathcal{O}(h^2). \tag{3.62}$$

Next we turn our attention to $(\mathcal{C}^n)^* r^n \overset{\text{def}}{=} g$. Referring to (3.3) we see that g depends linearly on $\{1/w_k\}$. We show that if the Simpson weights are replaced by the trapezoidal weights, then the result—call it $\tilde{g}$—satisfies $g = (3/2)\tilde{g}$. Indeed, this follows easily from Corollary 2(i) by which we are guaranteed that $\kappa(i)$ is *even* for all $i = 1, \dots, m$, so that the Simpson weights w_k satisfy $w_{\kappa(i)} = (2/3)h$ for all $i = 1, \dots, m$. However, $(2/3)h$ is simply $(3/2)$ times the $\kappa(i)$th *trapezoidal* weight. Thus the right-hand side of (3.62) is equal, to within $\mathcal{O}(h^2)$, to $(3/2)$ times the right side of (3.55). Thus the result of Theorem 6 applies, and we obtain the desired result. □

4. Numerical examples. To illustrate the results of the previous section we focus on a spline-based Galerkin approximation (with fixed dimensionality) of the diffusion equation (1.2) in one space dimension and $t_F = 1$, and directional derivatives for the corresponding least-squares functional (2.1). Accordingly we take $\Omega = (0,1)$, and for fixed $N \geq 1$ take as basis functions $\{b_i\}_{i=1}^N$ the usual piecewise linear "hat" functions on a uniform mesh. With the nodal points $x_i \overset{\text{def}}{=} i/(N+1)$, these are given, for $1 \leq i \leq N$, by

$$b_i(x) = \frac{1}{N+1}\begin{cases} x - x_{i-1} & \text{for } x \in [x_{i-1}, x_i), \\ x_{i+1} - x & \text{for } x \in [x_i, x_{i+1}), \\ 0 & \text{otherwise.} \end{cases}$$

We seek approximations $\mathbf{u}$ of the solution of the diffusion equation (1.2) as linear combinations of these basis functions, i.e.,

$$\mathbf{u}(x,t) = \sum_{i=1}^{N} u_i(t) b_i(x).$$

For fixed t we denote the coefficient vector $\{u_i(t)\}$ simply by $u(t)$. An N-dimensional linear system of the form (2.2) is obtained for u from the standard Galerkin method by requiring that $\mathbf{u}(x,0) \equiv 0$ and

$$\int_0^1 \dot{\mathbf{u}}(x,t) b_i(x)\, dx = \int_0^1 \left(-q(x) \frac{\partial \mathbf{u}}{\partial x}(x,t) \frac{db_i}{dx}(x) \; + \; f(x,t) b_i(x) \right) dx.$$

In our implementation, the integrals on the left-hand side of this equation are encoded exactly, and the ones on the right-hand side are approximated using the trapezoid rule.

We report results below for six numerical experiments. In each of these, $N = 3$. Thus H is the subspace of $L^2(0,1)$ consisting of continuous functions on $[0,1]$ that equal zero at the endpoints and are piecewise linear with interior nodes at $x_1 = \frac{1}{4}$, $x_2 = \frac{1}{2}$ and $x_3 = \frac{3}{4}$. One of the experiments is an example of "continuous time observations," and the rest are based on "discrete-time observations." In each of these cases, the spatial observation operator C, as in (3.27), is the same. It is defined as pointwise evaluation of functions in H at the points $x = \frac{1}{3}$ and $x = \frac{2}{3}$.

The function $q(x)$ and the perturbations $p(x)$ are represented as linear splines on the mesh ($N = 3$) described above. Perturbations p are elements of H; specifically, they are zero at the endpoints and can vary at any of the $x_i, i = 1,2,3$.

Thus the q-dependent linear evolution equation is given by $\mathbf{M}\dot{\mathbf{u}} = \mathbf{A}(\mathbf{q})\mathbf{u} + \mathbf{f}$, where

$$\mathbf{M} = \frac{1}{6}\begin{pmatrix} 4 & 1 & 0 \\ 1 & 4 & 1 \\ 0 & 1 & 4 \end{pmatrix},$$

$$\mathbf{A}(\mathbf{q}) = 8\begin{pmatrix} -(q_0 + 2q_1 + q_2) & (q_1 + q_2) & 0 \\ (q_1 + q_2) & -(q_1 + 2q_2 + q_3) & (q_2 + q_3) \\ 0 & (q_2 + q_3) & -(q_2 + 2q_3 + q_4) \end{pmatrix},$$

and $\mathbf{f}(t) = \{f(x_i,t)\}_{i=1}^3$.

The "data" z are generated by carrying out the approximations as outlined above, with

$$q(x) = 1 + \tfrac{1}{5} e^{-20(x-1/3)^2}$$

approximated by its linear spline interpolate. The source term $f(x,t)$ is chosen so that the solution of $u(x,t)$ of (1.2) is $u(x,t) = x(1-x)te^{-t}$. Then the directional derivatives (2.3) and their approximations (3.4) are investigated based on $q(x) \equiv 1$.

Three of the experiments are based on nonuniform meshes $\mathbf{t}^n$. To explain how these meshes are generated, it is sufficient that we explain how the vectors $\{h_k\}$ are generated. To highlight the dependence on n here we denote these vectors by $\mathbf{h}^{(n)}$, and the individual h-values by $h_k^{(n)}$. To generate nonuniform $\mathbf{h}^{(n)}$ we begin with a given n and create a uniform mesh by setting $h_k^{(n)} \equiv t_F/n$. Then with a given "mesh ratio" r satisfying $0 < r < 1$, the "mesh refinement scheme" for generating $\mathbf{h}^{(2n)}$ from $\mathbf{h}^{(n)}$ is

$$\begin{aligned} h_{2k-1}^{(2n)} &= (1-r)h_k^{(n)}, \\ h_{2k}^{(2n)} &= rh_k^{(n)} \end{aligned}$$

for $k = 1, \ldots, n$. The choice $r = \frac{1}{2}$ yields refined meshes which are each uniform. This is clearly not true for any other choice of r.

We examine the convergence properties of costate approximations using four different time-marching schemes in (3.2). For the state equation (2.2), these methods are as follows.

CN. The trapezoidal, or "Crank–Nicholson," method

$$u_{k+1} = (I - Ah_k/2)^{-1}\Big((I + Ah_k/2)u_k + h^k(f_k + f_{k+1})/2 \Big).$$

RK2e. The explicit second-order Runge–Kutta method

$$u_{k+1} = (I + Ah_k + (Ah_k)^2/2)u_k + h_k((I + Ah_k)f_k + f_{k+1})/2.$$

RK2i. The implicit second-order Runge–Kutta method

$$u_{k+1} = (I - Ah_k + (Ah_k)^2/2)^{-1}\Big(u_k + h_k((I - Ah_k)f_{k+1} + f_k)/2 \Big).$$

RK4. The explicit fourth-order Runge–Kutta method

$$u_{k+1} = u_k + (K_1 + 2K_2 + 2K_3 + k_4)/6,$$

where

$$\begin{aligned} K_1 &= h_k(Au_k + f_k), \\ K_2 &= h_k(A(u_k + K_1/2) + f_{k+1/2}), \\ K_3 &= h_k(A(u_k + K_2/2) + f_{k+1/2}), \\ K_4 &= h_k(A(u_k + K_3) + f_{k+1}). \end{aligned}$$

To express the RK4 method in a form compatible with Theorem 6(g), we first obtain a fourth-order accurate approximation of $f_{k+1/2}$. This we do by evaluating at $(t_k^n + t_{k+1}^n)/2$ the cubic polynomial that interpolates (t_{k+i}^n, f_{k+i}), $-1 \le i \le 2$. For a uniform mesh, this yields

$$\begin{aligned} B &= I + Ah + (Ah)^2/2 + (Ah)^3/6 + (Ah)^4/24, \\ [\mathcal{R}^n P_n f]_k &= \frac{1}{6}\Big(-C_1 f_{k-1} + (C_2 + 9C_1)f_k + (I + 9C_1)f_{k+1} - C_1 f_{k+2} \Big), \\ C_1 &\overset{\text{def}}{=} \frac{1}{4}(I + Ah/2 + (Ah)^2/8), \\ C_2 &\overset{\text{def}}{=} I + Ah + (Ah)^2/2 + (Ah)^3/4. \end{aligned}$$

We carried out a rather extensive numerical investigation, using the MATLAB software package [9] on a Sun Sparcstation 2 and a DECstation 5000/200. In all cases, we compare the costate approximation $\widetilde{\delta_p T^n}$ to centered-difference approximations $(\delta_p T^n)_{FD}$. These are computed by evaluation of the expression under the limit sign in (2.3) with T replaced by T^n, and with $\tau = \sqrt{\epsilon}$. The value of ϵ (called "*eps*" in MATLAB) is the machine-dependent constant defined as the smallest integer-power of 2 for which "$1+\epsilon > 1$" is true. On the Sparcstation 2, $\epsilon = 2^{-52} \approx 2\times 10^{-16}$, so that numbers are represented up to sixteen decimal places. The choice of $\tau = \sqrt{\epsilon}$ in the finite difference computation reflects a desire to strike a balance between discretization error and roundoff error. By heuristics in [8, pp. 31–32], we expect that $(\delta_p T^n)_{FD}$ thus computed should agree with the correct value of $\delta_p T^n$ up to roughly half of the sixteen decimal places.

By the result of Theorem 3, the true gradient of the discrete functional $\delta_p T^n$ converges with full accuracy to $\delta_p T(q)$ (under the assumptions (A5′) and (A6′), of course). Since $(\delta_p T^n)_{FD}$ equals $\delta_p T^n$ to within eight digits or so, the triangle inequality implies that the rate of convergence of $\widetilde{\delta_p T^n}$ to $\delta_p T$ is correctly indicated by the rate of convergence of $\widetilde{\delta_p T^n}$ to $(\delta_p T^n)_{FD}$ as long as the relative difference of these two quantities is larger than about 10^{-8}.

For each of the six experiments, then, we tabulate the relative error

$$E_{|\mathbf{h}|} = \max_{1\le i\le m} \left| \frac{\widetilde{\delta_p T^n} - (\delta_p T^n)_{FD}}{(\delta_p T^n)_{FD}} \right| \tag{4.1}$$

for the four methods as $|\mathbf{h}|$ varies. From these results we then estimate the actual rate of convergence $\bar{\nu}$ based on the assumption that $E_{|\mathbf{h}|} = C|\mathbf{h}|^{\bar{\nu}}$.

Case 1. Observations are continuous in time, and the meshes are non-uniform with refinement ratio $r = 1/3$. The trapezoidal quadrature rule is used for the $\mathcal{H}^n$ inner product for the three second-order methods, and Simpson's rule is used for the RK4 method. See Table 1.

TABLE 1

$\lvert\mathbf{h}\rvert$	CN	RK2e	RK2i	RK4
6.67e-3	1.6e-6	2.6e-4	2.0e-4	2.9e-6
4.44e-3	5.2e-7	7.8e-5	6.8e-5	2.1e-7
$\bar{\nu}$	2.7	2.7	2.7	6.4

Case 2. This experiment is motivated by Theorem 5. We take observations discrete in time on uniform meshes. The quadrature weights, as in (3.6), are those arising from the trapezoidal rule. The results are presented in the Table 2. The three Runge–Kutta methods appear to be converging at a second-order rate, as predicted by Theorem 5. Crank–Nicholson costate approximations agree with the forward-

TABLE 2

$\lvert\mathbf{h}\rvert$	CN	RK2e	RK2i	RK4
1.0e-2	8.8e-10	8.7e-5	6.9e-5	3.9e-6
0.5e-2	1.3e-9	2.0e-5	1.8e-5	9.1e-7
$\bar{\nu}$	*	2.1	1.9	2.1

difference approximations to within nine decimal places—roughly the expected accuracy of $(\delta_p T^n)_{FD}$. In fact, for uniform meshes, $\widetilde{\delta_p T^n}$ as computed by the Crank–Nicholson method *exactly* equals $\delta_p T^n$ (neglecting roundoff error). We outline a proof of this assertion. The recursion formula (3.12) for $u^n = \mathcal{S}^n \mathcal{R}^n P_n$ takes the form

$$u^n_{k+1} = B(q) u^n_k + hR(q)(f_k + f_{k+1}) \tag{4.2}$$

with

$$R(q) = \frac{1}{2}\left(I - \frac{h}{2}A\right)^{-1},$$
$$B(q) = 2R(q)\left(I + \frac{h}{2}A\right).$$

Then,

$$\delta_p R(q) = hR(q)\big(\delta_p A(q)\big)R(q),$$
$$\delta_p B(q) = hR(q)\big(\delta_p A(q)\big)\big(B(q) + I\big).$$

Using these in (3.19) leads to

$$\delta_p u^n_{k+1} = B(q)\delta_p u^n_k + hR(q)\big(\delta_p A(q)\big)(u^n_k + u^n_{k+1}).$$

Referring to (3.22) and (4.2) we see that this last equation is just the recursion formula for $\delta_p u^n = \mathcal{S}^n \mathcal{R}^n \phi^n$. Thus

$$\delta_p T^n(q) = \langle\!\langle\, r^n, \mathcal{C}^n \mathcal{S}^n \mathcal{R}^n \phi^n \,\rangle\!\rangle_{\mathcal{Z}}$$

However, by reasoning similar to that which led to (3.43), the right-hand side of this equation can be rewritten as

$$\langle\!\langle\, r^n, \mathcal{C}^n \mathcal{S}^n \mathcal{R}^n \phi^n \,\rangle\!\rangle_{\mathcal{Z}} = \langle\!\langle\, J^n \mathcal{S}^n_{(*)} \mathcal{R}^n_{(*)} J^n (\mathcal{C}^*)^n r^n,\ \phi^n \,\rangle\!\rangle_{\mathcal{H}^n},$$

which by (3.4), (3.22), and (3.26) is the same as $\widetilde{\delta_p T^n}$. Thus $\widetilde{\delta_p T^n} = \delta_p T^n$, so that the numerator in (4.1) is nonzero only to the extent that $(\delta_p T^n)_{FD} \neq \delta_p T^n$. This explains the Crank–Nicholson column in Table 2.

Case 3. This example is the same as Case 2 but with nonuniform meshes, generated with $r = 1/3$. In this case we have no theoretical basis on which to expect convergence, and indeed Table 3 suggests that convergence is not obtained.

TABLE 3

$\lvert\mathbf{h}\rvert$	CN	RK2e	RK2i	RK4
6.7e-3	1.6e-3	2.7e-2	1.9e-2	9.8e-3
4.4e-3	1.6e-3	2.4e-2	2.0e-2	4.2e-1
$\bar{\nu}$	0	0.3	-0.1	-9.6

Case 4. We conjecture that in some cases, even though the meshes $\mathbf{t}^n$ may be nonuniform, high-order convergence can still be achieved provided that the meshes are *"locally uniform"* in the sense that for a given n, all of the h_k's "near" a given observation point τ_i are constant. More precisely, there is a positive integer σ and a

TABLE 4

$\|h\|$	CN	RK2e	RK2i	RK4
6.7e-3	5.8e-5	6.2e-3	2.2e-3	9.8e-3
4.4e-3	2.5e-5	2.1e-3	1.2e-3	4.2e-1
$\bar{\nu}$	2.0	2.7	1.5	0.6

real $\tilde{h}$ for which $t_{k+1}^n - t_k^n = \tilde{h}$ whenever $(\kappa(i) - \sigma) \leq k \leq (\kappa(i) + \sigma)$, where $\kappa(i)$ is from (3.30).

In this example we repeat the experiment reported under Case 3, but took steps to ensure that the refined meshes were "locally uniform" with $\sigma = 1$.

Table 4 suggests that high-order convergence can sometimes be obtained on "locally uniform" meshes. However, we do not pursue this idea further.

Case 5. Table 5 presents an example illustrating Theorem 6. The forcing term f in the state equation is changed so that $f(0) = 0$. The RK4 method is used together with the quadrature weight matrix $\mathcal{W}_{\text{Trap}}$ arising from the trapezoidal quadrature rule. These weights appear in Theorem 6(d). The mesh refinement is uniform.

TABLE 5

$\|h\|$	RK4
1.0e-2	6.5e-8
0.5e-2	3.4e-9
$\bar{\nu}$	4.3

Case 6. Table 6 illustrates Corollary 2. We repeat experiment shown in Case 5 but using Simpson's rule instead of the trapezoidal rule, and we compare $\widetilde{\delta_p T^n}$ to $(3/2)(\delta_p T^n)_{FD}$.

TABLE 6

$\|h\|$	RK4
1.0e-2	3.1e-3
0.5e-2	6.7e-4
$\bar{\nu}$	2.2

5. Alternate approaches. Throughout the preceding sections we have discussed the use of the "discretized costate approximation" and have attempted to illustrate the delicacy of that procedure when pointwise (in time) observations are involved. There are at least three other possible approaches that also merit consideration when pointwise observations are involved, which we now briefly discuss.

The strategy pursued in §§2 and 3 was to take the *discretization of the adjoint* system; that is, to first derive the costate method for the continuous problem, and then to discretize the resulting equations and integrals. An alternative to this would be to use the *adjoint of the discrete* system. While this approach is straightforward when applied to boundary value problems, certain complications may arise in the implementation for evolution equations, particularly when time marching schemes are used for temporal discretization.

In particular, from the definition (3.1) of the discrete least-squares functional we have

$$\delta_p T^n(q) = \tfrac{1}{2}\langle\!\langle\, r^n, \mathcal{C}^n \delta_p u^n \,\rangle\!\rangle_{\mathcal{Z}^n}.$$

From the expression for components of $\delta_p u^n$ in (3.19) and the definition (3.11) of $\mathcal{S}^n(q)$, we see that

$$\delta_p u^n(q) = \mathcal{S}^n(q)\gamma^n(q;p),$$

where $\gamma^n(q;p)$ is given by

$$\gamma_k^n(q;p) \overset{\text{def}}{=} [\,\delta_p \mathcal{R}^n(q)\, P_n f]_k + \frac{1}{h_k}(\delta_p B_k) u_k^n. \tag{5.1}$$

Thus

$$\delta_p T^n(q) = \tfrac{1}{2}\langle\!\langle\, (\mathcal{S}^n)^*(\mathcal{C}^n)^* r^n, \gamma^n(q;p) \,\rangle\!\rangle_{\mathcal{Z}^n}.$$

The advantage of this approach is that it will yield *exact* values for the directional derivatives of T^n. Hence in numerical attempts to minimize T^n for any *fixed* n it would be ideal. However, it involves the use of $\gamma^n(q;p)$ as given in (5.1), which may require significant amounts of additional mathematical effort, code complexity, and computational execution time. These disadvantages may be negated in whole or in part for certain special cases if the expression for $\gamma^n(q;p)$ can be simplified. Such simplification occurred with the Crank–Nicholson method for uniform time meshes, as described in Case 2 of §4. However, we do not pursue this approach further here.

Another alternative [3, §V, (5.30)–(5.32)] is to retain the strategy of discretizing the continuous adjoint system, but to *transform* the costate equation (2.9) so as to increase the smoothness of the source term. This can be done by the introduction of two new variables ψ and ξ given by

$$\psi(t) \overset{\text{def}}{=} \int_0^t \big[J(\mathcal{C}^* r)\big](s)\, ds$$

and

$$\xi(t) \overset{\text{def}}{=} y(t) - \psi(t).$$

Then, ξ satisfies

$$\begin{aligned}\dot{\xi}(t) &= A^*(q)\xi(t) + A^*(q)\psi(t),\\ \xi(0) &= 0.\end{aligned}$$

The idea is that since ψ is piecewise constant (as opposed to $J(\mathcal{C}^*)r$, which is a linear combination of Dirac delta functions), it should be easier to obtain accurate numerical solutions of y by first approximating ξ as given here.

We briefly explored this idea numerically. We repeated the relevant numerical examples, which are Case 2 and Case 3 of §4. In both cases, all four of the TMSs were used to approximate ξ instead of y, and then we set $y^n = \xi^n + \psi$. In each of these runs, we observed *linear* convergence of the directional derivatives; e.g., $\bar{\nu}$ was approximately 1 in each case. Obviously, for uniform time meshes (as in Case 2) this represents a loss of accuracy, while for nonuniform meshes (Case 3) it is a definite improvement.

One other alternative is to consider modifying step (iii) of the discretized costate approximation along the following lines. Instead of attempting to integrate $\dot{y}$ over the whole interval $(0, t_F)$ in "one shot" with the TMS, we should treat each subinterval (τ_{i-1}, τ_i) separately, using the fact that $J^n(\mathcal{C}^*)^n r^n$ is zero in the interior of these subintervals and that y has jump discontinuities at τ_i. Correspondingly, the quadrature in (3.4) must be modified. Although the implementation is more complicated, analysis along the lines of the proof of Corollary 1 above then applies. Consequently, for sufficiently smooth state-variable forcing functions f, full accuracy of the TMS and the quadrature will yield full accuracy of the costate approximations.

6. Summary and conclusions. While the costate method offers potentially dramatic time savings when computing gradients in parameter estimation problems for time-dependent systems, we have indicated that care must be taken when using this approach. We have shown both analytically and numerically that the use of convergent time-marching and quadrature schemes in seemingly reasonable combinations can yield surprisingly poor results.

The difficulty lies in the fact that the forcing term of the costate equation may be very non-smooth, so that the assumptions under which the TMS and quadrature scheme converge are violated. In the description of the discretized costate approximation in §3.1.1, we assumed that the TMS is used in step (iii) as a "black box" without regard to the smoothness (or lack thereof) of $J^n(\mathcal{C}^*)^n r^n$; see (3.2).

The costate implementation that we have analyzed involves discretization of the (continuous) costate equation using standard time-marching schemes. Due to its simplicity, this approach is very desirable *when the resulting gradient approximation has the same order of accuracy as the particular time-marching scheme.* We presented conditions that guarantee this. Some of these conditions are straightforward. For instance, several popular second-order time-marching schemes yield second-order gradient approximations *provided* a uniform time mesh is used in combination with (second-order accurate) trapazoidal quadrature. Other conditions are less obvious (cf. Theorem 6). We also demonstrate that certain "reasonable" implementations yield gradient approximations that fail to converge at all, e.g., fourth-order Runge–Kutta time-marching on a uniform time mesh combined with (fourth-order) trapazoidal quadrature. Finally, several alternative (though less simple) costate implementations were presented.

REFERENCES

[1] H. T. Banks, *Computational issues in parameter estimation and feedback control problems for partial differential equations*, Physica D, 60 (1992), pp. 226–238.

[2] H. T. Banks, J. M. Crowley, and I. G. Rosen, *Methods for the identification of material parameters in distributed models for flexible structures*, Mat. Apl. Comput., 5 (1986), pp. 139–168.

[3] H. T. Banks and K. Kunisch, *Estimation Techniques for Distributed Parameter Systems*, Birkhäuser, Boston, MA, 1989.

[4] H. T. Banks and I. G. Rosen, *Numerical schemes for the estimation of functional parameters in distributed models for mixing mechanisms in lake and sea sediment cores*, Inverse Problems, 3 (1987), pp. 1–23.

[5] H. T. Banks, I. G. Rosen, and K. Ito, *A spline-based technique for computing Riccati operator and feedback controls in regulator problems for delay equations*, SIAM J. Sci. Stat. Comput., 5 (1984), pp. 830–855.

[6] J. A. Burns, K. Ito, and G. Propst, *On nonconvergence of adjoint semigroups for control systems with delays*, SIAM J. Control Optim., 26 (1988), pp. 1442–1454.

[7] G. Chavent and P. Lemonnier, *Identification de la non-linearité d'une équation parabolique quasilinéaire*, Applied Math. Optim., 1 (1974), pp. 121–162.

[8] J. E. Dennis and R.. B. Schnabel, *Numerical methods for unconstrained optimization and nonlinear equations*, Prentice Hall, Englewood Cliffs, 1983.

[9] MATLAB, The MathWorks, Inc., Cochituate Place, 24 Prime Park Way, South Natick, MA.

[10] R. D. Richtmyer and K. W. Morton, *Difference Methods for Initial Value Problems*, John Wiley, New York, 1967.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 33, No. 1, pp. 255–273, January 1995

©1995 Society for Industrial and Applied Mathematics
012

BOUNDARY STABILIZATION FOR THE VON KÁRMÁN EQUATIONS *

JEAN-PIERRE PUEL† AND MARIUS TUCSNAK†

Abstract. The boundary stabilization of a nonlinear plate model is studied. The equations used take in consideration the in-plane accelerations and the rotary inertia of the cross sections. Applying linear feedbacks, the authors obtain the exponential decay of the energy.

Key words. semilinear evolution equations, feedback boundary control, Lyapounov functionals

AMS subject classifications. 93D15, 35B40, 73K50

1. Introduction.

We consider the stabilization problem for the coupled in-plane and transversal vibrations of a nonlinear plate. As far as we know, all previous work on nonlinear plate vibrations is based on models neglecting the in-plane accelerations (e.g. [1], [3, Chap. 5], [6] for boundary stabilization problems). This assumption leads, for star-shaped domains and suitable boundary conditions, to the existence of an Airy stress function, which considerably simplifies the equations.

The goal of this paper is to investigate the boundary stabilization of the complete set of von Kármán equations, as they are given, for example, in [3, p. 19]. This model can be applied for domains that are not star-shaped, but it has the inconvenience that the nonlinear terms are not well defined in the energy space. This feature makes the proof of global existence results more delicate. To our knowledge, the only mathematical result concerning this plate model are the global existence theorems proved in [11, Part 1] (for weak solutions) and [10] (for strong solutions).

For the one-dimensional corresponding rod problem, the uniform stabilization was studied in [7], which inspired our work.

The paper is organized as follows. In §2 we formulate the initial and boundary value problem and we discuss the feedback boundary controls. For the sake of completeness, in §3, we outline the proof of the main global existence theorem (the complete proof is given in [10]). Section 4 contains the stabilization results. In §5 we give an energy identity, fundamental for the proof of the stabilization theorem, which is given in §6. Our method is based on the fact that the system we consider is obtained by coupling in a nonlinear way the equations of two-dimensional (2D) linear elasticity and of a linear Kirchhoff plate. This is why we apply the technique proposed in [2], [3, Chap. 4], [4], and [7], which is based on the construction of appropriate Lyapunov functionals. The desired differential inequalities are obtained as a consequence of the energy identity proven in §5, by the use of multiplier techniques. The ideas and computations we used are similar to the one contained in [5].

The results contained in this paper were announced in [9].

*Received by the editors March 26, 1992; accepted for publication (in revised form) September 10, 1993.

†Ecole Polytechnique, Centre de Mathématiques Appliquées, 91128 Palaiseau Cedex, France, and Université de Versailles.

2. Statement of the problem. Let $\Omega \subset \mathbf{R}^2$ be a bounded regular domain, $Q = \Omega \times (0, \infty)$, $\Gamma = \partial\Omega$, $\Gamma_0 \subset \Gamma$, $\Gamma_1 = \Gamma - \Gamma_0$, $\Sigma_i = \Gamma_i \times (0, \infty)$, $i = 0, 1$. We suppose that $\Gamma_0 \cap \Gamma_1 = \emptyset$. Let $\mathcal{S}$ be the set of all fourth-order symmetric tensors on $\mathbf{R}^2$ and $\mathcal{C} \in \mathcal{S}$ defined by

$$\mathcal{C}[\epsilon] = \frac{E}{d(1-\mu^2)}[\mu(tr\epsilon)I_{\mathcal{S}} + (1-\mu)\epsilon] \quad \forall \epsilon \in \mathcal{S},$$

where $I_{\mathcal{S}}$ is the identity of S, d is the density, E is Young's modulus, and μ is Poisson's ratio of the material. If $E > 0$ and $0 < \mu < \frac{1}{2}$ the tensor $\mathcal{C}$ satisfies the following condition:

$$\mathcal{C}[\epsilon] \cdot [\epsilon] \geq \lambda_0 |\epsilon|^2 \quad \forall \epsilon \in \mathcal{S}, \quad \text{where } \lambda_0 > 0. \tag{2.1}$$

We consider $f : \mathbf{R^2} \rightarrow \mathcal{S}$, defined by $f(x) = \frac{1}{2} x \otimes x$. With this notation the dynamic von Kármán equations can be written in the following form:

$$u'' - \operatorname{div}\{\mathcal{C}[\epsilon(u) + f(\nabla w)]\} = 0 \quad \text{in Q}, \tag{2.2}$$

$$w'' - \gamma\Delta w'' + D\Delta^2 w - div\{\mathcal{C}[\epsilon(u) + f(\nabla w)]\nabla w\} = 0 \quad \text{in Q}, \tag{2.3}$$

$$u = 0, \qquad w = \frac{\partial w}{\partial \nu} = 0 \quad \text{on } \Sigma_0, \tag{2.4}$$

$$\mathcal{C}[\epsilon(u) + f(\nabla w)]\nu = g \quad \text{on } \Sigma_1, \tag{2.5}$$

$$D[\Delta w + (1-\mu)B_1 w] = -M_s \quad \text{on } \Sigma_1, \tag{2.6}$$

(2.7)
$$D\left[\frac{\partial \Delta w}{\partial \nu} + (1-\mu)\frac{\partial B_2 w}{\partial s}\right] - \gamma \frac{\partial w''}{\partial \nu} - \mathcal{C}[\epsilon(u) + f(\nabla w)]\nu \cdot \nabla w = -\frac{\partial}{\partial s} M_\nu - h \quad \text{on } \Sigma_1,$$

$$u(0) = u^0, \quad u'(0) = u^1, \quad w(0) = w^0, \quad w'(0) = w^1, \quad \text{in } \Omega, \tag{2.8}$$

In the above equations $u = (u_1, u_2)$ is the plane displacement, $\gamma > 0$ is a constant, $\epsilon(u) = \frac{1}{2}[\nabla u + (\nabla u)^T]$, w is the transverse displacement, D represents the flexural rigidity of the plate, ν and s are the outward unit normal and tangent, respectively, to the boundary. The operators B_1, B_2 are given by

$$B_1 w = 2\nu_1\nu_2 \frac{\partial^2 w}{\partial x_1 \partial x_2} - \nu_1^2 \frac{\partial^2 w}{\partial x_2^2} - \nu_2^2 \frac{\partial^2 w}{\partial x_1^2},$$

$$B_2 w = (\nu_1^2 - \nu_2^2)\frac{\partial^2 w}{\partial x_1 \partial x_2} + \nu_1\nu_2 \left(\frac{\partial^2 w}{\partial x_2^2} - \frac{\partial^2 w}{\partial x_1^2}\right).$$

The quantities g, h, M_s, and M_ν are the boundary controls of the system. They correspond, respectively, to the tension in the plane of the plate, the effect of transverse shear force, the bending moment around the tangential vector to Γ, and the twisting moment about the normal to Γ.

Let $\{u, w\}$ be a classical solution of (2.2)–(2.8). The total energy of the plate is given (cf. [3, p. 18]) by

$$E(t) = \tfrac{1}{2}\{||u'(t)||^2 + ||w'(t)||^2 + \gamma||\nabla w'(t)||^2 + a(w(t), w(t)) \\ + (\mathcal{C}[\epsilon(u(t)) + f(\nabla w(t))], \epsilon(u(t)) + f(\nabla w(t)))\},$$

where $||\cdot||$ and $(\cdot,\cdot)$ represent the norm, respectively, the inner product, in $[L^2(\Omega)]^k$, $k \in \mathbf{N}$ and

$$\text{(2.9)} \qquad a(w,\tilde{w}) = D\int_\Omega \left[\frac{\partial^2 w}{\partial x_1^2}\frac{\partial^2 \tilde{w}}{\partial x_1^2} + \frac{\partial^2 w}{\partial x_2^2}\frac{\partial^2 \tilde{w}}{\partial x_2^2} + \mu\left(\frac{\partial^2 w}{\partial x_1^2}\frac{\partial^2 \tilde{w}}{\partial x_2^2} \right.\right. \\ \left.\left. + \frac{\partial^2 \tilde{w}}{\partial x_1^2}\frac{\partial^2 w}{\partial x_2^2}\right) + 2(1-\mu)\frac{\partial^2 w}{\partial x_1\partial x_2}\frac{\partial^2 \tilde{w}}{\partial x_1\partial x_2}\right]dx.$$

A simple calculation, based on the integration by parts formula,

$$\text{(2.10)} \quad D\int_\Omega (\Delta^2 w)\tilde{w}\,dX = a(w,\tilde{w}) - D\int_\Gamma \left\{[\Delta w + (1-\mu)B_1 w]\frac{\partial \tilde{w}}{\partial \nu} \right. \\ \left. - \left[\frac{\partial \Delta w}{\partial \nu} + (1-\mu)\frac{\partial B_2 w}{\partial s}\right]\tilde{w}\right\}d\Gamma,$$

shows that

$$\text{(2.11)} \qquad E'(t) = \int_{\Gamma_1}\left[g\cdot u' - M_s\frac{\partial w'}{\partial \nu} + \left(\frac{\partial}{\partial s}M_\nu + h\right)w'\right]d\Gamma.$$

Let $x_0 \in \mathbf{R}^2$ and $m(x) = x - x_0$. We suppose that Γ_0, Γ_1 have the property

$$\text{(2.12)} \qquad m(x)\cdot\nu(x) < 0, \quad \text{if } x \in \Gamma_0; \qquad m(x)\cdot\nu(x) \geq 0, \quad \text{if } x \in \Gamma_1.$$

If we put

$$\text{(2.13)} \qquad g = -c_1(m\cdot\nu)u', \qquad M_s = c_1(m\cdot\nu)\frac{\partial w'}{\partial \nu}, \\ \frac{\partial}{\partial s}M_\nu + h = -c_1(m\cdot\nu)w' + c_1\frac{\partial}{\partial s}\left[(m\cdot\nu)\frac{\partial w'}{\partial s}\right],$$

with $c_1 > 0$ we get

$$\text{(2.14)} \qquad E'(t) = -c_1\int_{\Gamma_1}(m\cdot\nu)(u'^2 + w'^2 + |\nabla w'|^2)d\Gamma \leq 0,$$

i.e., the total energy is nonincreasing. However, this type of feedback law does not seem to give the exponential decay of the energy, even in linear plane elasticity (see [4] for the details). This is why, following [4], we set

$$g = -c_1(m\cdot\nu)u' - c_2\left(\frac{\partial u_2}{\partial s}, -\frac{\partial u_1}{\partial s}\right), \text{ where } c_2 > 0,$$

$$M_s = c_1(m\cdot\nu)\frac{\partial w'}{\partial\nu}, \frac{\partial}{\partial s}M_\nu + h = -c_1(m\cdot\nu)w' + c_1\frac{\partial}{\partial s}\left[(m\cdot\nu)\frac{\partial w'}{\partial s}\right]. \tag{2.15}$$

The advantage of this feedback law is that, in our calculations, the classical energy density, $W(u) = \frac{1}{2}\mathcal{C}[\epsilon(u)]\cdot\epsilon(u)$ will be replaced by

$$\widetilde{W}(u) = W(u) + c_2\left(\frac{\partial u_1}{\partial x_1}\frac{\partial u_2}{\partial x_2} - \frac{\partial u_1}{\partial x_2}\frac{\partial u_2}{\partial x_1}\right),$$

which, for $c_2 > 0$, c_2 small, satisfies the condition

$$\widetilde{W}(u) \geq c_0|\nabla u|^2, \quad \text{with } c_0 > 0. \tag{2.16}$$

Define

$$\widetilde{E}(t) = E(t) + c_2\int_\Omega\left(\frac{\partial u_1}{\partial x_1}\frac{\partial u_2}{\partial x_2} - \frac{\partial u_1}{\partial x_2}\frac{\partial u_2}{\partial x_1}\right)dXdt. \tag{2.17}$$

Using the formula

$$\begin{aligned}&\int_\Gamma\left(\frac{\partial u_2}{\partial s}\phi_1 - \frac{\partial u_1}{\partial s}\phi_2\right)d\Gamma \\ &= \int_\Omega\left(\frac{\partial u_1}{\partial x_1}\frac{\partial \phi_2}{\partial x_2} + \frac{\partial u_2}{\partial x_2}\frac{\partial \phi_1}{\partial x_1} - \frac{\partial u_1}{\partial x_2}\frac{\partial \phi_2}{\partial x_1} - \frac{\partial u_2}{\partial x_1}\frac{\partial \phi_1}{\partial x_2}\right)dX,\end{aligned} \tag{2.18}$$

we get, with the feedback law (2.15),

$$\widetilde{E}'(t) = -c_1\int_{\Gamma_1}(m\cdot\nu)(u'^2 + w'^2 + |\nabla w'|^2)d\Gamma \leq 0. \tag{2.19}$$

It will be proved in §6 that the feedback law (2.15) uniformly stabilizes the system (2.2)–(2.8), in the sense that $\widetilde{E}(t) \to 0$ as $t \to \infty$ in an exponential way, on each bounded set $E(0) \leq M$ of initial data. On the other hand, it is easy to check that, for c_2 small enough, $\widetilde{E}(t)$ is equivalent to $E(t)$, in the sense that

$$K^{-1}E(t) \leq \widetilde{E}(t) \leq KE(t) \quad \forall t \geq 0, \tag{2.20}$$

for a suitable constant $K > 0$. In this way, the uniform exponential decay of the original energy functional will be established.

3. Existence of solutions. In this section we outline the proof of a global existence result for the problem (2.2)–(2.8). The complete proof is given in [10]. We begin by introducing the following function spaces:

$$W = \{w \in H^2(\Omega)|\ w_{|\Gamma_0} = \frac{\partial w}{\partial\nu}|\Gamma_0 = 0\};\ V = \{w \in H^1(\Omega)|\ w_{|\Gamma_0} = 0\};\ H = L^2(\Omega).$$

$$U = \{u \in [H^1(\Omega)]^2|\ u_{|\Gamma_0} = 0\}, \mathcal{K} = [L^2(\Omega)]^2.$$

Let $\{u, w\}$ be a classical solution of (2.2)–(2.8) with the feedback law (2.15), i.e.,

$$u \in C(0,T;[H^2(\Omega)]^2) \cap C^1(0,T;[H^1(\Omega)]^2), \qquad w \in C(0,T;H^4(\Omega) \cap C^1(0,T;H^3(\Omega)).$$

If we take the inner product in $[L^2(\Omega)]^2$ of (2.2) with $\phi \in U$ and of (2.3) with $\psi \in W$, by the use of (2.6)–(2.8) and(2.15) we get

$$(3.1)\qquad [(u',\phi)]' + (\mathcal{C}[\epsilon(u) + f(\nabla w)], \epsilon(\phi)) + c_2 \int_\Omega \left(\frac{\partial u_1}{\partial x_1}\frac{\partial \phi_2}{\partial x_2} + \frac{\partial u_2}{\partial x_2}\frac{\partial \phi_1}{\partial x_1} - \frac{\partial u_1}{\partial x_2}\frac{\partial \phi_2}{\partial x_1} - \frac{\partial u_2}{\partial x_1}\frac{\partial \phi_1}{\partial x_2} \right) dX + c_1 \left[\int_{\Gamma_1} (m\cdot\nu) u \cdot \phi d\Gamma \right]' = 0, \quad \forall \phi \in U,$$

$$(3.2)\qquad [(w',\psi) + \gamma(\nabla w', \nabla\psi)]' + a(w,\psi) + (\mathcal{C}[\epsilon(u) + f(\nabla w)]\nabla w, \nabla \psi) + c_1 \left[\int_{\Gamma_1} (m\cdot\nu)(w\psi + \nabla w \cdot \nabla\psi) d\Gamma \right]' = 0 \quad \forall \psi \in W.$$

DEFINITION 3.1. *We say that* $\{u, w\}$ *is a* strong *solution of* (2.2)–(2.8) *with the feedback law* (2.15) *if*

$$u \in C(0,T;[H^2(\Omega)]^2 \cap U) \cap C^1(0,T;U), \qquad w \in C(0,T;H^3(\Omega) \cap W) \cap C^1(0,T;W),$$

and they satisfy (3.1), (3.2), *and* (2.8).

DEFINITION 3.2. *We call* $\{u, w\}$ *a* weak *solution of* (2.2)–(2.8) *with the feedback law* (2.15) *if*

$$u \in C(0,T;U) \cap C^1(0,T;\mathcal{K}), \qquad w \in C(0,T;W) \cap C^1(0,T;V),$$

and they satisfy (3.1), (3.2), *and* (2.8).

The following theorem is the main result of this section.

THEOREM 3.1. *Suppose that* c_2 *is small enough and that* u^0, u^1, w^0, w^1 *satisfy the conditions*

$$(3.3)\qquad u^0 \in [H^2(\Omega)]^2 \cap U, \qquad u^1 \in U,$$

$$(3.4)\qquad w^0 \in H^3(\Omega) \cap W, \qquad w^1 \in W,$$

$$(3.5)\qquad \mathcal{C}[\epsilon(u^0) + f(\nabla w^0)]\nu + c_1(m\cdot\nu)u^1 + c_2\left(\frac{\partial u_2^0}{\partial s}, -\frac{\partial u_1^0}{\partial s}\right) = 0 \quad on\ \Gamma_1,$$

$$(3.6)\qquad D[\Delta w^0 + (1-\mu)B_1 w^0] = -c_1(m\cdot\nu)\frac{\partial w^1}{\partial \nu} \quad on\ \Sigma_1.$$

Then there exists a unique global strong solution $\{u, w\}$ *of* (2.2)–(2.8).

Sketch of the proof.

Step 1. The first step consists of the following local existence result.

LEMMA 3.1. *Suppose that* c_2, u^0, u^1, w^0, w^1 *satisfy the assumptions of Theorem* 3.1. *Then there exists* $T_0 > 0$ *such that problem* (2.2)–(2.8) *admits an unique strong solution for any* $T < T_0$. *Moreover, one of the following assertions holds*:

(A1) $T_0 = \infty$;

(A2) $\lim_{t\to T_0}(||u(t)||_{H^2} + ||u'(t)||_{H^1} + ||w(t)||_{H^3} + ||w'(t)||_{H^2}) = \infty$.

Idea of the proof. We first prove some well-possedness results for the linearization of (2.2)–(2.8). More precisely, we consider the following initial value problem:

$$
\begin{aligned}
(3.7)\qquad & [(u',\phi)]' + (\mathcal{C}[\epsilon(u)],\epsilon(\phi)) \\
& + c_2 \int_\Omega \left(\frac{\partial u_1}{\partial x_1}\frac{\partial \phi_2}{\partial x_2} + \frac{\partial u_2}{\partial x_2}\frac{\partial \phi_1}{\partial x_1} - \frac{\partial u_1}{\partial x_2}\frac{\partial \phi_2}{\partial x_1} - \frac{\partial u_2}{\partial x_1}\frac{\partial \phi_1}{\partial x_2} \right) dX \\
& + c_1 \left[\int_{\Gamma_1} (m\cdot\nu) u \cdot \phi d\Gamma \right]' + (f_1(t),\epsilon(\phi)) = 0 \quad \forall \phi \in U,
\end{aligned}
$$

$$(3.8)\qquad u(0) = u^0, u'(0) = u^1 \quad \text{in } \Omega,$$

where

$$f_1(t) : \Omega \to \mathcal{S} \quad \forall t \in [0,T),$$

and

$$
\begin{aligned}
(3.9)\ & [(w',\psi) + \gamma(\nabla w',\nabla\psi)]' + a(w,\psi) \\
& + c_1 \left[\int_{\Gamma_1} (m\cdot\nu)(w\psi + \nabla w\cdot\nabla\psi) d\Gamma \right]' + (f_2(t),\nabla\psi) = 0 \quad \forall \psi \in W,
\end{aligned}
$$

$$(3.10)\qquad w(0) = w^0, \qquad w'(0) = w^1,$$

where

$$f_2(t) : \Omega \to \mathbf{R}^2 \quad \forall t \geq 0.$$

The key point is to construct the operator generating the semigroup associated to (3.9) (see also [5, §4.2.1]). We present here the method we used to overcome this difficulty. The following two lemmas, which are proved in [10], are very important for our approach.

LEMMA 3.2. *For any $w \in W$ there exists at most one couple $(y,h) \in V \times H^{\frac{1}{2}}(\Gamma_1)$ such that*

$$(3.11)\qquad a(w,\psi) = (y,\psi) + \gamma(\nabla y,\nabla\psi) + \int_{\Gamma_1} h\frac{\partial\psi}{\partial\nu} \quad \forall\psi \in W.$$

LEMMA 3.3. *Suppose that* (3.11) *holds for some $(w,y,h) \in W \times V \times H^{1/2}(\Gamma_1)$. Then we have the regularity property $w \in H^3(\Omega)$. Moreover, there exists a constant $C > 0$ such that*

$$(3.12)\qquad \|w\|_{H^3(\Omega)} \leq C(\|y\|_{H^1(\Omega)} + \|h\|_{H^{1/2}(\Gamma)}),$$

and the relation $D[\Delta w + (1-\mu)B_1 w] = h$, holds in $H^{1/2}(\Gamma_1)$.

Let us now introduce the space

$$
\begin{aligned}
(3.13)\quad \mathcal{D}(A) = \Big\{ & w \in W \text{ such that it exists } (y,h) \in V \times H^{1/2}(\Gamma_1), \\
& \text{satisfying } a(w,\psi) = (y,\psi) + \gamma(\nabla y,\nabla\psi) + \int_{\Gamma_1} h\frac{\partial\psi}{\partial\nu} d\Gamma,\ \forall\psi \in W \Big\}.
\end{aligned}
$$

By Lemma 3.3 we know that $\mathcal{D}(A) \subset H^3(\Omega) \cap W$. On the other hand, for w fixed, by Lemma 3.2 we know that y and h in (3.13) are unique. This allows us to introduce the operator

$$A : \mathcal{D}(A) \to V, \qquad Aw = y. \tag{3.14}$$

We also define the operator $B : W \to V$ by
(3.15)

$$(By, \psi) + \gamma(\nabla(By), \nabla\psi) = c_1 \int_{\Gamma_1} (m \cdot \nu) y\psi d\Gamma + c_1 \int_{\Gamma_1} (m \cdot \nu) \frac{\partial y}{\partial s}\frac{\partial \psi}{\partial s} d\Gamma \quad \forall \psi \in V,$$

where the second integral is taken in the sense of the duality $H^{1/2}(\Gamma_1)$, $H^{-1/2}(\Gamma_1)$. Now consider the space

$$\mathcal{D}(\mathcal{A}) = \left\{ \begin{pmatrix} w \\ y \end{pmatrix} \in \mathcal{D}(A) \times W, \text{ s. t.} \right.$$
$$\left. D[\Delta w + (1-\mu)B_1 w] = -c_1(m \cdot \nu)\frac{\partial y}{\partial \nu} \text{ on } \Gamma_1 \right\},$$

and define the operator

$$\mathcal{A}\begin{pmatrix} w \\ y \end{pmatrix} = \begin{pmatrix} -y \\ Aw + By \end{pmatrix} \quad \forall \begin{pmatrix} w \\ y \end{pmatrix} \in \mathcal{D}(\mathcal{A}). \tag{3.16}$$

By using Lemma 3.3 we can easily obtain (see [10]) that the operator $-\mathcal{A}$ defined by (3.16) generates a strongly continous semigroup in $W \times V$. The well-possedness result for (3.9), (3.10) is now obtained by noting that (3.9), (3.10) are equivalent to the following initial value problem in $W \times V$:

$$Z' + \mathcal{A}Z + F = 0,$$

$$Z(0) = Z^0,$$

where

$$Z = \begin{pmatrix} w \\ w' \end{pmatrix}, \qquad F = \begin{pmatrix} 0 \\ \widetilde{f_2}(t) \end{pmatrix},$$

and $\widetilde{f_2}$ is defined by

$$(\widetilde{f_2}(t), \psi) + \gamma(\nabla \widetilde{f_2}(t), \nabla\psi) = (f_2(t), \nabla\psi) \quad \forall \psi \in V,$$

and by using classical results on semigroups (see [8]). The linear estimates allow us to apply a fixed point technique. As the nonlinear terms in (2.2)–(2.8) are not Lipschitz in $W \times V$, we must work in the space $[H^2(\Omega)]^2 \cap U \times U \times H^3(\Omega) \cap W \times W$, suggested by Lemma 3.3. The local character of the solutions comes from the fact that even in this more regular space the nonlinearities are only locally Lipschitz. □

Step 2. The second step consists in proving the following estimates.

LEMMA 3.4. *Suppose that* (u, w) *is a strong solution of* (2.2)–(2.8) *defined on* $[0, T)$. *Then there exists a constant* M_T *depending on* T *such that*

$$||u''(t)||^2 + ||u'(t)||^2_{H^1} + ||w''(t)||^2_{H^1} + ||w'(t)||^2_{H^2} \le M_T \quad \forall t \in [0, T).$$

LEMMA 3.5. *Suppose that* (u, w) *is a strong solution of* (2.2)–(2.8) *defined on* $[0, T)$. *Then there exists a constant* M_T *depending on* T *such that*

$$||u(t)||_{H^2}^2 + ||w(t)||_{H^3}^2 \leq M_T \quad \forall t \in [0, T).$$

As for the proofs of Lemmas 3.4 and 3.5 we only mention two ideas.

(a) To prove Lemma 3.4 we take the derivative of (3.1), (3.2) with respect to time and we choose $\phi = u''$, $\psi = w''$ in the resulting equations. By the use of the divergence structure of the nonlinearities and of some Gagliardo–Nirenberg inequalities, we obtain that the expresion

$$\tfrac{1}{2}(||u''(t)||^2 + ||u'(t)||_{H^1}^2 + ||w''(t)||_{H^1}^2 + ||w'(t)||_{H^2}^2)$$

is uniformly bounded with respect to t varying in any compact set.

(b) For the proof of Lemma 3.5 we use Lemma 3.4 combined with the elliptic estimates provided by Lemma 3.3.

Step 3. It is now sufficient to note that, by Lemmas 3.4 and 3.5, the assertion (A2) in Lemma 3.1 cannot hold. As a consequence we obtain that (A1) takes place so any local solution can be extended to a global one. □

As a consequence of Theorem 3.1 we easily obtain global existence of weak solutions of (2.2)–(2.8) (for a complete proof see [10] or [11, Part 1]).

THEOREM 3.2. *Suppose that*

$$u^0 \in U, \quad u^1 \in \mathcal{K}, \quad w^0 \in W, \quad w^1 \in V.$$

Then there exists at least one weak solution of (*2.2*)*-*(*2.8*).

Remark. The solutions provided by our proof of Theorem 3.2 are weak limits of a sequence of strong solutions. We also remark that the uniqueness of weak solutions is an open problem.

4. The stabilization results.

Our main results assert that strong solutions of (2.2)-(2.8) satisfy the following energy estimate.

THEOREM 4.1. *Let* $\{u, w\}$ *be any strong solution of* (*2.2*)*-*(*2.8*), *with the feedback law* (*2.15*) *and let* $B > 0$. *Then there exist constants* $K > 0$, *and* $\omega = \omega(B) > 0$ *such that the following estimate holds, provided that* $E(0) \leq B$:

$$E(t) \leq Ke^{-\omega t}E(0) \quad \forall t \geq 0. \tag{4.1}$$

We can hope that estimate (4.1) might be extended for weak solutions of (2.2)–(2.8). We can easily check that any sequence of strong solutions that is bounded in the energy space converges weakly to a finite energy solution (cf. [10], [11, Part 1]). However, due to the lack of Lipschitzianity there is no uniqueness result for weak solutions, so we cannot assert that any weak solution can be obtained as above. This is why the next stabilization result considers only the weak solutions that can be approached by a sequence of strong solutions. By using the lower semicontinuity of the energy functional we obtain the following result.

THEOREM 4.2. *Suppose that* b, u^0, u^1, w^0, w^1 *satisfy the assumptions of Theorem* 3.2. *Then, among the weak solution*'*of* (2.2)–(2.8), *there exists at least one satisfying the energy estimate* (4.1).

The rest of our paper is devoted to the proof of Theorem 4.1. We begin by noting that any strong solution of (2.2)–(2.8) satisfies a slight generalization of (3.1), (3.2). More precisely, we have the following result.

LEMMA 4.1. *If $\{u,w\}$ is a strong solution of* (2.2)–(2.8), *with the feedback law* (2.15), *then the following relations hold*:

$$\begin{aligned}(4.2)\qquad & [(u',\phi)]' + (\mathcal{C}[\epsilon(u)+f(\nabla w)],\epsilon(\phi)) \\ & + c_2\int_\Omega\left(\frac{\partial u_1}{\partial x_1}\frac{\partial \phi_2}{\partial x_2}+\frac{\partial u_2}{\partial x_2}\frac{\partial \phi_1}{\partial x_1}-\frac{\partial u_1}{\partial x_2}\frac{\partial \phi_2}{\partial x_1}-\frac{\partial u_2}{\partial x_1}\frac{\partial \phi_1}{\partial x_2}\right)dX \\ & + c_1\left[\int_{\Gamma_1}(m\cdot\nu)u\cdot\phi d\Gamma\right]' \\ & - \int_{\Gamma_0}\mathcal{C}[\epsilon(u)]\nu\cdot\phi d\Gamma = 0 \quad \forall\phi\in[H^1(\Omega)]^2,\end{aligned}$$

$$\begin{aligned}(4.3)\ & [(w',\psi)+\gamma(\nabla w',\nabla\psi)]' + a(w,\psi) \\ & + (\mathcal{C}[\epsilon(u)+f(\nabla w)]\nabla w,\nabla\psi) - D\int_{\Gamma_0}\Delta w\frac{\partial\psi}{\partial\nu}d\Gamma \\ & + c_1\left[\int_{\Gamma_1}(m\cdot\nu)(w\psi+\nabla w\cdot\nabla\psi)d\Gamma\right]' = 0 \quad \forall\psi\in H^2(\Omega), \psi_{|\Gamma_0}=0.\end{aligned}$$

Proof. Let us first note that, for (u,w), the strong solution of (2.2)–(2.8), the relation (3.1) implies that (3.2) holds in $C([0,T],L^2(\Omega))$. This is why, by taking the scalar product of (2.2) with $\phi\in[H^1(\Omega)]^2$, we easily obtain that (u,w) satisfy (4.2). On the other hand, according to (2.10) (see also [3, p. 71]), the relation

$$\begin{aligned}(4.4)\qquad a(w,\psi) = & -\int_\Omega\nabla(\Delta w)\cdot\nabla\psi dX + D\int_\Gamma[\Delta w+(1-\mu)B_1 w]\frac{\partial\psi}{\partial\nu}d\Gamma \\ & - D(1-\mu)\int_\Gamma\frac{\partial B_2 w}{\partial s}\psi d\Gamma,\end{aligned}$$

holds for any $w\in H^3(\Omega)$ and $\psi\in H^2(\Omega)$, where the last integral may be interpreted as the duality between $H^{1/2}(\Gamma)$ and $H^{-1/2}(\Gamma)$. As w,w' satisfy (2.6) and (2.15) from (4.4) we get

(4.5)

$$\begin{aligned} a(w,\psi) + c_1\int_{\Gamma_1}(m\cdot\nu)(w'\psi+\nabla w'\cdot\nabla\psi)d\Gamma = & -\int_\Omega\nabla(\Delta w)\cdot\nabla\psi dX \\ - D(1-\mu)\int_{\Gamma_1}\frac{\partial B_2 w}{\partial s}\psi d\Gamma + c_1\int_{\Gamma_1} & \left\{(m\cdot\nu)w' - \frac{\partial}{\partial s}\left[(m\cdot\nu)\frac{\partial w'}{\partial s}\right]\right\}\psi d\Gamma,\end{aligned}$$

for any

$$\psi\in H^2(\Omega), \qquad \psi_{|\Gamma_0} = \frac{\partial\psi}{\partial\nu}_{|\Gamma_0} = 0.$$

From (3.2) and (4.5) it follows that u,w satisfy the following relation:

(4.6)

$$\begin{aligned}(w'',\psi) + \gamma(\nabla w'',\nabla\psi) - \int_\Omega\nabla(\Delta w)\cdot\nabla\psi dX + (\mathcal{C}[\epsilon(u)+f(\nabla w)]\nabla w,\nabla\psi) \\ -D(1-\mu)\int_{\Gamma_1}\frac{\partial B_2 w}{\partial s}\psi d\Gamma + c_1\int_{\Gamma_1}\left\{(m\cdot\nu)w' - \frac{\partial}{\partial s}\left[(m\cdot\nu)\frac{\partial w'}{\partial s}\right]\right\}\psi d\Gamma = 0,\end{aligned}$$

for all

$$\psi \in H^2(\Omega), \qquad \psi_{|\Gamma_0} = \frac{\partial \psi}{\partial \nu}_{|\Gamma_0} = 0,$$

and the last two integrals may be interpreted as the duality beetween $H^{1/2}(\Gamma)$ and $H^{-1/2}(\Gamma)$. It is obvious that, by density, (4.6) still holds for $\psi \in H^1(\Omega), \psi_{|\Gamma_0} = 0$. In particular (4.6) is true for $\psi \in H^2(\Omega)$, $\psi_{|\Gamma_0} = 0$. This implies, by using once again (4.4) and the relation $B_1 w = 0$ on Γ_0, that (4.3) holds. □

In the next section we shall give an energy identity that plays a fundamental role in proving Theorem 3.1.

5. The energy identity. For $u : \Omega \to \mathbf{R}^2$ we denote by $\nabla u m$ the product of the matrix ∇u with the column vector $m \in \mathbf{R}^2$. The following result will be essential for the proof of the stabilization result.

LEMMA 5.1. *Let* $\{u, w\}$ *be a strong solution of* (2.2)–(2.8), (2.15) *and let* $t > 0$, $\alpha \in \mathbf{R}$. *Define*

$$\begin{aligned}\rho(t) = &\int_\Omega u' \cdot [\nabla u m - (2\alpha - 1)u] dX - \frac{c_1(2\alpha-1)}{2}\int_{\Gamma_1} (m\cdot\nu)u^2 d\Gamma \\ &+ \int_\Omega [w'(m\cdot\nabla w) + \gamma\nabla w' \cdot \nabla(m\cdot\nabla w)]dX - \alpha\int_\Omega (w'w + \gamma\nabla w'\cdot\nabla w)dX \\ &- \frac{c_1\alpha}{2}\int_{\Gamma_1}(m\cdot\nu)(w'^2 + |\nabla w|^2)d\Gamma.\end{aligned}$$

Then we have

$$\begin{aligned}\rho(t) - \rho(0) &+ 2\alpha\int_0^t\int_\Omega u'^2 dX ds \\ &+ (\alpha+1)\int_0^t\int_\Omega w'^2 dXds + \alpha\gamma\int_0^t\int_\Omega |\nabla w'|^2 dXds \\ &+ (1-\alpha)\int_0^t a(w,w)ds \\ &+ (1-2\alpha)\int_0^t\int_\Omega \mathcal{C}[\epsilon(u) + f(\nabla w)]\cdot[\epsilon(u) + f(\nabla w)]dXds \\ &- c_2(2\alpha-1)\int_0^t\int_\Omega\left(\frac{\partial u_1}{\partial x_1}\frac{\partial u_2}{\partial x_2} - \frac{\partial u_1}{\partial x_2}\frac{\partial u_2}{\partial x_1}\right)dXds \\ =&\frac{1}{2}\int_0^t\int_{\Gamma_1}(m\cdot\nu)(u'^2 + w'^2 + \gamma|\nabla w'|^2)d\Gamma ds \\ &+ \frac{1}{2}\int_0^t\int_{\Gamma_0}\mathcal{C}[\epsilon(u)]\cdot\epsilon(u)(m\cdot\nu)d\Gamma ds \\ &- \frac{1}{2}\int_0^t\int_{\Gamma_1}\mathcal{C}[\epsilon(u) + f(\nabla w)]\cdot[\epsilon(u) + f(\nabla w)](m\cdot\nu)d\Gamma ds \\ &- c_2\int_0^t\int_{\Gamma_1}(m\cdot\nu)\left(\frac{\partial u_1}{\partial x_1}\frac{\partial u_2}{\partial x_2} - \frac{\partial u_1}{\partial x_2}\frac{\partial u_2}{\partial x_1}\right)d\Gamma ds \\ &+ \frac{D}{2}\int_0^t\int_{\Gamma_0}(m\cdot\nu)(\Delta w)^2 d\Gamma ds \\ &- \frac{D}{2}\int_0^t\int_{\Gamma_1}(m\cdot\nu)\left[\left(\frac{\partial^2 w}{\partial x_1^2}\right)^2 + \left(\frac{\partial^2 w}{\partial x_2^2}\right)^2 + 2\mu\frac{\partial^2 w}{\partial x_1^2}\frac{\partial^2 w}{\partial x_2^2}\right.\end{aligned} \tag{5.1}$$

$$+ 2(1-\mu)\left(\frac{\partial^2 w}{\partial x_1 \partial x_2}\right)^2\Bigg] d\Gamma ds - c_1 \int_0^t \int_{\Gamma_1} (m\cdot\nu)u' \cdot (\nabla u m) d\Gamma ds$$
$$- c_1 \int_0^t \int_{\Gamma_1} (m\cdot\nu)[w'(m\cdot\nabla w) + \nabla w' \cdot \nabla(m\cdot\nabla w)] d\Gamma ds.$$

Proof. We set $\phi = \nabla u m - \beta u$, with arbitrary β, in (3.1) and integrate that equation over $(0,t)$. For the first term we have

$$\text{(5.2)}\qquad \int_0^t \int_\Omega u'' \cdot (\nabla u m - \beta u) dX ds$$
$$= \int_\Omega u' \cdot (\nabla u m - \beta u) dX|_0^t - \int_0^t \int_\Omega u' \cdot (\nabla u' m - \beta u') dX ds.$$

We notice that

$$\text{(5.3)}\; - \int_\Omega u' \cdot (\nabla u' m) dX = - \int_\Omega u_i' \frac{\partial u_i'}{\partial x_j} m_j dX$$
$$= -\frac{1}{2}\int_\Omega m_j \frac{\partial}{\partial x_j}(u_i')^2 dX = -\frac{1}{2}\int_{\Gamma_1} (m\cdot\nu)u'^2 + \int_\Omega u'^2.$$

Taking in consideration (5.2), (5.3) we have

$$\text{(5.4)}\; \int_0^t \int_\Omega u'' \cdot (\nabla u m - \beta u) dX ds = \int_\Omega u' \cdot (\nabla u m - \beta u) dX|_0^t$$
$$- \frac{1}{2}\int_0^t \int_{\Gamma_1} (m\cdot\nu) u'^2 d\Gamma ds + \beta + 1) \int_0^t \int_\Omega u'^2 dX ds.$$

For the following terms we note that

$$\text{(5.5)}\qquad \int_\Omega \Bigg[\frac{\partial u_1}{\partial x_1}\frac{\partial(m\cdot\nabla u_2)}{\partial x_2} + \frac{\partial u_2}{\partial x_2}\frac{\partial(m\cdot\nabla u_1)}{\partial x_1}$$
$$- \frac{\partial u_1}{\partial x_2}\frac{\partial(m\cdot\nabla u_2)}{\partial x_1} - \frac{\partial u_2}{\partial x_1}\frac{\partial(m\cdot\nabla u_1)}{\partial x_2}\Bigg] dX$$
$$= \int_\Gamma (m\cdot\nu)\left(\frac{\partial u_1}{\partial x_1}\frac{\partial u_2}{\partial x_2} - \frac{\partial u_1}{\partial x_2}\frac{\partial u_2}{\partial x_1}\right) d\Gamma.$$

Let us make the notation

$$\text{(5.6)}\qquad \rho_1(t) = \int_\Omega u' \cdot (\nabla u m - \beta u) dX - \frac{\beta c_1}{2}\int_0^t \int_{\Gamma_1} (m\cdot\nu) u^2 d\Gamma.$$

As a consequence of (3.1), (5.5), and (5.6) we get

$$\text{(5.7)}\qquad \rho_1(t) - \rho_1(0) + (\beta+1)\int_0^t \int_\Omega u'^2 dX ds$$
$$- c_2\beta \int_\Omega \left(\frac{\partial u_1}{\partial x_1}\frac{\partial u_2}{\partial x_2} - \frac{\partial u_1}{\partial x_2}\frac{\partial u_2}{\partial x_1}\right) dX ds$$

$$+\int_0^t\int_\Omega \mathcal{C}[\epsilon(u)+f(\nabla w)]\cdot\epsilon(\nabla um-\beta u)dXds$$
$$=\frac{1}{2}\int_0^t\int_{\Gamma_1}(m\cdot\nu)u'^2d\Gamma ds+\int_0^t\int_{\Gamma_0}(m\cdot\nu)\mathcal{C}[\epsilon(u)]\cdot\epsilon(u)d\Gamma ds$$
$$-c_1\int_0^t\int_{\Gamma_1}(m\cdot\nu)u'\cdot(\nabla um)d\Gamma ds$$
$$-c_2\int_0^t\int_{\Gamma_1}(m\cdot\nu)\left(\frac{\partial u_1}{\partial x_1}\frac{\partial u_2}{\partial x_2}-\frac{\partial u_1}{\partial x_2}\frac{\partial u_2}{\partial x_1}\right)d\Gamma ds.$$

We now set $\psi = m\cdot\nabla w-\alpha w$ in (3.4) and integrate over $[0,t]$. We obtain

$$\begin{aligned}(5.8)\quad c(w',m\cdot\nabla w-\alpha w)|_0^t &- \int_0^t c(w',m\cdot\nabla w'-\alpha w')ds\\ &+\int_0^t a(w,m\cdot\nabla w-\alpha w)ds\\ &+\int_0^t(\mathcal{C}[\epsilon(u)+f(\nabla w)]\nabla w\cdot\nabla(m\cdot\nabla w-\alpha w)ds\\ &-D\int_0^t\int_{\Gamma_0}(m\cdot\nu)(\Delta w)^2d\Gamma ds+\int_0^t b(w',m\cdot\nabla w)ds\\ &-\alpha\int_0^t b(w',w)ds=0,\end{aligned}$$

where

$$(5.9)\qquad c(v_1,v_2)=\int_\Omega(v_1v_2+\gamma\nabla v_1\cdot\nabla v_2)dX,$$

$$(5.10)\qquad b(v_1,v_2)=c_1\int_{\Gamma_1}(m\cdot\nu)(v_1v_2+\nabla v_1\cdot\nabla v_2)d\Gamma.$$

The relation

$$\begin{aligned}c(w',m\cdot\nabla w')&=\int_\Omega[w'(m\cdot\nabla w')+\gamma\nabla w'\cdot\nabla(m\cdot\nabla w')]dX\\ &=\frac{1}{2}\int_\Omega \mathrm{div}(w'^2m)dX-\int_\Omega w'^2dX+\frac{\gamma}{2}\int_\Omega\mathrm{div}(|\nabla w'|^2m)dX\\ &=\frac{1}{2}\int_{\Gamma_1}(m\cdot\nu)(w'^2+\gamma|\nabla w'|^2)d\Gamma-\int_\Omega w'^2dX,\end{aligned}$$

allows us to write in a more convenient way the sum of the first two terms in (5.8), that is,

$$\begin{aligned}(5.11)\quad c(w',m\cdot\nabla w-\alpha w)|_0^t&-\int_0^t c(w',m\cdot\nabla w'-\alpha w')ds\\ &=c(w',m\cdot\nabla w-\alpha w)|_0^t+(\alpha+1)\int_0^t\int_\Omega w'^2dXds\\ +\alpha\gamma\int_0^t\int_\Omega|\nabla w'|^2dXds&-\frac{1}{2}\int_0^t\int_{\Gamma_1}(m\cdot\nu)(w'^2+\gamma|\nabla w'|^2)d\Gamma ds.\end{aligned}$$

We note that, according to [3, p. 81],

$$
\begin{aligned}
(5.12)\qquad a(w, m\cdot\nabla w) =& a(w,w)\\
&+\frac{D}{2}\int_{\Gamma}(m\cdot\nu)\left[\left(\frac{\partial^2 w}{\partial x_1^2}\right)^2+\left(\frac{\partial^2 w}{\partial x_2^2}\right)^2+2\mu\frac{\partial^2 w}{\partial x_1^2}\frac{\partial^2 w}{\partial x_1^2}\right.\\
&\qquad\left.+2(1-\mu)\left(\frac{\partial^2 w}{\partial x_1\partial x_2}\right)^2\right]d\Gamma.
\end{aligned}
$$

Taking in consideration the boundary conditions we obtain

$$
\begin{aligned}
(5.13)\ &\frac{D}{2}\int_{\Gamma_0}(m\cdot\nu)\left[\left(\frac{\partial^2 w}{\partial x_1^2}\right)^2+\left(\frac{\partial^2 w}{\partial x_2^2}\right)^2+2\mu\frac{\partial^2 w}{\partial x_1^2}\frac{\partial^2 w}{\partial x_1^2}+2(1-\mu)\left(\frac{\partial^2 w}{\partial x_1\partial x_2}\right)^2\right]d\Gamma\\
&=\frac{D}{2}\int_{\Gamma_0}(m\cdot\nu)(\Delta w)^2 d\Gamma,
\end{aligned}
$$

The relations (5.12) and (5.13) imply

$$
\begin{aligned}
(5.14)\ a(w, m\cdot\nabla w-\alpha w)&-D\int_{\Gamma_0}(m\cdot\nu)(\Delta w)^2 d\Gamma\\
&=(1-\alpha)a(w,w)-\frac{D}{2}\int_{\Gamma_0}(m\cdot\nu)(\Delta w)^2 d\Gamma\\
&\quad+\frac{D}{2}\int_{\Gamma_1}(m\cdot\nu)\left[\left(\frac{\partial^2 w}{\partial x_1^2}\right)^2+\left(\frac{\partial^2 w}{\partial x_2^2}\right)^2+2\mu\frac{\partial^2 w}{\partial x_1^2}\frac{\partial^2 w}{\partial x_1^2}\right.\\
&\qquad\left.+2(1-\mu)\left(\frac{\partial^2 w}{\partial x_1\partial x_2}\right)^2\right]d\Gamma.
\end{aligned}
$$

Let $\rho_2(t)=c(w', m\cdot\nabla w)-\alpha c(w', w)-\frac{\alpha}{2}b(w,w)$. The relations (5.8), (5.11), (5.14) imply that

$$
\begin{aligned}
(5.15)\,&\rho_2(t)-\rho_2(0)+(\alpha+1)\int_0^t\int_\Omega w'^2 dX ds+\alpha\gamma\int_0^t\int_\Omega|\nabla w'|^2 dX ds\\
&+(1-\alpha)\int_0^t a(w,w)ds\\
&+\int_0^t\int_\Omega\{\mathcal{C}[\epsilon(u)+f(\nabla w)]\nabla w\}\cdot\nabla(m\cdot\nabla w-\alpha w)dX ds\\
=&\frac{1}{2}\int_0^t\int_{\Gamma_1}(m\cdot\nu)(w'^2+\gamma|\nabla w'|^2)d\Gamma ds+\frac{D}{2}\int_0^t\int_{\Gamma_0}(m\cdot\nu)(\Delta w)^2 d\Gamma\\
&-\frac{D}{2}\int_0^t\int_{\Gamma}(m\cdot\nu)\left[\left(\frac{\partial^2 w}{\partial x_1^2}\right)^2+\left(\frac{\partial^2 w}{\partial x_2^2}\right)^2+2\mu\frac{\partial^2 w}{\partial x_1^2}\frac{\partial^2 w}{\partial x_1^2}\right.\\
&\qquad\left.+2(1-\mu)\left(\frac{\partial^2 w}{\partial x_1\partial x_2}\right)^2\right]d\Gamma-\int_0^t b(w', m\cdot\nabla w)ds.
\end{aligned}
$$

As $\rho(t)=\rho_1(t)+\rho_2(t)$ from (5.7) and (5.15), with $\beta=2\alpha-1$, we obtain

$$
\begin{aligned}
(5.16)\quad & \rho(t) - \rho(0) \\
& + \int_0^t \int_\Omega [(\beta+1)u'^2 + (\alpha+1)w'^2 + \alpha\gamma|\nabla w'|^2] dX ds \\
& + (1-\alpha) \int_0^t a(w,w) ds \\
& + \int_0^t \int_\Omega \{\mathcal{C}[\epsilon(u) + f(\nabla w)]\nabla w\} \cdot \nabla(m \cdot \nabla w - \alpha w) dX ds \\
& + \int_0^t \int_\Omega \mathcal{C}[\epsilon(u) + f(\nabla w)] \cdot \epsilon(\nabla u m - \beta u) dX ds \\
& - c_2 \beta \int_0^t \int_\Omega \left(\frac{\partial u_1}{\partial x_1} \frac{\partial u_2}{\partial x_2} - \frac{\partial u_1}{\partial x_2} \frac{\partial u_2}{\partial x_1} \right) dX ds \\
= & \frac{1}{2} \int_0^t \int_{\Gamma_1} (m \cdot \nu)(u'^2 + w'^2 + \gamma|\nabla w'|^2) d\Gamma ds \\
& - c_2 \int_0^t \int_{\Gamma_1} (m \cdot \nu) (\frac{\partial u_1}{\partial x_1} \frac{\partial u_2}{\partial x_2} - \frac{\partial u_1}{\partial x_2} \frac{\partial u_2}{\partial x_1}) d\Gamma ds \\
& + \int_0^t \int_{\Gamma_0} (m \cdot \nu) \mathcal{C}[\epsilon(u)] \cdot \epsilon(u) dX ds + \frac{D}{2} \int_0^t \int_{\Gamma_0} (m \cdot \nu)(\Delta w)^2 d\Gamma \\
& - \frac{D}{2} \int_0^t \int_{\Gamma_1} (m \cdot \nu) \Bigg[\left(\frac{\partial^2 w}{\partial x_1^2} \right)^2 + \left(\frac{\partial^2 w}{\partial x_2^2} \right)^2 \\
& \qquad\qquad + 2\mu \frac{\partial^2 w}{\partial x_1^2} \frac{\partial^2 w}{\partial x_1^2} + 2(1-\mu) \left(\frac{\partial^2 w}{\partial x_1 \partial x_2} \right)^2 \Bigg] d\Gamma ds \\
& - c_1 \int_0^t \int_{\Gamma_1} (m \cdot \nu) u' \cdot (\nabla u m) d\Gamma ds - \int_0^t b(w', m \cdot \nabla w) ds.
\end{aligned}
$$

Let us consider the terms containing both u and w from the left side of (5.16). We note that

$$
\begin{aligned}
(5.17)\quad & \int_\Omega \mathcal{C}[\epsilon(u) + f(\nabla w)] \cdot \epsilon(\nabla u m - \beta u) dX \\
& = (1-\beta) \int_\Omega \mathcal{C}[\epsilon(u) + f(\nabla w)] \cdot \epsilon(u) dX + \int_\Omega a_{ij} \frac{\partial^2 u_i}{\partial x_k \partial x_j} m_k dX,
\end{aligned}
$$

where $a_{ij}(\epsilon(u), f(\nabla w)) = \{\mathcal{C}[\epsilon(u) + f(\nabla w)]\}_{ij}, i, j = 1, 2$. For the other coupled term we have

$$
\begin{aligned}
(5.18)\quad & \int_\Omega \{\mathcal{C}[\epsilon(u) + f(\nabla w)]\nabla w\} \cdot \nabla(m \cdot \nabla w - \alpha w) dX \\
& = 2(1-\alpha) \int_\Omega \{\mathcal{C}[\epsilon(u) + f(\nabla w)]\} \cdot f(\nabla w) dX + \int_\Omega a_{ij} \frac{\partial w}{\partial x_j} \frac{\partial^2 w}{\partial x_i \partial x_k} m_k dX.
\end{aligned}
$$

To have a good coupling of (5.17) and (5.18) we use again the condition $\beta = 2\alpha - 1$. Then (5.17), (5.18) give

$$
(5.19)\qquad \int_\Omega \mathcal{C}[\epsilon(u) + f(\nabla w)] \cdot \epsilon[\nabla u m - (2\alpha - 1)u] dX
$$

$$
\begin{aligned}
&+\int_\Omega \{\mathcal{C}[\epsilon(u)+f(\nabla w)]\nabla w\}\cdot\nabla(m\cdot\nabla w-\alpha w)dX\\
&=2(1-\alpha)\int_\Omega\{\mathcal{C}[\epsilon(u)+f(\nabla w)]\cdot[\epsilon(u)+f(\nabla w)]dX\\
&+\int_\Omega a_{ij}\left(\frac{\partial^2 u_i}{\partial x_k x_j}+\frac{\partial w}{\partial x_j}\frac{\partial^2 w}{\partial x_i\partial x_k}\right)m_k dX.
\end{aligned}
$$

By using in (5.19) the identity

$$
\begin{aligned}
&a_{ij}(\frac{\partial^2 u_i}{\partial x_k x_j}+\frac{\partial w}{\partial x_j}\frac{\partial^2 w}{\partial x_i\partial x_k})m_k\\
&=\frac{1}{2}\mathrm{div}\{\mathcal{C}[\epsilon(u)+f(\nabla w)]\cdot[\epsilon(u)+f(\nabla w)]m\}\\
&-\mathcal{C}[\epsilon(u)+f(\nabla w)]\cdot[\epsilon(u)+f(\nabla w)],
\end{aligned}
$$

we obtain

$$
\begin{aligned}
(5.20)\qquad &\int_\Omega \mathcal{C}[\epsilon(u)+f(\nabla w)]\cdot\epsilon[\nabla u m-(2\alpha-1)u]dX\\
&+\int_\Omega\{\mathcal{C}[\epsilon(u)+f(\nabla w)]\nabla w\}\cdot\nabla(m\cdot\nabla w-\alpha w)dX\\
&=\int_\Omega(1-2\alpha)\mathcal{C}[\epsilon(u)+f(\nabla w)]\cdot[\epsilon(u)+f(\nabla w)]dX\\
&+\frac{1}{2}\int_{\Gamma_0}\mathcal{C}[\epsilon(u)]\cdot\epsilon(u)(m\cdot\nu)dX\\
&+\frac{1}{2}\int_{\Gamma_1}\mathcal{C}[\epsilon(u)+f(\nabla w)]\cdot[\epsilon(u)+f(\nabla w)](m\cdot\nu)dX.
\end{aligned}
$$

Now the conclusion (5.1) follows from (5.16) and (5.20). □

6. Proof of the stabilization results. We begin this section by noting that $\rho(t)$ has another important property given by the following lemma.

LEMMA 6.1. *For all $M>0$ there is a constant $C_0>0$ such that $\rho(t)\le C_0\widetilde{E}(t)$, for all $t\ge 0$, where $\widetilde{E}(t)$ is defined by* (2.17) *and u^0,u^1,w^0,w^1 are such that $\widetilde{E}(0)\le M$.*

Proof. We obviously have

$$
\left(\frac{\partial u_i}{\partial x_j}+\frac{\partial u_j}{\partial x_i}\right)^2\le 2\left(\frac{\partial u_i}{\partial x_j}+\frac{\partial u_j}{\partial x_i}+\frac{\partial w}{\partial x_i}\frac{\partial w}{\partial x_j}\right)^2+2\left(\frac{\partial w}{\partial x_i}\frac{\partial w}{\partial x_j}\right)^2,
$$

which implies that

$$
\begin{aligned}
\int_\Omega|\epsilon(u)|^2dX&\le 2\int_\Omega[\epsilon(u)+f(\nabla w)]^2dX+\frac{1}{2}||\frac{\partial w}{\partial x_i}||^2_{L^4(\Omega)}||\frac{\partial w}{\partial x_j}||^2_{L^4(\Omega)}\\
&\le 2\int_\Omega[\epsilon(u)+f(\nabla w)]^2dX+C_1||w||^4_{H^2(\Omega)}.
\end{aligned}
$$

Here we also used the imbedding $H^1(\Omega)\subset L^4(\Omega)$ and the energy inequality (2.14). The last inequality, (2.1), the H^2 coercivity of $a(w,\widetilde{w})$, and the fact that $\widetilde{E}(t)$ is nonincreasing imply

$$
\int_\Omega|\epsilon(u)|^2dX\le K_0\widetilde{E}(t)+K_1M\widetilde{E}(t),
$$

with K_O, K_1 positive constants. From (2.17), the above inequality, the continuity of the trace mapping $\gamma : H^1(\Omega) \to L^2(\Gamma)$, and Korn inequality, we obtain the following estimate:

$$\int_{\Gamma_1} (m\cdot\nu)u^2 d\Gamma \le K_2 \int_\Omega |\epsilon(u)|^2 dX \le K_3\widetilde{E}(t),$$

with K_2 and K_3 two positive constants. In a similar way we can prove that there is a positive constant K_4 such that

$$\int_\Omega u'\cdot[\nabla u m - (2\alpha-1)u]dX \le K_4\widetilde{E}(t).$$

As the other terms in ρ are obviously dominated by $E(t)$ and (according to (2.20)) by $\widetilde{E}(t)$, the conclusion of the lemma easily follows. □

We can now give the proof of the stabilization result.

Proof of Theorem 3.1. We set $\alpha \in (0,\frac{1}{2})$ in (5.1) and by taking the derivative we obtain

$$(6.1)\ \rho'(t) = -\int_\Omega \Bigg[2\alpha u'^2 + (\alpha+1)w'^2 + \alpha\gamma|\nabla w'|^2 + c_2(1-2\alpha)\left(\frac{\partial u_1}{\partial x_1}\frac{\partial u_2}{\partial x_2} - \frac{\partial u_1}{\partial x_2}\frac{\partial u_2}{\partial x_1}\right)\Bigg]dX$$
$$- (1-\alpha)a(w,w) - (1-2\alpha)\int_\Omega \mathcal{C}[\epsilon(u)+f(\nabla w)]\cdot[\epsilon(u)+f(\nabla w)]dX$$
$$+\frac{1}{2}\int_{\Gamma_1}(m\cdot\nu)(u'^2+w'^2+\gamma|\nabla w'|^2)d\Gamma - P_\Gamma(u,w)$$
$$- c_1\int_{\Gamma_1}(m\cdot\nu)u'\cdot(\nabla u m)d\Gamma - b(w', m\cdot\nabla w),$$

where

$$P_\Gamma(u,w) = \frac{1}{2}\int_\Gamma \mathcal{C}[\epsilon(u)+f(\nabla w)]\cdot[\epsilon(u)+f(\nabla w)]|m\cdot\nu|d\Gamma$$
$$+ c_2\int_\Gamma (m\cdot\nu)(\frac{\partial u_1}{\partial x_1}\frac{\partial u_2}{\partial x_2} - \frac{\partial u_1}{\partial x_2}\frac{\partial u_2}{\partial x_1})d\Gamma$$
$$+\frac{D}{2}\int_\Gamma |m\cdot\nu|\left[\left(\frac{\partial^2 w}{\partial x_1^2}\right)^2 + \left(\frac{\partial^2 w}{\partial x_2^2}\right)^2 + 2\mu\frac{\partial^2 w}{\partial x_1^2}\frac{\partial^2 w}{\partial x_2^2} + 2(1-\mu)\left(\frac{\partial^2 w}{\partial x_1\partial x_2}\right)^2\right]d\Gamma.$$

We obviously have

(6.2)

$$P_\Gamma(u,w) \ge \frac{1}{2}\int_{\Gamma_1} \mathcal{C}[\epsilon(u)+f(\nabla w)]\cdot[\epsilon(u)+f(\nabla w)](m\cdot\nu)d\Gamma$$
$$+ c_2\int_{\Gamma_1} (m\cdot\nu)\left(\frac{\partial u_1}{\partial x_1}\frac{\partial u_2}{\partial x_2} - \frac{\partial u_1}{\partial x_2}\frac{\partial u_2}{\partial x_1}\right)d\Gamma$$
$$+\frac{D}{2}\int_{\Gamma_1} (m\cdot\nu)\left[\left(\frac{\partial^2 w}{\partial x_1^2}\right)^2 + \left(\frac{\partial^2 w}{\partial x_2^2}\right)^2 + 2\mu\frac{\partial^2 w}{\partial x_1^2}\frac{\partial^2 w}{\partial x_2^2} + 2(1-\mu)\left(\frac{\partial^2 w}{\partial x_1\partial x_2}\right)^2\right]d\Gamma.$$

By applying (2.1) and the inequality

$$(x+y)^2 \ge \delta x^2 - \frac{\delta}{1-\delta}y^2, \quad \text{for all } x,y\in\mathbf{R} \quad \text{and} \quad \delta\in(0,1),$$

we obtain

$$\frac{1}{2}\int_{\Gamma_1} \mathcal{C}[\epsilon(u)+f(\nabla w)]\cdot[\epsilon(u)+f(\nabla w)](m\cdot\nu)d\Gamma$$
$$\geq \lambda_0\delta\int_{\Gamma_1}(m\cdot\nu)|\epsilon(u)|^2 d\Gamma - \frac{\lambda_0\delta}{1-\delta}\int_{\Gamma_1}(m\cdot\nu)[f(\nabla w)]^2 d\Gamma \quad \forall\delta\in(0,1).$$

As the trace mapping from $H^2(\Omega)$ to $H^1(\Gamma)$ and the imbedding $H^{1/2}(\Gamma)\subset L^4(\Gamma)$ are continous, the inequality above and (6.2) imply that there is a constant $K_5>0$ such that

(6.3)
$$P_\Gamma(u,w)\geq\lambda_0\delta\int_{\Gamma_1}[\epsilon(u)]^2(m\cdot\nu)d\Gamma-\frac{K_5\delta}{1-\delta}a(w,w)$$
$$+c_2\int_{\Gamma_1}(m\cdot\nu)\left(\frac{\partial u_1}{\partial x_1}\frac{\partial u_2}{\partial x_2}-\frac{\partial u_1}{\partial x_2}\frac{\partial u_2}{\partial x_1}\right)d\Gamma$$
$$+\frac{D}{2}\int_{\Gamma_1}(m\cdot\nu)\left[\left(\frac{\partial^2 w}{\partial x_1^2}\right)^2+\left(\frac{\partial^2 w}{\partial x_2^2}\right)^2+2\mu\frac{\partial^2 w}{\partial x_1^2}\frac{\partial^2 w}{\partial x_2^2}+2(1-\mu)\left(\frac{\partial^2 w}{\partial x_1\partial x_2}\right)^2\right]d\Gamma.$$

Let us now choose $\delta\in(0,1)$ such that

$$1-\alpha-\frac{K_5\delta}{1-\delta}\geq\frac{1-\alpha}{2}. \tag{6.4}$$

We note that, for b small enough (with respect to δ), (6.4) implies

(6.5)
$$P_\Gamma(u,w)\geq K_6\delta\int_{\Gamma_1}|\nabla u|^2(m\cdot\nu)d\Gamma-\frac{K_5\delta}{1-\delta}a(w,w)$$
$$+\frac{D}{2}\int_{\Gamma_1}(m\cdot\nu)\left[\left(\frac{\partial^2 w}{\partial x_1^2}\right)^2+\left(\frac{\partial^2 w}{\partial x_2^2}\right)^2+2\mu\frac{\partial^2 w}{\partial x_1^2}\frac{\partial^2 w}{\partial x_2^2}+2(1-\mu)\left(\frac{\partial^2 w}{\partial x_1\partial x_2}\right)^2\right]d\Gamma,$$

where $K_6>0$. From (6.1), (6.4) and (6.5) it follows that

$$\text{(6.6)}\ \rho'(t)\leq-\int_\Omega\Big[2\alpha u'^2+(\alpha+1)w'^2+\alpha\gamma|\nabla w'|^2$$
$$+c_2(1-2\alpha)\left(\frac{\partial u_1}{\partial x_1}\frac{\partial u_2}{\partial x_2}-\frac{\partial u_1}{\partial x_2}\frac{\partial u_2}{\partial x_1}\right)\Big]dX$$
$$-\frac{1-\alpha}{2}a(w,w)-(1-2\alpha)\int_\Omega\mathcal{C}[\epsilon(u)+f(\nabla w)]\cdot[\epsilon(u)+f(\nabla w)]dX$$
$$+\frac{1}{2}\int_{\Gamma_1}(m\cdot\nu)(u'^2+w'^2+\gamma|\nabla w'|^2)d\Gamma-V_{\Gamma_1}(u,w)$$
$$-c_1\int_{\Gamma_1}(m\cdot\nu)u'\cdot(\nabla u m)d\Gamma-b(w',m\cdot\nabla w),$$

where

(6.7)
$$V_{\Gamma_1}(u,w)=K_6\delta\int_{\Gamma_1}|\nabla u|^2(m\cdot\nu)d\Gamma$$
$$+\frac{D}{2}\int_{\Gamma_1}(m\cdot\nu)\left[\left(\frac{\partial^2 w}{\partial x_1^2}\right)^2+\left(\frac{\partial^2 w}{\partial x_2^2}\right)^2+2\mu\frac{\partial^2 w}{\partial x_1^2}\frac{\partial^2 w}{\partial x_2^2}+2(1-\mu)\left(\frac{\partial^2 w}{\partial x_1\partial x_2}\right)^2\right]d\Gamma.$$

We also note that for every $\eta > 0$ we have the estimates

$$(6.8)\qquad \begin{aligned}|b(w', m\cdot\nabla w)| &\leq [b(w',w')b(m\cdot\nabla w, m\cdot\nabla w)]^{1/2}\\ &\leq \frac{1}{2\eta}b(w',w') + \frac{\eta}{2}b(m\cdot\nabla w, m\cdot\nabla w) \quad \forall\eta > 0,\end{aligned}$$

$$(6.9)\ |c_1\int_{\Gamma_1}(m\cdot\nu)u'\cdot(\nabla um)d\Gamma| \leq \frac{c_1}{2\eta}\int_{\Gamma_1}(m\cdot\nu)u'^2d\Gamma + \frac{c_1\eta}{2}\int_{\Gamma_1}(m\cdot\nu)(\nabla um)^2d\Gamma.$$

Taking in consideration (2.19), (6.8), and (6.9) it follows that

$$(6.10)\qquad \begin{aligned}&-c_1\int_{\Gamma_1}(m\cdot\nu)u'\cdot(\nabla um)d\Gamma - b(w', m\cdot\nabla w) \leq -\frac{1}{2\eta}\widetilde{E}'(t)\\ &+\frac{c_1\eta}{2}\int_{\Gamma_1}(m\cdot\nu)(\nabla um)^2d\Gamma + \frac{\eta}{2}b(m\cdot\nabla w, m\cdot\nabla w) \quad \forall\eta > 0.\end{aligned}$$

We also note that

$$(6.11)\qquad \frac{c_1\eta}{2}\int_{\Gamma_1}(m\cdot\nu)(\nabla um)^2d\Gamma \leq C_1\eta V_{\Gamma_1}(u,w),$$

with $C_1 > 0$. By using the fact that $0 < \mu < \frac{1}{2}$ we obtain

$$(6.12)\qquad b(m\cdot\nabla w, m\cdot\nabla w) \leq C_2[a(w,w) + V_{\Gamma_1}(u,w)].$$

From (6.6)–(6.12), it follows that there is a constant $k > 0$ such that

$$(6.13)\qquad \begin{aligned}\rho'(t) \leq &-(k - C_3\eta)\widetilde{E}(t)\\ &+\frac{1}{2}\int_{\Gamma_1}(m\cdot\nu)(u'^2 + w'^2 + \gamma|\nabla w'|^2)\\ &-[1-(C_1+C_2)\eta]V_{\Gamma_1}(u,w) - \frac{1}{2\eta}\widetilde{E}'(t).\end{aligned}$$

For η small enough relation (6.13) becomes

$$(6.14)\qquad \begin{aligned}\rho'(t) \leq &-\frac{1}{2\eta}\widetilde{E}'(t) - \frac{k}{2}\widetilde{E}(t)\\ &+\frac{1}{2}\int_{\Gamma_1}(m\cdot\nu)(u'^2 + w'^2 + \gamma|\nabla w'|^2) - \frac{1}{2}V_{\Gamma_1}(u,w).\end{aligned}$$

We now introduce a Lyapunov functional obtained by slightly perturbating $\widetilde{E}$ in the direction of ρ. Let $F_\epsilon(t) = \widetilde{E}(t) + \epsilon\rho(t)$. From (6.14) we obtain

$$(6.15)\qquad \begin{aligned}F_\epsilon'(t) \leq &(1 - \frac{\epsilon}{2\eta})\widetilde{E}'(t) - \frac{k\epsilon}{2}\widetilde{E}(t)\\ &+\frac{\epsilon}{2}\int_{\Gamma_1}(m\cdot\nu)(u'^2 + w'^2 + \gamma|\nabla w'|^2)d\Gamma - \frac{\epsilon}{2}V_{\Gamma_1}(u,w).\end{aligned}$$

But, for $\gamma \leq 1$,

$$\widetilde{E}'(t) = -a\int_{\Gamma_1}(m\cdot\nu)(u'^2 + w'^2 + |\nabla w'|^2)d\Gamma$$

$$\leq -c_1 \int_{\Gamma_1} (m \cdot \nu)(u'^2 + w'^2 + \gamma|\nabla w'|^2) d\Gamma,$$

so (6.15) becomes

$$F_\epsilon'(t) \leq \left[\frac{\epsilon}{2} - c_1\left(1 - \frac{\epsilon}{2\eta}\right)\right] \int_{\Gamma_1} (m \cdot \nu)(u'^2 + w'^2 + \gamma|\nabla w'|^2) d\Gamma \\ - \frac{k\epsilon}{2}\widetilde{E}(t) - \frac{\epsilon}{2} V_{\Gamma_1}(u, w). \tag{6.16}$$

We now choose ϵ in such way that $\epsilon \leq 4c_1\eta/(c_1 + \eta)$ which implies that the inequality $\epsilon/2 - c_1(1 - \epsilon/2\eta) \leq -\epsilon/4$ also holds. In this case (6.16) becomes

$$F_\epsilon'(t) \leq -\frac{k\epsilon}{2}\widetilde{E}(t) - \frac{\epsilon}{2} E_\Gamma(u, w), \tag{6.17}$$

where $E_\Gamma(u,w) = \frac{1}{2}\int_{\Gamma_1}(m \cdot \nu)(u'^2 + w'^2 + \gamma|\nabla w'|^2)d\Gamma + V_{\Gamma_1}(u,w)$. From (6.17) we obtain

$$F_\epsilon'(t) \leq -\frac{k\epsilon}{2}\widetilde{E}(t). \tag{6.18}$$

We note now that (2.20), (6.18), and Lemma 6.1 easily imply (4.1). □

REFERENCES

[1] M. Bradley and I. Lasiecka, *Local exponential stabilization of a nonlinearly perturbed plate*, Nonlinear Anal., 18 (1991), pp. 333–343.

[2] V. Komornik and E. Zuazua, *Stabilisation frontière de l'èquation des ondes: Une méthode directe*, C. R. Acad. Sci. Paris Sér. I Math., 305 (1987), pp. 605–608.

[3] J. E. Lagnese, *Boundary Stabilization of Thin Elastic Plates*, SIAM Stud. Appl. Math., Vol. 10, Society for Industrial and Applied Mathematics, Philadelphia, PA, 1987.

[4] ———, *Uniform asymptotic energy estimates for solutions of the equations of dynamic plane elasticity with nonlinear dissipation at the boundary*, in Nonlinear Anal., 16 (1991), pp. 35–54.

[5] ———, *Modelling and stabilization of nonlinear plates*, Internat. Ser. Numer. Math., 100 (1991), pp. 247–264.

[6] J. E. Lagnese and J. L. Lions, *Modelling, Analysis and Control of Thin Plates*, Collection Recherche en Mathematiques Appliquées, Paris, 1988.

[7] J. E. Lagnese and G. Leugering, *Uniform stabilization of a nonlinear beam by nonlinear boundary feedback*, J. Differential Equations, 91 (1991), pp. 355–388.

[8] A. Pazy, *Semigroups of Linear Operators and Applications to Partial Differential Equations*, Springer-Verlag, New York, Berlin, 1983.

[9] J. P. Puel and M. Tucsnak, *Stabilisation frontière pour les équations de von Kármán*, Comptes Rendus de l'Académie des Sciences, Série I, 314 (1992), pp. 609–612.

[10] ———*Global existence for the full von Kármán system*, unpublished manuscript.

[11] M. Tucsnak, *Résultats de stabilisation sur quelques modèles non linéaires de poutres et de plaques élastiques*, thèse de l'Université d'Orléans, 1992.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 33, No. 1, pp. 274-298, January 1995

© 1995 Society for Industrial and Applied Mathematics
013

AN EXTENSION OF PONTRYAGIN'S PRINCIPLE FOR STATE-CONSTRAINED OPTIMAL CONTROL OF SEMILINEAR ELLIPTIC EQUATIONS AND VARIATIONAL INEQUALITIES*

FRÉDÉRIC BONNANS† AND EDUARDO CASAS‡

Abstract. This paper deals with state-constrained optimal control problems governed by semilinear elliptic equations or variational inequalities. By using Ekeland's principle, a minimum principle of Pontryagin's type under some stability conditions of the optimal cost with respect to the state constraints is derived.

Key words. Pontryagin's principle, boundary control, semilinear elliptic operators, optimality conditions, state constraints

AMS subject classifications. 49K20, 35J65, 35J85

1. Introduction. There exists a vast literature devoted to Pontryagin's principle for optimal control problems governed by ordinary differential equations or evolution partial differential equations, but very few papers have considered the case of elliptic equations. A simple case corresponding to a linear equation was studied by Lions [19]. More recently, the authors derived Pontryagin's principle for semilinear monotone elliptic equations in [7]. Here we extend the results of the last work by letting the existence of pointwise state constraints generalize some preliminary results of Bonnans [3], [4]; in [8] we considered, assuming continuity of the data, the case of boundary as well as distributed control and obtained a "symmetric" formulation of the optimality system involving boundary and interior hamiltonians. See Bonnans and Casas [5] for a different approach to the optimality conditions of state-constrained control problems.

The difficulty of deriving the optimality conditions for control problems associated with variational inequalities is well known; see the works of Mignot [20], Mignot and Puel [21], and Barbu [2]. Zheng-Xu He [18] obtained the optimality conditions for state-constrained problems governed by variational inequalities, and Bonnans and Tiba [9] proved Pontryagin's principle for control problems of semilinear elliptic variational inequalities. Here we will derive a principle of Pontryagin's type for state-constrained control problems of semilinear elliptic variational inequalities.

In this article we prove Pontryagin's principle as follows: with the aid of Ekeland's principle, we introduce a family of control problems without state constraints for which some approximate solutions converge toward the optimal control of the initial problem; we derive the optimality conditions for the problems of this family by using some results on problems without state constraints that generalize those of Bonnans and Casas [7] and Bonnans and Tiba [9] and finally pass to the limit. In order to apply Ekeland's principle we need to assume some stability conditions of the optimal cost with respect to small perturbations of the feasible state set. We distinguish two different stability conditions, called weak and strong, respectively. Under

* Received by the editors October 1, 1992; accepted for publication (in revised form) September 24, 1993.

† Institut National de Recherche en Informatique et en Automatique, Domaine de Voluceau, BP 105, Rocquencourt, 78153 Le Chesnay, France.

‡ Departamento de Matemática Aplicada y Ciencias de la Computación, E.T.S.I. de Caminos, C. y P., Universidad de Cantabria, 39071 Santander, Spain. The research of this author was partially supported by Dirección General de Investigación Científica y Técnica (Madrid).

a weak stability condition we derive the optimality conditions in a nonqualified form, while the strong stability allows us to prove a qualified Pontryagin's principle. The weak stability condition has been used by Casas [13] to prove the convergence of the numerical approximations of state-constrained control problems.

The paper is organized as follows: in the next section we formulate the control problem associated with a monotone semilinear elliptic equation, and in §3 the statements of the weak and strong Pontryagin's principles are presented; in §4 we give some technical results used in §§5 and 6 to prove the theorems stated in the third section; finally §7 is devoted to the control of variational inequalities.

2. Setting of the problem. Let Ω be an open and bounded subset of R^n, $n \geq 1$, with a Lipschitz boundary Γ. Given a nonempty bounded set $K \subset R^m$, $m \geq 1$, and $f : \Omega \times R \times K \longrightarrow R$ we consider the following boundary value problem:

$$\tag{2.1} \left\{ \begin{array}{l} Ay = f(x, y(x), u(x)) \quad \text{in } \Omega, \\ y = 0 \quad \text{on } \Gamma, \end{array} \right.$$

where

$$Ay = -\sum_{i,j=1}^{n} \partial_{x_j} \left(a_{ij}(x) \partial_{x_i} y(x) \right),$$

$$\tag{2.2} \left\{ \begin{array}{l} a_{ij} \in C^{0,1}(\overline{\Omega}) \text{ and} \\ \exists \Lambda > 0 \text{ such that } \displaystyle\sum_{i,j=1}^{n} a_{ij}(x)\xi_i\xi_j \geq \Lambda |\xi|^2 \quad \forall \xi \in R^n, \ \forall x \in \Omega. \end{array} \right.$$

We recall that $C^{0,\alpha}(\overline{\Omega})$, with $\alpha \in (0,1]$, is the space of all continuous functions in $\overline{\Omega}$ that satisfy the Hölder condition

$$\sup_{x_1, x_2 \in \overline{\Omega}} \frac{|u(x_2) - u(x_1)|}{|x_2 - x_1|^\alpha} < +\infty.$$

Given two measurable functions $L : \Omega \times R \times K \longrightarrow R$ and $g : \overline{\Omega} \times R \longrightarrow R$, for every $\delta > 0$ we formulate the control problem

$$(P_\delta) \left\{ \begin{array}{l} \min J(y, u) = \displaystyle\int_\Omega L(x, y(x), u(x))\, dx \\ (y, u) \text{ satisfies (2.1)}, \ u(x) \in K \text{ a.e. } x \in \Omega \text{ and } g(x, y(x)) \leq \delta \ \forall x \in \Omega. \end{array} \right.$$

We will make the following assumptions on the functions defining the problem (P_δ): $g \in C(\overline{\Omega} \times R)$; g, L, and f are continuously differentiable with respect to the second variable for every $(x, u) \in \Omega \times K$; and there exist functions $M_1 \in L^s(\Omega)$, $s > n/2$, and $s \geq 2$, $M_2 \in L^1(\Omega)$, and η increasing monotone verifying for every $(x, y, u) \in \Omega \times R \times K$

$$\tag{2.3} \left\{ \begin{array}{l} |f(x,0,u)| + \left| \dfrac{\partial f}{\partial y}(x, y, u) \right| \leq M_1(x) + \eta(|y|), \\ \\ |L(x,0,u)| + \left| \dfrac{\partial L}{\partial y}(x, y, u) \right| \leq M_2(x) + \eta(|y|), \\ \\ \dfrac{\partial g}{\partial y} \in C(\overline{\Omega} \times R), \quad \dfrac{\partial f}{\partial y}(x, y, u) \leq 0. \end{array} \right.$$

We will say that a control $u : \Omega \longrightarrow R^m$ is feasible if $u(x) \in K$ a.e. $x \in \Omega$ and the mapping $(x, y) \longrightarrow (f(x, y, u(x)), L(x, y, u(x)))$ is measurable in $\Omega \times R$. The set of feasible controls is denoted by $\mathcal{K}$. In this set we define the distance, called Ekeland's distance, as

$$d(u, v) = m(\{x \in \Omega : u(x) \neq v(x)\}),$$

where m denotes the Lebesgue measure. Adapting the proof of Ekeland [16] to our case it is easy to check that $(\mathcal{K}, d)$ is a complete metric space (the only difference is that we have to check the feasibility of the limit of a Cauchy sequence, which is immediate from the definition and the fact that a limit of measurable functions is measurable).

Under the previous hypotheses and thanks to the boundedness of K, we can deduce the following theorem.

THEOREM 2.1. *There exist constants* $C_1 > 0$ *and* $\alpha \in (0, 1)$ *such that for every* $u \in \mathcal{K}$ *equation* (2.1) *has a unique solution* $y_u \in H_0^1(\Omega) \cap C^{0,\alpha}(\overline{\Omega})$ *satisfying*

$$\|y_u\|_{H_0^1(\Omega)} + \|y_u\|_{C^{0,\alpha}(\overline{\Omega})} \leq C_1. \tag{2.4}$$

Furthermore the mapping $u \in (\mathcal{K}, d) \longrightarrow y_u \in H_0^1(\Omega) \cap C^{0,\alpha}(\overline{\Omega})$ *is continuous.*

Before proving this theorem we state the following lemma.

LEMMA 2.2. *There exist* $\alpha \in (0, 1)$, C_2, *and* C_3 *such that for every* $a, b \in L^s(\Omega)$, $a(x) \geq 0$, *the problem*

$$\begin{cases} Ay + ay = b & \textit{in } \Omega, \\ y = 0 & \textit{on } \Gamma \end{cases} \tag{2.5}$$

has a unique solution $y \in H_0^1(\Omega) \cap C^{0,\alpha}(\overline{\Omega})$ *verifying*

$$\|y\|_{H_0^1(\Omega)} + \|y\|_{L^\infty(\Omega)} \leq C_2 \|b\|_{L^s(\Omega)}, \tag{2.6}$$

$$\|y\|_{C^{0,\alpha}(\overline{\Omega})} \leq C_3 \|b\|_{L^s(\Omega)} \left(1 + \|a\|_{L^s(\Omega)}\right). \tag{2.7}$$

This result follows from classical estimations in the spaces $C^{0,\alpha}(\overline{\Omega})$ (Gilbarg and Trudinger [17] or Stampacchia [22]); see Bonnans and Casas [7] for details. Now we prove Theorem 2.1.

Proof. The first part of the theorem is also proved in Bonnans and Casas [7]. Let us prove the continuity of $u \longrightarrow y_u$. Let $\{u_k\}_{k=1}^\infty \subset \mathcal{K}$ be a sequence converging to $u \in \mathcal{K}$, i.e., $d(u_k, u) \to 0$. Denote by y_k and y the states corresponding to u_k and u, respectively. From (2.3) and (2.4) we deduce the existence of $M > 0$ such that

$$b_k(x) = f(x, y_k(x), u(x)) - f(x, y_k(x), u_k(x))$$

satisfies

$$\|b_k\|_{L^s(\Omega)} = \left(\int_{\{x \,:\, u_k(x) \neq u(x)\}} |f(x, y_k(x), u(x)) - f(x, y_k(x), u_k(x))|^s \, dx \right)^{1/s} \to 0.$$

Now applying the mean value theorem we get for some function $\theta_k : \overline{\Omega} \longrightarrow (0, 1)$

$$A(y - y_k) - \frac{\partial f}{\partial y}(x, y + \theta_k(y - y_k), u(x))(y - y_k) = b_k.$$

Therefore, from Lemma 2.2 and using (2.3) we obtain

$$\|y-y_k\|_{H_0^1(\Omega)}+\|y-y_k\|_{C^{0,\alpha}(\overline{\Omega})}\le M'\|b_k\|_{L^s(\Omega)}\to 0,$$

which completes the proof. □

Remark 2.3. In order to use Lemma 2.2 in the above proof we have to check that

$$x\longrightarrow \frac{\partial f}{\partial y}(x,y(x)+\theta_k(x)(y(x)-y_k(x)),u(x))$$

is a measurable function. Although $\theta_k(x)$ itself might be nonmeasurable, this is true because by definition of θ_k as

$$\frac{\partial f}{\partial y}(x,y(x)+\theta_k(x)(y(x)-y_k(x)),u(x))$$
$$=\begin{cases}\dfrac{f(x,y(x),u(x))-f(x,y_k(x),u(x))}{y(x)-y_k(x)} & \text{if } y(x)\ne y_k(x),\\[2ex] \dfrac{\partial f}{\partial y}(x,y(x),u(x)) & \text{if } y(x)=y_k(x).\end{cases}$$

We finish this section by proving a lemma that will be used several times in this paper. First let us introduce some notation. In the sequel $M(\Omega)$ will denote the space of real regular Borel measures in Ω, which is identified with the dual space of $C_0(\Omega)$, the space formed by the real continuous functions defined in $\overline{\Omega}$ and vanishing on Γ. Let $A^\star$ denote the formal adjoint operator of A:

$$A^\star y=-\sum_{i,j=1}^{n}\partial_{x_j}\left(a_{ji}(x)\partial_{x_i}y(x)\right).$$

LEMMA 2.4. *For every function $a\in L^s(\Omega)$, with $a(x)\ge 0$ a.e. $x\in\Omega$, and every Borel measure $\mu\in M(\Omega)$ there exists a unique solution in $W_0^{1,\sigma}(\Omega)$, for all $\sigma<n/(n-1)$, of problem*

$$\begin{cases}A^\star p+ap=\mu & \text{in } \Omega,\\ y=0 & \text{on } \Gamma.\end{cases}$$

Moreover there exists a constant $M>0$ independent of a such that

$$\|p\|_{W_0^{1,\sigma}(\Omega)}\le M\|\mu\|_{M(\Omega)}. \tag{2.8}$$

Proof. The existence and uniqueness in $W_0^{1,\sigma}(\Omega)$ of solution p of the above Dirichlet problem is well known; see Stampacchia [22] or Casas [12]. Let us prove (2.8). Let t be the conjugate of σ, $1/t+1/\sigma=1$, thus $t>n$. For every $\psi\in W^{-1,t}(\Omega)=(W_0^{1,s}(\Omega))'$, the equation

$$\begin{cases}A\varphi+a\varphi=\psi & \text{in } \Omega,\\ \varphi=0 & \text{on } \Gamma\end{cases}$$

has a unique solution in $H_0^1(\Omega)\cap C_0(\Omega)$, and proceeding as in [7, Lemma 3.2], there exists $M>0$ independent of a such that

$$\|\varphi\|_\infty\le M\|\psi\|_{W^{-1,t}(\Omega)}.$$

Hence

$$\left|\int_\Omega p\psi\,dx\right| = \left|\int_\Omega p(A\varphi + a\varphi)\,dx\right| = \left|\int_\Omega \varphi\,d\mu\right|$$
$$\leq \|\varphi\|_\infty \|\mu\|_{M(\Omega)} \leq M\|\psi\|_{W^{-1,t}(\Omega)}\|\mu\|_{M(\Omega)},$$

which proves the desired inequality. □

3. The weak and strong Pontryagin's minimum principle. In this section we present the statements of the weak and strong Pontryagin's principles. First let us introduce some notation and definitions.

DEFINITION 3.1. *We will say that problem (P_δ) is weakly stable on the right if*

$$\lim_{\delta' \searrow \delta} \inf(P_{\delta'}) = \inf(P_\delta), \tag{3.9}$$

and weakly stable on the left if

$$\lim_{\delta' \nearrow \delta} \inf(P_{\delta'}) = \inf(P_\delta). \tag{3.10}$$

(P_δ) is said to be strongly stable on the right (respectively, left) if there exist $\epsilon > 0$ and $r > 0$ such that:

$$\inf(P_\delta) - \inf(P_{\delta'}) \leq r(\delta' - \delta) \quad \forall\delta' \in [\delta, \delta + \epsilon], \tag{3.11}$$

respectively,

$$\inf(P_{\delta'}) - \inf(P_\delta) \leq r(\delta - \delta') \quad \forall\delta' \in [\delta - \epsilon, \delta]. \tag{3.12}$$

If (P_δ) is weakly (respectively, strongly) stable on the left and on the right, it will be called weakly (respectively, strongly) stable.

Sufficient conditions for the weak stability were given by Casas [14] under additional regularity hypotheses on the functions L and f. In particular, if they are continuous with respect to the third variable, L is convex with respect to the same variable, K is convex and closed, and (P_{δ_0}) has a feasible pair (y, u), then (P_δ) is stable on the right for every $\delta > \delta_0$. In spite of these results, in general it is difficult to establish the stability of a problem, mainly the strong stability. However most of problems (P_δ) are weak and strongly stable. More precisely, we get the following proposition.

PROPOSITION 3.2. *Let us denote by δ_0 a real number such that (P_{δ_0}) has at least one feasible pair (y, u). Then for every $\delta \geq \delta_0$, except at most a countable number of them (respectively, a set of zero measure), the problem (P_δ) is weakly (respectively, strongly) stable.*

Proof. If we define $\phi : [\delta_0, +\infty) \longrightarrow R$ by $\phi(\delta) = \inf(P_\delta)$, then ϕ is a decreasing monotone function and therefore ϕ is continuous (respectively, differentiable) at each point except at most a countable number of them (respectively, a set of zero measure). Finally it is obvious that the continuity (respectively, differentiability) of ϕ at δ implies the weak (respectively, strong) stability of (P_δ). □

Given a number $\alpha \geq 0$, we define the Hamiltonian associated to (P_δ) by

$$H_\alpha(x, y, u, p) = \alpha L(x, y, u) + pf(x, y, u).$$

If $\alpha = 1$, we simply write H instead of H_1. Now we can formulate the following theorems.

THEOREM 3.3 (Weak Pontryagin's principle). *Let $\overline{u}$ be a solution of (P_δ) in $(\mathcal{K}, d)$, with $\overline{y}$ its associated state. If (P_δ) is weakly stable on the right, then there exist $\overline{\alpha} \geq 0$, $\overline{p} \in W_0^{1,\sigma}(\Omega)$ for every $\sigma < \frac{n}{n-1}$ and $\overline{\mu} \in M(\Omega)$ such that*

$$\overline{\alpha} + \|\overline{\mu}\|_{M(\Omega)} > 0, \tag{3.13}$$

$$\begin{cases} A^\star \overline{p} = \dfrac{\partial f}{\partial y}(x, \overline{y}(x), \overline{u}(x)) + \overline{\alpha}\dfrac{\partial L}{\partial y}(x, \overline{y}(x), \overline{u}(x)) + \dfrac{\partial g}{\partial y}(x, \overline{y}(x))\overline{\mu} & \text{in } \Omega, \\ \overline{p} = 0 \quad \text{on } \Gamma, \end{cases} \tag{3.14}$$

$$\int_\Omega (z(x) - g(x, \overline{y}(x)))d\overline{\mu}(x) \leq 0 \quad \forall z \in C_0(\Omega) \text{ with } z(x) \leq \delta \ \forall x \in \Omega, \tag{3.15}$$

and for every $v \in K$

$$H_{\overline{\alpha}}(x, \overline{y}(x), \overline{u}(x), \overline{p}(x)) \leq H_{\overline{\alpha}}(x, \overline{y}(x), v, \overline{p}(x)) \quad a.e.\ x \in \Omega. \tag{3.16}$$

Moreover if there exists a Lebesgue measurable set $\Omega_0 \subset \Omega$, with $m(\Omega_0) = m(\Omega)$, of such a kind that one of the two following conditions is satisfied

(H1) *for each $y \in C_0(\Omega)$ and $\forall v \in K$ the set of Lebesgue points of the functions $x \longrightarrow f(x, y(x), v)$ and $x \longrightarrow L(x, y(x), v)$ contains Ω_0,*

(H2) *the functions L and f are continuous with respect to the third variable for every $x \in \Omega_0$,*

then

$$H_{\overline{\alpha}}(x, \overline{y}(x), \overline{u}(x), \overline{p}(x)) = \min_{v \in K} H_{\overline{\alpha}}(x, \overline{y}(x), v, \overline{p}(x)) \quad a.e.\ x \in \Omega. \tag{3.17}$$

THEOREM 3.4 (Strong Pontryagin's principle). *Under the assumptions of Theorem* 3.3 *and assuming that (P_δ) is strongly stable on the right, there exist $\overline{p} \in W_0^{1,\sigma}(\Omega)$ and $\overline{\mu} \in M(\Omega)$ satisfying* (3.14)–(3.16), *or* (3.17) *if the conditions* (H1) *or* (H2) *hold, with $\overline{\alpha} = 1$.*

A first version of these theorems (with stronger hypotheses) was given by Bonnans in [3] and [4]. Since we will use penalization techniques to prove these theorems, the stability on the right is the proper condition to obtain the desired result. However the Slater condition, which is a stability condition on the left, is the usual hypothesis to derive the optimality conditions (different of Pontryagin's principle) in a qualified form; see Bonnans and Casas [6]. Weak stability on the left also was the assumption in [13] to prove the convergence of the numerical approximations.

4. Hamiltonian formulation of the cost variation. In this section we generalize some results of [7] that we will use later. Let us denote by $h : \overline{\Omega} \times R \longrightarrow R$ and $\phi : R \longrightarrow R$ two functions satisfying the condition that ϕ is of class C^1 and h is continuous, differentiable with respect to the second variable and $\frac{\partial h}{\partial y} \in C(\overline{\Omega} \times R)$. Now we consider the functional

$$\hat{J}(y, u) = \int_\Omega L(x, y(x), u(x))\, dx + \phi\left(\int_\Omega h(x, y(x))\, dx\right).$$

We are interested in studying this type of functionals because it plays an important role in the proof of Pontryagin's principle, the second term being particularized later to some penalization of state constraints. As in the previous section

$$H(x, y, u, p) = L(x, y, u) + pf(x, y, u).$$

In the first part of this section we will assume that the following regularity condition holds:

$$\left|\frac{\partial L}{\partial y}(x,y,u)\right| \le M_3(x) + \eta(|y|) \quad \forall (x,y,u) \in \Omega \times R \times K, \tag{4.18}$$

with $M_3 \in L^s(\Omega)$.

Let $u, v \in \mathcal{K}$ be two controls and y_u and y_v the associated states. From the mean value theorem it follows that there exist the intermediate states $\check{y}$, $\hat{y}$, and $\tilde{y}$ satisfying

$$\phi\left(\int_\Omega h(x,y_v(x))\,dx\right) = \phi\left(\int_\Omega h(x,y_u(x))\,dx\right)$$

$$+\phi'\left(\int_\Omega h(x,\check{y}(x))\,dx\right)\int_\Omega \frac{\partial h}{\partial y}(x,\check{y}(x))(y_v(x)-y_u(x))\,dx,$$

$$f(\cdot,y_v,v) = f(\cdot,y_u,v) + \frac{\partial f}{\partial y}(\cdot,\hat{y},v)(y_v - y_u),$$

$$L(\cdot,y_v,v) = L(\cdot,y_u,v) + \frac{\partial L}{\partial y}(\cdot,\tilde{y},v)(y_v - y_u),$$

with $\check{y}(x), \hat{y}(x), \tilde{y}(x) \in [y_u(x), y_v(x)]\ \forall x \in \overline{\Omega}$. Since y_u and y_v are bounded, it follows that $\tilde{y}, \hat{y}$, and $\tilde{y}$ also are bounded. Now we define the intermediate adjoint state $p_{u,v}$ as the solution of

$$\left\{\begin{array}{l} A^\star p_{u,v} = \frac{\partial f}{\partial y}(\cdot,\hat{y},v)p_{u,v} + \frac{\partial L}{\partial y}(\cdot,\tilde{y},v) + \phi'\left(\int_\Omega h(x,\check{y})dx\right)\frac{\partial h}{\partial y}(x,\check{y}) \quad \text{in } \Omega, \\ p_{u,v} = 0 \quad \text{on}\Gamma. \end{array}\right. \tag{4.19}$$

Note that if $u = v$, then $\tilde{y}, \hat{y} = \tilde{y} = y_u$ and $p_{u,v} = p_u$ is the adjoint state associated to u. Let us verify that (4.19) is well posed.

LEMMA 4.1. *If* (4.18) *holds, then equation* (4.19) *has a unique solution* $p_{u,v} \in H_0^1(\Omega) \cap C^{0,\alpha}(\overline{\Omega})$ *that moreover satisfies*

$$\|p_{u,v}\|_{H_0^1(\Omega)} + \|p_{u,v}\|_{C^{0,\alpha}(\overline{\Omega})} \le C_4 \quad \forall u, v \in \mathcal{K}. \tag{4.20}$$

Proof. The proof is a straightforward consequence of Lemma 2.2 and inequalities (2.3). □

Now we have the following Hamiltonian formulation of the cost variation.

PROPOSITION 4.2. *Assume that* (4.18) *is satisfied, and let* $u, v \in \mathcal{K}$ *and* $p_{u,v}$ *be the intermediate adjoint state associated. Then*

$$\hat{J}(y_v,v) = \hat{J}(y_u,u) + \int_\Omega [H(x,y_u(x),v(x),p_{u,v}(x)) - H(x,y_u(x),u(x),p_{u,v}(x))]\,dx.$$

Proof. We have

$$\hat{J}(y_v,v) - \hat{J}(y_u,u) = \int_\Omega [L(x,y_u(x),v(x)) - L(x,y_u(x),u(x))]dx$$

$$+\int_\Omega [L(x,y_v(x),v(x)) - L(x,y_u(x),v(x))]\,dx$$

$$+\phi\left(\int_\Omega h(x,y_v(x))\,dx\right) - \phi\left(\int_\Omega h(x,y_u(x))\,dx\right).$$

From (4.19) we deduce

$$\int_\Omega [L(x,y_v(x),v(x)) - L(x,y_u(x),v(x))]\,dx$$

$$+\phi\left(\int_\Omega h(x,y_v(x))dx\right) - \phi\left(\int_\Omega h(x,y_u(x))\,dx\right)$$

$$= \int_\Omega \frac{\partial L}{\partial y}(x,\tilde{y}(x),v(x))(y_v(x) - y_u(x))\,dx$$

$$+\phi'\left(\int_\Omega h(x,\check{y}(x))dx\right)\int_\Omega \frac{\partial h}{\partial y}(x,\check{y}(x))(y_v(x) - y_u(x))\,dx$$

$$= \int_\Omega \left[A^\star p_{u,v} - \frac{\partial f}{\partial y}(x,\hat{y}(x),v(x))p_{u,v}\right](y_v - y_u)\,dx$$

$$= \int_\Omega A(y_v - y_u)p_{u,v}dx - \int_\Omega \frac{\partial f}{\partial y}(x,\hat{y}(x),v(x))p_{u,v}(y_v - y_u)\,dx$$

$$= \int_\Omega [f(x,y_v(x),v(x)) - f(x,y_u(x),u(x))]p_{u,v}\,dx$$

$$+\int_\Omega [f(x,y_u(x),v(x)) - f(x,y_v(x),v(x))]p_{u,v}\,dx$$

$$= \int_\Omega [f(x,y_u(x),v(x)) - f(x,y_u(x),u(x))]p_{u,v}\,dx,$$

which proves the proposition. □

PROPOSITION 4.3. *Assume that* (4.18) *holds, and let* $\{v_k\}_{k=1}^\infty \subset \mathcal{K}$ *be a sequence converging to u in the topology defined by Ekeland's distance. Then the states and the adjoint states associated* $y_k = y_{v_k}$ *and* $p_k = p_{u,v_k}$ *converge to* y_u *and* p_u*, respectively, in* $H_0^1(\Omega) \cap C^{0,\alpha}(\overline{\Omega})$.

Proof. The convergence $y_k \to y_u$ follows from Theorem 2.1. The convergence of $\{p_k\}$ follows from the continuity of $v \in \mathcal{K} \longrightarrow p_{u,v} \in H_0^1(\Omega) \cap C^{0,\alpha}(\overline{\Omega})$, which can be proved arguing in a similar way to the proof of Theorem 2.1. □

Given a point $x_0 \in \Omega$, we will denote

$$\omega_k(x_0) = \{x \in \Omega : \|x - x_0\| \leq 1/k\}$$

and $m_k(x_0) = m(\omega_k(x_0))^{-1}$. We will say that a sequence $\{v_k\}$ in $\mathcal{K}$ is a spike perturbation of $u \in \mathcal{K}$ around x_0 associated to $v \in K$ if

$$v_k(x) = \begin{cases} v & \text{if } x \in \omega_k(x_0), \\ u(x) & \text{otherwise.} \end{cases}$$

PROPOSITION 4.4. *Assume that (4.18) holds, let v_k be a spike perturbation of u around x_0 associated to $v \in K$, and let y_k be the associated state. Then for every $u \in \mathcal{K}$ there exists a set $\Omega(u,v) \subset \Omega$, with $m(\Omega(u,v)) = m(\Omega)$, such that*

$$\lim_{k \to \infty} m_k(x_0)[\hat{J}(y_k, v_k) - \hat{J}(y_u, u)]$$

$$= H(x_0, y_u(x_0), v, p_u(x_0)) - H(x_0, y_u(x_0), u(x_0), p_u(x_0)) \quad \forall x_0 \in \Omega(u,v).$$

Proof. From Proposition 4.2 we have

$$\hat{J}(y_k, v_k) - \hat{J}(y_u, u) = \int_{\omega_k(x_0)} [H(x, y_u(x), v, p_k(x)) - H(x, y_u(x), u(x), p_k(x))]\, dx,$$

where $p_k = p_{u,v_k}$ converges to p_u in $H_0^1(\Omega) \cap C^{0,\alpha}(\overline{\Omega})$ as stated in Proposition 4.3. Then

$$\hat{J}(y_k, v_k) - \hat{J}(y_u, u) = \int_{\omega_k(x_0)} [H(x, y_u(x), v, p_u(x)) - H(x, y_u(x), u(x), p_u(x))]\, dx$$

$$+ \int_{\omega_k(x_0)} f(x, y_u(x), v)(p_k(x) - p_u(x))\, dx + \int_{\omega_k(x_0)} f(x, y_u(x), u)(x)(p_u(x) - p_k(x))\, dx.$$

Let $\Omega(u,v)$ be the intersection of the Lebesgue points of the following mappings:

$$\begin{aligned} x &\longrightarrow f(x, y_u(x), v), \\ x &\longrightarrow f(x, y_u(x), u(x)), \\ x &\longrightarrow H(x, y_u(x), v, p_u(x)), \\ x &\longrightarrow H(x, y_u(x), u(x), p_u(x)). \end{aligned}$$

Then $m(\Omega(u,v)) = m(\Omega)$. Now using the uniform convergence of $p_k \to p_u$, it follows that for every $x_0 \in \Omega(u,v)$

$$m_k(x_0) \left| \int_{\omega_k(x_0)} f(x, y_u(x), v)(p_u(x) - p_k(x))\, dx \right|$$

$$\leq m_k(x_0) \int_{\omega_k(x_0)} |f(x, y_u(x), v)|\, dx \|p_u - p_k\|_{L^\infty(\Omega)} \to 0.$$

Analogously

$$m_k(x_0) \left| \int_{\omega_k(x_0)} f(x, y_u(x), u(x))(p_u(x) - p_k(x))\, d\dot{x} \right|$$

$$\leq m_k(x_0)\int_{\omega_k(x_0)}|f(x,y_u(x),u(x))|\,dx\|p_u-p_k\|_{L^\infty(\Omega)}\to 0.$$

Therefore we deduce that

$$\lim_{k\to\infty} m_k(x_0)[\hat{J}(y_k,v_k)-\hat{J}(y_u,u)]$$

$$=\lim_{k\to\infty} m_k(x_0)\int_{\omega_k(x_0)}[H(x,y_u(x),v,p_u(x))-H(x,y_u(x),u(x),p_u(x))]\,dx$$

$$=H(x_0,y_u(x_0),v,p_u(x_0))-H(x_0,y_u(x_0),u(x_0),p_u(x_0)),$$

which concludes the proof. □

The last proposition allows us to deduce easily Pontryagin's principle for control problems without state constraints. In fact it is enough to suppose that $\overline{u}\in\mathcal{K}$ is a stationary point to derive a minimum principle.

DEFINITION 4.5. *We say that $\overline{u}$ is a stationary point of the control problem*

$$(P)\begin{cases}\min \hat{J}(y,u)\\ (y,u)\ \textit{satisfies}\ (2.1)\ \textit{and}\ \ u(x)\in K\ \textit{a.e.}\ x\in\Omega\end{cases}$$

if

$$\liminf_{d(u,\overline{u})\to 0}\frac{\hat{J}(y,u)-\hat{J}(\overline{y},\overline{u})}{d(u,\overline{u})}\geq 0.$$

Obviously, every local solution in $(\mathcal{K},d)$ is a stationary point. Now we can prove the following proposition

PROPOSITION 4.6. *Let us suppose that* (4.18) *holds, and let $\overline{u}$ be a stationary point of (P). Then for every $v\in K$*

$$H(x,\overline{y}(x),\overline{u}(x),\overline{p}(x))\leq H(x,\overline{y}(x),v,\overline{p}(x))\quad \textit{a.e. } x\in\Omega,$$

where $\overline{y}$ and $\overline{p}$ are the state and adjoint state associated to $\overline{u}$. Moreover, if condition (H1) *or* (H2) *is verified, then $\overline{u}$ satisfies Pontryagin's principle:*

$$H(x,\overline{y}(x),\overline{u}(x),\overline{p}(x))=\min_{v\in K}H(x,\overline{y}(x),v,\overline{p}(x))\quad \textit{a.e. } x\in\Omega.$$

Proof. The first part of the proof is an immediate consequence of Proposition 4.4; it is enough to remark that $d(v_k,\overline{u})\leq m_k(x_0)^{-1}$. To derive Pontryagin's principle under condition (H1) we use the fact that the set $\Omega(\overline{u},v)$, defined in the proof of Proposition 4.4, contains the intersection of Ω_0 and the set of Lebesgue points of the functions:

$$\begin{aligned}x&\longrightarrow f(x,\overline{y}(x),\overline{u}(x)),\\ x&\longrightarrow H(x,\overline{y}(x),\overline{u}(x),\overline{p}(x)).\end{aligned}$$

Indeed the continuity of $\overline{p}$ and condition (H1) imply that Ω_0 is a subset of the Lebesgue point set of the functions:

$$\begin{aligned}x&\longrightarrow f(x,\overline{y}(x),v),\\ x&\longrightarrow H(x,\overline{y}(x),v,\overline{p}(x)).\end{aligned}$$

Therefore

$$H(x,\overline{y}(x),\overline{u}(x),\overline{p}(x)) \leq H(x,\overline{y}(x),v,\overline{p}(x)) \ \ \forall x \in \Omega_0 \text{ and } \forall v \in K.$$

Then

$$H(x,\overline{y}(x),\overline{u}(x),\overline{p}(x)) = \min_{v\in K} H(x,\overline{y}(x),v,\overline{p}(x)) \ \ \forall x \in \Omega_0.$$

In the case of (H2), let us take a sequence $\{v_k\}_{k=1}^{\infty}$ dense in K and $\Omega_1 = \cap_k \Omega(\overline{u}, v_k)$. Then

$$H(x,\overline{y}(x),\overline{u}(x),\overline{p}(x)) \leq H(x,\overline{y}(x),v_k,\overline{p}(x)) \ \ \forall x \in \Omega_1 \text{ and } \forall k.$$

Finally the continuity of the Hamiltonian with respect to the control and the last inequality imply Pontryagin's principle in the points $x \in \Omega_0 \cap \Omega_1$. □

We now get rid of the regularity hypothesis (4.18).

PROPOSITION 4.7. *Let $\overline{u}$ be a local solution of* (P). *Then the conclusions of Proposition* 4.6 *remain true without hypothesis* (4.18).

To prove this proposition we will use Ekeland's principle.

THEOREM 4.8 (Ekeland [16]). *Let (E,d) be a complete metric space, $F : E \longrightarrow R \cup \{+\infty\}$ be a lower semicontinuous function, and let $e_\epsilon \in E$ satisfy*

$$F(e_\epsilon) \leq \inf_{e\in E} F(e) + \epsilon^2.$$

Then there exists an element $\overline{e}_\epsilon \in E$ such that

$$F(\overline{e}_\epsilon) \leq F(e_\epsilon), \quad d(\overline{e}_\epsilon, e_\epsilon) \leq \epsilon,$$

and

$$F(\overline{e}_\epsilon) \leq F(e) + \epsilon d(e, \overline{e}_\epsilon) \ \ \forall e \in E.$$

Now we proceed to prove Proposition 4.7.

Proof. The idea is to regularize L and to check that $\overline{u}$ is an approximate solution of the regularized problem. Using Ekeland's principle we get some optimality conditions as in Proposition 4.7 and finally pass to the limit in these optimality conditions and get the desired result.

The regularization is as follows. By Proj_ϵ we denote the projection onto the segment $[-1/\epsilon, +1/\epsilon]$, i.e.,

$$\text{Proj}_\epsilon(t) = \max\{-1/\epsilon, \min\{t, +1/\epsilon\}\}.$$

We define

$$\varphi_\epsilon(x,t,u) = \text{Proj}_\epsilon\left(\frac{\partial L}{\partial y}(x,t,u)\right)$$

and

$$L_\epsilon(x,t,u) = L(x,0,u) + \int_0^t \varphi_\epsilon(x,t,u)\,dt.$$

We can now state the problem

$$(P^\epsilon)\begin{cases} \min \hat{J}_\epsilon(y,u) = \displaystyle\int_\Omega L_\epsilon(x,y(x),u(x))\,dx + \phi\left(\int_\Omega h(x,y(x))\,dx\right) \\ (y,u) \text{ satisfies (2.1)}, \ u(x) \in K \text{ a.e. } x \in \Omega. \end{cases}$$

We claim that $\inf(P^\epsilon) \to \inf(P)$ when $\epsilon \searrow 0$. To prove this it is enough to check that

$$(4.21) \qquad |\hat{J}_\epsilon(y,u) - \hat{J}(y,u)| \le r_\epsilon \quad \forall u \in K \text{ and } (y,u) \text{ satisfying (2.1)},$$

with $r_\epsilon \searrow 0$ when $\epsilon \searrow 0$ and r_ϵ independent of u. Indeed, if $u_\epsilon \in \mathcal{K}$ satisfies that $\hat{J}_\epsilon(y_\epsilon,u_\epsilon) \le \inf(P_\epsilon) + \varepsilon$, y_ϵ being the state associated with u_ϵ, then by (4.21)

$$\liminf_{\epsilon\searrow 0} \inf(P_\epsilon) \ge \liminf_{\epsilon\searrow 0}(\hat{J}_\epsilon(y_\epsilon,u_\epsilon) - \epsilon) \ge \liminf_{\epsilon\searrow 0}(\hat{J}_\epsilon(\overline{y},\overline{u}) - r_\epsilon) = \inf(P)$$

and also

$$\limsup_{\epsilon\searrow 0} \inf(P_\epsilon) \le \limsup_{\epsilon\searrow 0} \hat{J}_\epsilon(\overline{y},\overline{u}) \le \limsup_{\epsilon\searrow 0}[\hat{J}(\overline{y},\overline{u}) + r_\epsilon] = \inf(P),$$

which proves that $\inf(P_\epsilon) \to \inf(P)$, as desired. Now let us check that (4.21) holds. Indeed

$$L(x,y,u) - L_\epsilon(x,y,u) = \int_0^y \left[\frac{\partial L}{\partial y}(x,t,u) - \varphi_\epsilon(x,t,u)\right] dt.$$

Let $M > 0$ be such that $|y(x)| \le M$ whenever (y,u) is solution of (2.1) and $u \in \mathcal{K}$. Then

$$|J_\epsilon(y,u) - J(y,u)| \le \int_\Omega \int_{-M}^{+M} \left|\varphi_\epsilon(x,t,u(x)) - \frac{\partial L}{\partial y}(x,t,u(x))\right| dt\,dx$$

$$= \int_{-M}^{+M} \int_\Omega \left|\varphi_\epsilon(x,t,u(x)) - \frac{\partial L}{\partial y}(x,t,u(x))\right| dx\,dt$$

$$\le 2M \int_\Omega \sup_{|t|\le M,\ v\in K} \left|\varphi_\epsilon(x,t,u(x)) - \frac{\partial L}{\partial y}(x,t,u(x))\right| dx.$$

Put

$$\Omega_\epsilon = \left\{ x \in \Omega : \sup_{|t|\le M, v\in K} \left|\frac{\partial L}{\partial y}(x,t,u(x))\right| \ge \frac{1}{\epsilon} \right\}.$$

As

$$\left|\frac{\partial L}{\partial y}(x,t,u(x))\right| \le M_2 \in L^1(\Omega),$$

it follows that $m(\Omega_\epsilon) \searrow 0$ as $\epsilon \searrow 0$ and

$$\left|\hat{J}_\epsilon(y,u) - \hat{J}(y,u)\right| \le r_\epsilon = 2M \int_{\Omega_\epsilon} M_2(x)dx \to 0.$$

As a consequence of the previous results, we can get a family of real numbers $\{\delta_\epsilon\}_{\epsilon>0}$, with $\delta_\epsilon \searrow 0$ when $\epsilon \searrow 0$, such that

$$\hat{J}_\epsilon(\overline{y},\overline{u}) \leq \inf(P_\epsilon) + \delta_\epsilon^2.$$

Therefore we can apply Ekeland's principle with $F(u) = \hat{J}_\epsilon(y_u, u)$ defined in the complete metric space $(\mathcal{K}, d)$ and deduce the existence of a control $u_\epsilon \in \mathcal{K}$ such that $d(\overline{u}, u_\epsilon) \leq \delta_\epsilon$ and

$$\hat{J}_\epsilon(y_\epsilon, u_\epsilon) \leq \hat{J}_\epsilon(y_u, u) + \delta_\epsilon d(u, u_\epsilon) \quad \forall u \in \mathcal{K}, \tag{4.22}$$

where y_ϵ is the state associated with u_ϵ. To apply Proposition 4.6 we must put the cost given by the right-hand side of inequality (4.22) into the framework of this proposition. For it we introduce the function $\chi_\epsilon : \Omega \times K \longrightarrow R$ by

$$\chi_\epsilon(x, v) = \begin{cases} 0 & \text{if } v = u_\epsilon(x), \\ 1 & \text{otherwise.} \end{cases}$$

Then (y_ϵ, u_ϵ) is the solution of the problem

$$(Q^\epsilon) \begin{cases} \min \hat{J}_\epsilon(y,u) = \hat{J}_\epsilon(y,u) + \delta_\epsilon \int_\Omega \chi_\epsilon(x, u(x)) \\ \\ (y,u) \text{ satisfies (2.1)}, \ u(x) \in K \text{ a.e. } x \in \Omega. \end{cases}$$

Then Proposition 4.6 implies that for every $v \in K$

$$H^\epsilon(x, y_\epsilon(x), u_\epsilon(x), p_\epsilon(x)) \leq H^\epsilon(x, y_\epsilon(x), v, p_\epsilon(x)) + \delta_\epsilon \quad \text{a.e. } x \in \Omega,$$

where p_ϵ is the adjoint state

$$\begin{cases} A^\star p_\epsilon = \dfrac{\partial f}{\partial y}(x, y_\epsilon, u_\epsilon)p_\epsilon + \dfrac{\partial L_\epsilon}{\partial y}(x, y_\epsilon, u_\epsilon) \text{ in } \Omega, \\ \\ p_\epsilon = 0 \text{ on } \Gamma \end{cases} \tag{4.23}$$

and

$$H^\epsilon(x, y, u, p) = L_\epsilon(x, y, u) + pf(x, y, u) + \delta_\epsilon \chi_\epsilon(x, u). \tag{4.24}$$

From Theorem 2.1 it follows that $y_\epsilon \rightarrow \overline{y}$ in $H_0^1(\Omega) \cap C^{0,\alpha}(\overline{\Omega})$. Then, thanks to hypothesis (2.3) and the definition of L_ϵ, we have

$$f(x, y_\epsilon, u_\epsilon) \rightarrow f(x, \overline{y}, \overline{u}) \quad \text{and} \quad \frac{\partial f}{\partial y}(x, y_\epsilon, u_\epsilon) \rightarrow \frac{\partial f}{\partial y}(x, \overline{y}, \overline{u}) \quad \text{in } L^s(\Omega),$$

$$L_\epsilon(x, y_\epsilon, u_\epsilon) \rightarrow L(x, \overline{y}, \overline{u}) \quad \text{and} \quad \frac{\partial L_\epsilon}{\partial y}(x, y_\epsilon, u_\epsilon) \rightarrow \frac{\partial L}{\partial y}(x, \overline{y}, \overline{u}) \quad \text{in } L^1(\Omega).$$

With the aid of these relations and Lemma 2.4 we can pass to the limit in (4.23) and (4.24) and deduce the first conclusion of Proposition 4.6. To prove the second conclusion, i.e., the Pontryagin's principles, we argue as follows. Under condition (H2), the argument used in the proof of Proposition 4.6 can be repeated here without modifications. If condition (H1) is satisfied, then, thanks to Proposition 4.6, we can take a sequence $\epsilon_j \searrow 0$ and $\delta_j = \delta_{\epsilon_j} \searrow 0$ such that

$$H^{\epsilon_j}(x, y_{\epsilon_j}(x), u_{\epsilon_j}(x), p_{\epsilon_j}(x)) \leq \min_{v\in K} H^{\epsilon_j}(x, y_{\epsilon_j}(x), v, p_{\epsilon_j}(x)) + \delta_j \quad \text{a.e.} \quad x \in \Omega_j,$$

with $m(\Omega_j) = m(\Omega)$. Now we pass to the limit and get

$$H(x, \overline{y}(x), \overline{u}(x), \overline{p}(x)) = \min_{v\in K} H(x, \overline{y}(x), v, \overline{p}(x)) \quad \text{a.e.} \quad x \in \tilde{\Omega} = \bigcap_{j=1}^{\infty} \Omega_j,$$

which concludes the proof because $m(\tilde{\Omega}) = m(\Omega)$. □

THEOREM 4.9. *The statement of Proposition* 4.6 *is still valid without hypothesis* (4.18).

Proof. From the definition of stationary point we deduce that for every $\epsilon > 0$ there exists $r_\epsilon > 0$ such that

$$\frac{\hat{J}(y,u) - \hat{J}(\overline{y},\overline{u})}{d(u,\overline{u})} \geq -\epsilon \quad \forall u \in B_{r_\epsilon}(\overline{u}),$$

where $B_{r_\epsilon}(\overline{u})$ is the open ball of $(\mathcal{K}, d)$ of radius r_ϵ and center at $\overline{u}$. Hence

$$\hat{J}(\overline{y},\overline{u}) \leq \hat{J}(y,u) + \epsilon \int_\Omega \chi(x, u(x))dx \quad \forall u \in B_{r_\epsilon}(\overline{u}),$$

with

$$\chi(x,v) = \begin{cases} 0 & \text{if } v = \overline{u}(x), \\ 1 & \text{otherwise.} \end{cases}$$

Then it is enough to apply Proposition 4.7 to the problem

$$\begin{cases} \min \hat{J}(y,u) + \epsilon \displaystyle\int_\Omega \chi(x, u(x))\, dx, \\ u \in B_{r_\epsilon}(\overline{u}), \; u(x) \in K, \end{cases}$$

and pass to the limit when $\epsilon \searrow 0$ to deduce the desired result. □

The hypotheses made about K, L, and f do not allow one to assure the existence of a solution of control problem (P). Here we will prove a principle of Pontryagin's type for ϵ-solutions.

DEFINITION 4.10. *A control* $u \in \mathcal{K}$ *is called an* ϵ*-solution of* (P) *if* $\hat{J}(y_u, u) \leq \inf(P) + \epsilon$.

THEOREM 4.11. *For every* $\epsilon > 0$ *there exists at least one* ϵ^2*-solution of* (P) *in* $\mathcal{K}$. *Furthermore for every* ϵ^2*-solution of* (P), $\overline{u}_\epsilon$, *there exists another* ϵ^2*-solution* u_ϵ *such that* $d(u_\epsilon, \overline{u}_\epsilon) \leq \epsilon$ *and for every* $v \in K$

$$H(x, y_\epsilon(x), u_\epsilon(x), p_\epsilon(x)) \leq H(x, y_\epsilon(x), v, p_\epsilon(x)) + \epsilon \quad a.e. \; x \in \Omega,$$

where $y_\epsilon = y_{u_\epsilon}$ *and* $p_\epsilon = p_{u_\epsilon}$. *Moreover, if there exists a Lebesgue measurable set* $\Omega_0 \subset \Omega$, *with* $m(\Omega_0) = m(\Omega)$, *in such a way that* (H1) *or* (H2) *holds, then* $\overline{u}$ *satisfies Pontryagin's principle*:

$$H(x, y_\epsilon(x), u_\epsilon(x), p_\epsilon(x)) = \min_{v\in K} H(x, y_\epsilon(x), v, p_\epsilon(x)) + \epsilon \quad a.e. \; x \in \Omega.$$

Proof. Thanks to hypothesis (2.3), we have that $\inf(P) \in R$. Therefore there exists at least one ϵ^2-solution of (P). Let $\overline{u}_\epsilon$ be one of them. Then we can apply Theorem 4.8, with $F(u) = \hat{J}(y_u, u)$ defined in the metric space $(\mathcal{K}, d)$, and deduce the existence of a control $u_\epsilon \in \mathcal{K}$, ϵ^2-solution of (P), such that $d(u_\epsilon, \overline{u}_\epsilon) \leq \epsilon$ and

$$\hat{J}(y_{u_\epsilon}, u_\epsilon) \leq \hat{J}(y_u, u) + \epsilon d(u, u_\epsilon) \ \ \forall u \in \mathcal{K}. \tag{4.25}$$

Now we put the cost into the framework of Proposition 4.7, using the function χ_ϵ as in its proof, and apply it to get the result. □

5. Proof of the weak minimum principle. Let $\overline{u}$ be a solution of (P_δ) and $\overline{y}$ its associated state. For every $\gamma > 0$ we define the problems

$$(Q_\gamma) \begin{cases} \min J_\gamma(y, u) = \displaystyle\int_\Omega \left[L(x, y(x), u(x)) + \frac{1}{2\gamma}((g(x, y(x)) - \delta)^+)^2 \right] dx \\ (y, u) \text{ satisfies (2.1) and } u(x) \in K \text{ a.e. } x \in \Omega. \end{cases}$$

The first issue to remark is the following.

PROPOSITION 5.1. *Let (P_δ) be weakly stable on the right. Then*

$$\lim_{\gamma \searrow 0} \inf(Q_\gamma) = \inf(P_\delta).$$

Proof. Let $\{u_\gamma\}$ be a family of γ-solutions of problems (Q_γ) and $\{y_\gamma\}$ be the associated states:

$$J(y_\gamma, u_\gamma) \leq \inf(Q_\gamma) + \gamma.$$

From the definition of (Q_γ) it follows that $(g(x, y_\gamma(x)) - \delta)^+ \to 0$ in $L^2(\Omega)$, which, together with (2.4) and the compactness of the inclusion $H_0^1(\Omega) \cap C^{0,\alpha}(\overline{\Omega}) \subset C_0(\Omega)$, implies the convergence $(g(x, y_\gamma(x)) - \delta)^+ \to 0$ in $C_0(\Omega)$. Therefore

$$\delta_\gamma = \|(g(x, y_\gamma(x)) - \delta)^+\|_{L^\infty(\Omega)} + \delta \to \delta \text{ if } \gamma \searrow 0.$$

As (y_γ, u_γ) is a feasible pair for (P_{δ_γ}) we deduce that

$$\inf(P_{\delta_\gamma}) \leq J(y_\gamma, u_\gamma) \leq \inf(Q_\gamma) + \gamma.$$

Then, using the weak stability of (P_δ) on the right, we obtain

$$\inf(P_\delta) = \lim_{\gamma \searrow 0} \inf(P_{\delta_\gamma}) \leq \lim_{\gamma \searrow 0} \{\inf(Q_\gamma) + \gamma\} = \lim_{\gamma \searrow 0} \inf(Q_\gamma) \leq \inf(P_\delta),$$

with the last inequality due to the fact that (y_u, u) is feasible for (Q_γ) whenever it is feasible for (P_δ), with the same cost. □

Proof of Theorem 3.3. Thanks to Proposition 5.1 we deduce that $\overline{u}$ is a ϵ_γ^2-solution of (Q_γ), with $\epsilon_\gamma \searrow 0$ when $\gamma \searrow 0$. Applying Theorem 4.11 we obtain the existence of a control $u_\gamma \in \mathcal{K}$, ϵ_γ^2-solution of (Q_γ), with $d(u_\gamma, \overline{u}) \leq \epsilon_\gamma$ and such that for every $v \in K$

$$H(x, y_\gamma(x), u_\gamma(x), p_\gamma(x)) \leq H(x, y_\gamma(x), v, p_\gamma(x)) + \epsilon_\gamma \ \text{ a.e. } x \in \Omega,$$

where $y_\gamma = y_{u_\gamma}$ and p_γ is the adjoint state:

$$\begin{cases} A^\star p_\gamma = \dfrac{\partial f}{\partial y}(x, y_\gamma, u_\gamma)p_\gamma + \dfrac{\partial L}{\partial y}(x, y_\gamma, u_\gamma) + \dfrac{1}{\gamma}(g(x, y_\gamma) - \delta)^+ \dfrac{\partial g}{\partial y}(x, y_\gamma) \text{ in } \Omega, \\ p_\gamma = 0 \text{ on } \Gamma. \end{cases}$$

Defining

$$\overline{\alpha}_\gamma = \left(1 + \|\frac{1}{\gamma}(g(x, y_\gamma) - \delta)^+\|_{M(\Omega)}\right)^{-1},$$

$$\overline{\mu}_\gamma = \frac{\overline{\alpha}_\gamma}{\gamma}(g(x, y_\gamma) - \delta)^+, \quad \text{and} \quad \overline{p}_\gamma = \overline{\alpha}_\gamma p_\gamma,$$

we get

$$\overline{\alpha}_\gamma + \|\overline{\mu}_\gamma\|_{M(\Omega)} = 1, \tag{5.26}$$

$$\begin{cases} A^\star \overline{p}_\gamma = \dfrac{\partial f}{\partial y}(x, y_\gamma, u_\gamma)\overline{p}_\gamma + \overline{\alpha}_\gamma \dfrac{\partial L}{\partial y}(x, y_\gamma, u_\gamma) + \overline{\mu}_\gamma \dfrac{\partial g}{\partial y}(x, y_\gamma) \text{ in } \Omega, \\ \overline{p}_\gamma = 0 \text{ on } \Gamma, \end{cases} \tag{5.27}$$

and for every $v \in K$

$$H_{\overline{\alpha}_\gamma}(x, y_\gamma(x), u_\gamma(x), \overline{p}_\gamma(x)) \leq H_{\overline{\alpha}_\gamma}(x, y_\gamma(x), v, \overline{p}_\gamma(x)) + \epsilon_\gamma \text{ a.e. } x \in \Omega. \tag{5.28}$$

If $\gamma \searrow 0$, then $d(u_\gamma, \overline{u}) \leq \epsilon_\gamma \to 0$; therefore from Theorem 2.1 we get $y_\gamma \to \overline{y}$ in $H_0^1(\Omega) \cap C^{0,\alpha}(\overline{\Omega})$. Now using (2.3), we deduce the convergences

$$f(x, y_\gamma, u_\gamma) \to f(x, \overline{y}, \overline{u}), \qquad \frac{\partial f}{\partial y}(x, y_\gamma, u_\gamma) \to \frac{\partial f}{\partial y}(x, \overline{y}, \overline{u}) \text{ in } L^s(\Omega),$$

$$L(x, y_\gamma, u_\gamma) \to L(x, \overline{y}, \overline{u}), \qquad \frac{\partial L}{\partial y}(x, y_\gamma, u_\gamma) \to \frac{\partial L}{\partial y}(x, \overline{y}, \overline{u}) \text{ in } L^s(\Omega).$$

Applying Lemma 2.4, we obtain the following estimation for $\overline{p}_\gamma$:

$$\|\overline{p}_\gamma\|_{W_0^{1,\sigma}(\Omega)} \leq M\|\overline{\alpha}_\gamma \frac{\partial L}{\partial y}(x, y_\gamma, u_\gamma) + \overline{\mu}_\gamma \frac{\partial g}{\partial y}(x, y_\gamma(x))\|_{M(\Omega)} \leq M' < +\infty$$

for every $\sigma < n/(n-1)$. Therefore, remembering (5.26), we can extract subsequences, denoted in the same way, such that $\overline{\mu}_\gamma \to \overline{\mu}$ in $M(\Omega)$ *weakly and $\overline{p}_\gamma \to \overline{p}$ weakly in $W_0^{1,\sigma}(\Omega)$. From Rellich's theorem (Adams [1]) it follows that $\overline{p}_\gamma \to \overline{p}$ strongly in $L^q(\Omega)$ for each $q < n/(n-2)$. Then

$$\frac{\partial f}{\partial y}(x, y_\gamma, u_\gamma)\overline{p}_\gamma \to \frac{\partial f}{\partial y}(x, \overline{y}, \overline{u})\overline{p} \text{ in } L^1(\Omega).$$

We can pass to the limit in (5.27) and derive (3.14). Relation (3.16) follows from (5.28). Relation (3.15) is obtained as follows: for every $z \in C_0(\Omega)$ with $z(x) \leq \delta$ for all $x \in \Omega$

$$\int_\Omega (z(x) - g(x, \overline{y}(x)))\, d\overline{\mu}(x) = \lim_{\gamma \to 0} \frac{\overline{\alpha}_\gamma}{\gamma} \int_\Omega (z(x) - g(x, y_\gamma(x)))(g(x, y_\gamma(x)) - \delta)^+ \, dx \leq 0.$$

It follows that

$$\langle \overline{\mu}, g(\cdot, \overline{y})\rangle = \max\{\langle \overline{\mu}, z\rangle : z \in C_0(\Omega), z(x) \leq \delta\};$$

hence, $\overline{\mu}$ is nonnegative and the value of the max is $\delta\|\overline{\mu}\|_{M(\Omega)}$.

To obtain (3.13) it is enough to remember (5.26) and remark that

$$\|\overline{\mu}\|_{M(\Omega)} = \frac{1}{\delta}\langle \overline{\mu}, g(\cdot, \overline{y})\rangle = \lim_{\gamma\to 0}\frac{1}{\delta}\langle \overline{\mu}_\gamma, g(\cdot, y_\gamma)\rangle = \lim_{\gamma\to 0}\|\overline{\mu}_\gamma\|_{M(\Omega)}.$$

Finally we must prove (3.17). If we assume that (H1) is satisfied, then, thanks to Theorem 4.11, (5.28) can be written as

$$H_{\overline{\alpha}_\gamma}(x, y_\gamma(x), u_\gamma(x), \overline{p}_\gamma(x)) = \min_{v\in K} H_{\overline{\alpha}_\gamma}(x, y_\gamma(x), v, \overline{p}_\gamma(x)) + \epsilon_\gamma \quad \text{a.e. } x \in \Omega.$$

Taking a subsequence $\{\gamma_j\}_{j=1}^\infty$, we pass to the limit as above and get (3.17).

If (H2) is satisfied, we can argue as in the proof of Theorem 4.9 to conclude (3.17). □

6. Proof of the strong minimum principle. In this section we establish the existence of a certain link between the stability of the cost with respect to small perturbations of the feasible state set and the viability of the exact penalization procedure of the state constraints. In the context of the abstract optimization Burke [11], generalizing an idea of Clarke [15], proved an equivalence result between stability and exact penalization. Since we are assuming the hypotheses of Theorem 3.4, we have that (P_δ) is strongly stable on the right and $(\overline{y}, \overline{u})$ is a solution of this problem. Now we consider the exact penalization of state constraints.

PROPOSITION 6.1. *If* $r > 0$ *satisfies* (3.11), *then* $\overline{u}$ *is a local solution in* $(\mathcal{K}, d)$ *of the penalized control problem*

$$\begin{cases} \min J_r(u) = \displaystyle\int_\Omega L(x, y_u(x), u(x))\, dx + r\|(g(x, y_u) - \delta)^+\|_\infty, \\ u \in \mathcal{K}. \end{cases}$$

Proof. From (3.11) it follows that

$$\inf(P_\delta) = \inf\{J(y_u, u) + r(\delta' - \delta) : u \in \mathcal{K},\ g(x, y_u(x)) \leq \delta',\ \delta' \in [\delta, \delta + \epsilon]\}.$$

Minimizing first with respect to δ' for fixed u we find

$$\inf(P_\delta) = \inf\left\{J(y_u, u) + r\|(g(x, y_u) - \delta)^+\|_\infty : u \in \mathcal{K},\ g(x, y_u(x)) \leq \delta + \epsilon\right\}.$$

Since the mapping $u \in (\mathcal{K}, d) \longrightarrow y_u \in C_0(\Omega)$ is continuous, we deduce the existence of a ball $B_\lambda(\overline{u})$, $\lambda > 0$, such that

$$\|g(x, y_u)\|_\infty < \delta + \epsilon \quad \forall u \in B_\lambda(\overline{u}),$$

which together with the previous identity proves that $\overline{u}$ is a local solution of the penalized control problem. □

Take $\lambda > 0$ as in the proof of the previous proposition and $r > 0$ verifying (3.11). We introduce the problem

$$(Q_r)\begin{cases} \min\ J_r(u) = \displaystyle\int_\Omega L(x, y_u(x), u(x))\, dx + r\|(g(x, y_u) - \delta)^+\|_\infty, \\ u \in B_\lambda(\overline{u}). \end{cases}$$

Then $\overline{u}$ is a solution of this problem. We passed from a state-constrained control problem to another control problem without state constraints. The difficulty in this new problem is that the penalization term is not differentiable. To overcome this difficulty we define

$$J_{r,q}(u) = \int_\Omega L(x, y_u(x), u(x))\, dx + r\left(q^{-q} + \int_\Omega [(g(x, y_u(x)) - \delta)^+]^q\, dx\right)^{1/q}$$

and

$$(Q_{r,q})\begin{cases} \min J_{r,q}(u) \\ u \in B_\lambda(\overline{u}), \end{cases}$$

with $q > 1$. Note that $(Q_{r,q})$ has a differentiable cost and, moreover, it represents an approximation of (Q_r) given in the following proposition.

PROPOSITION 6.2. *The following identity holds:*

$$\inf(Q_r) = \lim_{q\to\infty} \inf(Q_{r,q}).$$

Proof. From the convergence $\|z\|_{L^q(\Omega)} \to \|z\|_\infty$ for every $z \in L^\infty(\Omega)$ and the inequalities

$$\|(g(x, y_u) - \delta)^+\|_{L^q(\Omega)} \le \left(q^{-q} + \int_\Omega [(g(x, y_u(x)) - \delta)^+]^q dx\right)^{1/q}$$

$$\le \frac{1}{q} + \|(g(x, y_u) - \delta)^+\|_{L^q(\Omega)}$$

we deduce that $J_{r,q}(u) \to J_r(u)$ when $q \to +\infty$. Therefore for every $u \in B_\lambda(\overline{u})$

$$\limsup_{q\to+\infty} \inf(Q_{r,q}) \le \limsup_{q\to+\infty} J_{r,q}(u) = J_r(u);$$

hence,

$$\limsup_{q\to+\infty} \inf(Q_{r,q}) \le \inf(Q_r). \tag{6.29}$$

Now we prove the converse inequality. Let us take $\epsilon > 0$ arbitrary, and let $C_1 > 0$ be the constant given in Theorem 2.1. Since $g : \overline{\Omega} \times R \longrightarrow R$ is continuous, it follows that the existence of a constant $\rho \in (0, \epsilon)$ such that $\forall x, x' \in \overline{\Omega}$

$$|g(x,t) - g(x',t)| < \epsilon \quad \text{if} \quad |x - x'| \le \rho \text{ and } |t| \le C_1. \tag{6.30}$$

Moreover we assume ρ small enough in such a way that $m(\{x : |x| \le \rho\}) < 1$. Now we define $\Omega_\rho(x_0) = \Omega \cap B_\rho(x_0)$. Since the boundary of Γ is Lipschitz, there exists a number $\beta \in (0,1)$ verifying

$$m(\Omega_\rho(x_0)) \ge \beta m(\{x : |x| \le \rho\}) \quad \forall x_0 \in \overline{\Omega}.$$

On the other hand, from the continuity of $\frac{\partial g}{\partial y}$ we deduce the existence of another constant $M > 0$ such that

$$|g(x,t)| + \left|\frac{\partial g}{\partial y}(x,t)\right| \le M \quad \forall (x,t) \in \overline{\Omega} \times [-C_1, +C_1]. \tag{6.31}$$

Pick $u \in B_\lambda(\overline{u})$. If $\|(g(x,y_u)-\delta)^+\|_\infty = 0$, then $J_{r,q}(u) = J_r(u)$. Let us suppose that $\|(g(x,y_u)-\delta)^+\|_\infty > 0$, and take $x_0 \in \overline{\Omega}$ verifying

$$(g(x_0,y_u(x_0))-\delta)^+ = \|(g(x,y_u)-\delta)^+\|_\infty.$$

Then for each $x \in \Omega_\rho(x_0)$ we get with the aid of (2.4), (6.30), and (6.31) that

$$|g(x,y_u(x)) - g(x_0,y_u(x_0))|$$

$$\leq |g(x,y_u(x)) - g(x_0,y_u(x))| + |g(x_0,y_u(x)) - g(x_0,y_u(x_0))|$$

$$\leq \epsilon + M|y_u(x) - y_u(x_0)| \leq \epsilon + MC_1\rho^\alpha \leq M'\epsilon;$$

hence,

$$(g(x,y_u(x))-\delta)^+ \geq (g(x_0,y_u(x_0)) - \delta - M'\epsilon)^+ \quad \forall x \in \Omega_\rho(x_0).$$

Therefore, we obtain

$$\|(g(x,y_u)-\delta)^+\|_{L^q(\Omega)} \geq \left(\int_{\Omega_\rho(x_0)} [(g(x,y_u(x))-\delta)^+]^q \, dx\right)^{1/q}$$

$$\geq m(\Omega_\rho(x_0))^{1/q}(g(x_0,y_u(x_0)) - \delta - M'\epsilon)^+ \geq \|(g(x,y_u)-\delta)^+\|_\infty$$

$$+(m(\Omega_\rho(x_0))^{1/q} - 1)\|(g(x,y_u)-\delta)^+\|_\infty - M'\epsilon m(\Omega_\rho(x_0))^{1/q}$$

$$\geq \|(g(x,y_u)-\delta)^+\|_\infty + M(m(\Omega_\rho(x_0))^{1/q} - 1) - M'\epsilon.$$

Choosing $q_\epsilon > 1$ such that

$$1 - m(\Omega_\rho(x_0))^{1/q} \leq 1 - [\beta m(\{x : |x| \leq \rho\})]^{1/q} < \epsilon \quad \forall q > q_\epsilon,$$

it follows that

$$\|(g(x,y_u)-\delta)^+\|_{L^q(\Omega)} \geq \|(g(x,y_u)-\delta)^+\|_\infty - (M+M')\epsilon \quad \forall q > q_\epsilon.$$

We have proved that

$$J_{r,q}(u) \geq J_r(u) - (M+M')\epsilon$$

for each $u \in B_\lambda(\overline{u})$ and all $q \geq q_\epsilon$; hence,

$$\liminf_{q \to +\infty} \inf(Q_{r,q}) \geq \inf(Q_r) - (M+M')\epsilon$$

for $\epsilon > 0$ arbitrary; consequently,

$$\liminf_{q \to +\infty} \inf(Q_{r,q}) \geq \inf(Q_r). \tag{6.32}$$

So the proposition follows from (6.29) and (6.32). □

Proof of Theorem 3.4. Thanks to Proposition 6.2 we deduce that $\overline{u}$ is an ϵ_q^2-solution of $(Q_{r,q})$, with $\epsilon_q \to 0$ as $q \to \infty$. Then Theorem 4.11 states the existence of a control $u_q \in \mathcal{K}$, with $d(u_q, \overline{u}) \leq \epsilon_q$, satisfying for every $v \in K$

$$H(x, y_q(x), u_q(x), p_q(x)) \leq H(x, y_q(x), v, p_q(x)) + \epsilon_q \quad \text{a.e. } x \in \Omega, \tag{6.33}$$

where $y_q = y_{u_q}$ and p_q is the adjoint state:

$$\begin{cases} A^\star p_q = \dfrac{\partial f}{\partial y}(x, y_q, u_q)p_q + \dfrac{\partial L}{\partial y}(x, y_q, u_q) + \mu_q \dfrac{\partial g}{\partial y}(x, y_q) & \text{in } \Omega, \\ p_q = 0 \quad \text{on } \Gamma, \end{cases} \tag{6.34}$$

with

$$\mu_q = r\left(q^{-q} + \int_\Omega [(g(x, y_q(x)) - \delta)^+]^q \, dx\right)^{1/q-1} [(g(x, y_q) - \delta)^+]^{q-1}.$$

Now we must pass to the limit. From $d(u_q, \overline{u}) \leq \epsilon_q \to 0$ and Theorem 2.1 we obtain that $y_q \to \overline{y}$ in $H_0^1(\Omega) \cap C^{0,\alpha}(\overline{\Omega})$. On the other hand, from the definition of μ_q we get

$$\|\mu_q\|_{M(\Omega)} = \|\mu_q\|_{L^1(\Omega)} \leq r\left(\int_\Omega [g(x, y_q(x))^+]^q \, dx\right)^{1/q-1} \int_\Omega [g(x, y_q(x))^+]^{q-1} \, dx$$

$$= r\|(g(x, y_q) - \delta)^+\|_{L^q(\Omega)}^{1-q} \|(g(x, y_q) - \delta)^+\|_{L^{q-1}(\Omega)}^{q-1}.$$

Applying Hölder's inequality with exponents $q/(q-1)$ and q it follows that

$$\|z\|_{L^{q-1}(\Omega)} \leq m(\Omega)^{1/q} \|z\|_{L^q(\Omega)} \quad \forall z \in L^q(\Omega),$$

which together with the previous relation leads to

$$\|\mu_q\|_{M(\Omega)} \leq m(\Omega)^{1/q} r \|(g(x, y_q) - \delta)^+\|_{L^q(\Omega)}$$

$$\leq m(\Omega)^{2/q} r \|(g(x, y_q) - \delta)^+\|_\infty \leq M < +\infty \quad \forall q > 1.$$

As in the proof of Theorem 3.3, the boundedness of $\{\mu_q\}_{q \geq 1}$, the convergence of $\{(y_q, u_q)\}_{q \geq 1}$, and assumptions (2.3) imply the boundedness of $\{p_q\}_{q \geq 1}$ in $W_0^{1,\sigma}(\Omega)$ for every $\sigma < n/(n-1)$. Therefore we can extract subsequences $\{p_{q_k}\}$ and $\{\mu_{q_k}\}$, with $q_k \to +\infty$, converging to $\overline{p}$ and $\overline{\mu}$ in $W_0^{1,\sigma}(\Omega)$ weakly and $M(\Omega)$ $\star$-weakly, respectively. Now it is easy to pass to the limit in (6.33) and (6.34) and to obtain (3.16) and (3.14). As in the proof of Theorem 3.3 we derive (3.15) from the definition of μ_q.

Finally, as stated in Theorem 4.11, under conditions (H1) or (H2) given in Theorem 3.3, the relation (6.33) becomes

$$H(x, y_q(x), u_q(x), p_q(x)) = \min_{v \in K} H(x, y_q(x), v, p_q(x)) + \epsilon_q \quad \text{a.e. } x \in \Omega.$$

Therefore, passing to the limit in this inequality, we get (3.17). □

7. Pontryagin's principle in the control of variational inequalities. In this section we will consider the following control system:

$$(7.35)\qquad \begin{cases} Ay+\beta(y)=f(x,y(x),u(x)) \text{ in } \Omega,\\ y=0 \text{ on } \Gamma,\end{cases}$$

where A and f are as in §2 and β is a maximal monotone graph in $R\times R$ (see Brezis [10] and Barbu [2]) such that $\operatorname{dom}(\beta)\ni 0$. The control problem is

$$(P_\delta)\begin{cases} \min J(y,u)=\displaystyle\int_\Omega L(x,y(x),u(x))\,dx\\ (y,u) \text{ satisfies (7.35)}, \ u(x)\in K \text{ a.e. } x\in\Omega \text{ and } g(x,y(x))\le\delta\ \forall x\in\Omega.\end{cases}$$

We keep the assumptions stated in §2 on the data of this problem. Then we have the following result about the state equation analogous to Theorem 2.1.

THEOREM 7.1. *There exist constants $C_5>0$ and $\alpha\in(0,1)$ such that for every $u\in\mathcal{K}$ (7.35) has a unique solution $y_u\in H_0^1(\Omega)\cap C^{0,\alpha}(\overline{\Omega})$ satisfying*

$$(7.36)\qquad \|y_u\|_{H_0^1(\Omega)}+\|y_u\|_{C^{0,\alpha}(\overline{\Omega})}\le C_5.$$

Furthermore the mapping $u\in(\mathcal{K},d)\longrightarrow y_u\in H_0^1(\Omega)\cap C^{0,\alpha}(\overline{\Omega})$ is continuous.

Proof. We may assume that $\beta(0)\ni 0$. If $\operatorname{dom}(\beta)=R$ and β is Lipschitz and of class C^1, the result is consequence of Theorem 2.1. When β is a general maximal monotone graph in $R\times R$ it is enough to apply the standard procedure that consists in approximating (via Yosida's approximation and convolution with a smoothing kernel: see [2]) with a Lipschitz C^1 monotone function β_ϵ. In this way we obtain solutions $y_\epsilon\in H_0^1(\Omega)\cap C^{0,\alpha}(\overline{\Omega})$ of

$$\begin{cases} Ay+\beta_\epsilon(y)=f(x,y(x),u(x)) \text{ in } \Omega,\\ y=0 \text{ on } \Gamma.\end{cases}$$

In order to pass to the limit and derive (7.36) we need a uniform estimate of y_ϵ in $H_0^1(\Omega)\cap C^{0,\alpha}(\overline{\Omega})$. Using the mean value theorem we can write

$$\beta_\epsilon(y_\epsilon)=\beta_\epsilon'(\hat{y}_\epsilon)y_\epsilon,$$
$$f(x,y_\epsilon(x),u(x))=f(x,0,u(x))+\frac{\partial f}{\partial y}(x,\tilde{y}_\epsilon,u(x))y_\epsilon,$$

with $|\hat{y}_\epsilon(x)|\le|y_\epsilon(x)|$ and $|\tilde{y}_\epsilon(x)|\le|y_\epsilon(x)|$ for all $x\in\Omega$. Hence

$$\begin{cases} Ay_\epsilon+\left(\beta_\epsilon'(\hat{y}_\epsilon)(x))-\dfrac{\partial f}{\partial y}(x,\tilde{y}_\epsilon,u(x))\right)y_\epsilon=f(x,0,u(x)) \text{ in } \Omega,\\ y=0 \text{ on } \Gamma.\end{cases}$$

Now applying Lemma 2.2 to the above equation we get a uniform estimate of y_ϵ in $H_0^1(\Omega)\cap L^\infty(\Omega)$. Then hypotheses (2.3) on f imply that $\{f(x,y_\epsilon(x),u(x))\}_{\epsilon>0}$ is uniformly bounded in $L^s(\Omega)$. On the other hand, arguing as in [9, Appendix] we also deduce the boundedness of $\{\beta_\epsilon(y_\epsilon)\}_{\epsilon>0}$ in $L^s(\Omega)$. Therefore $\{Ay_\epsilon\}_{\epsilon>0}$ is uniformly bounded in $L^s(\Omega)$; applying Lemma 2.2 again we obtain the desired result. □

The aim of this section is to prove Pontryagin's principle for control problem (P_δ). For this purpose we need the following approximation scheme. First let us observe that Proposition 6.1 is still valid and consequently there exists a number $\lambda > 0$ such that $\overline{u}$ is a solution of the problem

$$(Q_r)\begin{cases} \min J_r(u) = \int_\Omega L(x, y_u(x), u(x))\,dx + r\|(g(x,y_u) - \delta)^+\|_\infty, \\ u \in B_\lambda(\overline{u}). \end{cases}$$

Here y_u denotes the solution of (7.35) corresponding to the control u.

The next step consists of approximating (Q_r) by a new control problem with a differentiable cost functional and a state equation with a C^1 monotone term $\beta_q(y)$. Let us begin with the last question. Following Bonnans and Tiba [9] we will say that a maximal monotone graph in $R \times R$ β_q, with $q > 1$, is an $(1/q)$-uniform approximation to β if β_q satisfies the following two conditions:

1. $\beta(t + 1/q) \geq \beta_q(t) \geq \beta(t - 1/q)$, $\forall t \in R$,
2. $\mathrm{dom}(\beta_q) \supset \mathrm{dom}(\beta)$.

Here we view β and β_q as multivalued operators extended on R with values $-\infty$ on the left of their domains and $+\infty$ on the right of their domains, and the inequality for sets means

$$\xi \geq \eta \geq \nu, \quad \forall \xi \in \beta(t + 1/q),\ \eta \in \beta_q(t),\ \nu \in \beta(t - 1/q).$$

A constructive procedure for $(1/q)$-uniform approximations of class C^1 was given in [9], and the following result was proved.

PROPOSITION 7.2. *Let $u \in \mathcal{K}$. Then the problem*

$$\begin{cases} Ay + \beta_q(y) = f(x, y(x), u(x)) & \text{in } \Omega, \\ y = 0 \quad \text{on } \Gamma \end{cases}$$

has a unique solution $y_{q,u} \in H_0^1(\Omega) \cap C^\alpha(\overline{\Omega})$ and $\|y_{q,u} - y_u\|_\infty \leq 1/q$.

Now we consider the following approximation of (Q_r):

$$(Q_{rq})\begin{cases} \min J_{r,q}(u) \\ u \in B_\lambda(\overline{u}), \end{cases}$$

where

$$J_{r,q}(u) = \int_\Omega L(x, y_{q,u}(x), u(x))dx + r\left(q^{-q} + \int_\Omega [(g(x,y_{q,u}) - \delta)^+]^q\,dx\right)^{1/q}.$$

PROPOSITION 7.3. *The following identity holds:*

$$\inf(Q_r) = \lim_{q \to \infty} \inf(Q_{r,q}).$$

Proof. With the aid of Proposition 7.2 we get

$$\left|\|(g(x,y_{q,u}) - \delta)^+\|_{L^q(\Omega)} - \|(g(x,y_{q,u}) - \delta)^+\|_\infty\right|$$

$$\leq \left|\|(g(x,y_u) - \delta)^+\|_{L^q(\Omega)} - \|(g(x,y_u) - \delta)^+\|_\infty\right|$$

$$+ \left|\|(g(x,y_{q,u}) - \delta)^+\|_{L^q(\Omega)} - \|(g(x,y_u) - \delta)^+\|_{L^q(\Omega)}\right|$$

$$+\left|\|(g(x,y_{q,u})-\delta)^+\|_\infty-\|(g(x,y_u)-\delta)^+\|_\infty\right|$$

$$\leq\left|\|(g(x,y_u)-\delta)^+\|_{L^q(\Omega)}-\|(g(x,y_u)-\delta)^+\|_\infty\right|$$

$$+\|g(x,y_{q,u})-g(x,y_u)\|_{L^q(\Omega)}+\|g(x,y_{q,u})-g(x,y_u)\|_\infty$$

$$\leq\left|\|(g(x,y_{q,u})-\delta)^+\|_{L^q(\Omega)}-\|(g(x,y_u)-\delta)^+\|_\infty\right|+\frac{m(\Omega)^{1/q}+1}{q}\longrightarrow 0 \text{ as } q\to\infty.$$

Therefore we can argue as in the proof of Proposition 6.2 to deduce that

$$\limsup_{q\to\infty}\inf(Q_{r,q})\leq\inf(Q_r). \tag{7.37}$$

Let us prove the converse inequality. Let $C_5>0$ be the constant given in Theorem 7.1. From the properties of g we obtain

$$|g(x,t)|+\left|\frac{\partial g}{\partial y}(x,t)\right|\leq M\quad\forall(x,t)\in\overline{\Omega}\times[-C_5,+C_5] \tag{7.38}$$

for some constant $M>0$. Applying the mean value theorem and using the hypotheses (2.3), (7.38), and Proposition 7.2 we get for some constant M' and all $u\in B_\lambda(\bar u)$:

$$\int_\Omega L(x,y_{q,u}(x),u(x))\,dx+r\left(q^{-q}+\int_\Omega[(g(x,y_{q,u})-\delta)^+]^q\,dx\right)^{1/q}$$

$$\geq\int_\Omega L(x,y_u(x),u(x))\,dx+r\left(q^{-q}+\int_\Omega[(g(x,y_u)-\delta)^+]^q\,dx\right)^{1/q}-\frac{M'}{q}.$$

Therefore

$$\liminf_{q\to\infty}\inf(Q_{r,q})$$

$$\geq\liminf_{q\to\infty}\inf\left\{\int_\Omega L(x,y_u(x),u(x))\,dx\right.$$

$$\left.+r\left(q^{-q}+\int_\Omega[(g(x,y_u)-\delta)^+]^q\,dx\right)^{1/q}:u\in B_\lambda(\bar u)\right\}.$$

The proof is concluded by noting that the second term of this inequality is greater than or equal to $\inf(Q_r)$, which is proved exactly in the same way than in Proposition 6.2. □

Now we are ready to state the extension of Pontryagin's principle.

THEOREM 7.4. *Let $\bar u$ be a solution of (P_δ) in $(\mathcal{K},d)$, with $\bar y$ its associated state. If (P_δ) is strongly stable on the right, then there exist $\chi\in W^{-1,\sigma}(\Omega)$, $\bar p\in W_0^{1,\sigma}(\Omega)$ for every $\sigma<\frac{n}{n-1}$, and $\bar\mu\in M(\Omega)$ such that χ is a limit point in $W^{-1,\sigma}(\Omega)$ weak of $\{\beta'_q(y_q)p_q\}$,*

$$\left\{\begin{array}{l}A^\star\bar p+\chi=\dfrac{\partial f}{\partial y}(x,\bar y(x),\bar u(x))+\dfrac{\partial L}{\partial y}(x,\bar y(x),\bar u(x))+\dfrac{\partial g}{\partial y}(x,\bar y(x))\bar\mu \textit{ in } \Omega,\\ \bar p=0 \textit{ on } \Gamma,\end{array}\right. \tag{7.39}$$

$$\int_\Omega (z(x) - g(x,\overline{y}(x)))\,d\overline{\mu}(x) \leq 0 \quad \forall z \in C_0(\Omega) \text{ with } z(x) \leq \delta \ \forall x \in \Omega, \tag{7.40}$$

and for every $v \in K$

$$H(x,\overline{y}(x),\overline{u}(x),\overline{p}(x)) \leq H(x,\overline{y}(x),v,\overline{p}(x)) \quad a.e. \ x \in \Omega. \tag{7.41}$$

Moreover if conditions (H1) *or* (H2) *of Theorem* 3.3 *hold, then*

$$H(x,\overline{y}(x),\overline{u}(x),\overline{p}(x)) = \min_{v\in K} H(x,\overline{y}(x),v,\overline{p}(x)) \quad a.e. \ x \in \Omega. \tag{7.42}$$

Proof. This theorem can be proved in the same manner as Theorem 3.4: applying Proposition 7.2 we deduce that $\overline{u}$ is a solution of (Q_r) and then Theorem 4.11 provides a minimum principle for an ϵ_q^2-solution u_q. The adjoint state corresponding to u_q satisfies the equation

$$\begin{cases} A^\star p_q + \beta_q'(y_q)p_q = \dfrac{\partial f}{\partial y}(x,y_q,u_q)p_q + \dfrac{\partial L}{\partial y}(x,y_q,u_q) + \mu_q \dfrac{\partial g}{\partial y}(x,y_q) \text{ in } \Omega, \\ p_q = 0 \text{ on } \Gamma, \end{cases}$$

with

$$\mu_q = r\left(q^{-q} + \int_\Omega [(g(x,y_q(x)) - \delta)^+]^q\,dx\right)^{1/q-1} [(g(x,y_q) - \delta)^+]^{q-1}.$$

The passage to the limit is carried out as in the proof of Theorem 3.4 with the only modification due to the term $\beta_q'(y_q)p_q$. That $\{\mu_q\}$ is bounded in $L^1(\Omega)$ can be proved as in §7; therefore, the boundedness of $\{p_q\}$ in $W_0^{1,s}(\Omega)$ is a consequence of Lemma 2.4. Finally from the adjoint state equation it follows that $\{\beta_q'(y_q)p_q\}$ is bounded in $W^{-1,\sigma}(\Omega)$, for all $\sigma < n/(n-1)$. Then there exists a subsequence, denoted in the same way, and an element $\chi \in W^{-1,\sigma}(\Omega)$ such that $\beta_q'(y_q)p_q \to \chi$ weakly in $W^{-1,\sigma}(\Omega)$ when $q \to \infty$. □

Remark 7.5. Additional information on χ can be derived from Theorem 7.4 for particular choices of β. For instance if β is Lipschitz near $y_0 \in R$ and $x_0 \in \Omega$ is such that $y(x_0) = y_0$, then $\chi(x) \in \partial_c\beta(y(x))$ with $\partial_c\beta$ the Clarke gradient [14] of β, for x close to x_0. See, for example, [9] for other illustrations.

REFERENCES

[1] R. Adams, *Sobolev Spaces*, Academic Press, New York, 1975.

[2] V. Barbu, *Optimal control of variational inequalities*, Lecture Notes in Mathematics, 100, Pitman, London, 1984.

[3] J. F. Bonnans, *El principio de Pontryagine para el control de sistemas elípticos con restricciones sobre el estado: el método de penalización*, in Jornadas Hispano-Francesas sobre Control de Sistemas Distribuidos, A. Valle, ed., Málaga, 1990, 1991, pp. 13–19.

[4] ———, *Pontryagin's principle for the optimal control of semilinear elliptic systems with state constraints*, in 30th IEEE Conference on Control and Decision, Brighton, England, 1991, pp. 1976–1979.

[5] J. F. Bonnans and E. Casas, *Contrôle de systèmes elliptiques semilinéaires comportant des contraintes sur l'état*, in Nonlinear Partial Differential Equations and Their Applications. Collège de France Seminar, H. Brezis and J. Lions, eds., vol. 8, Longman Scientific & Technical, New York, 1988, pp. 69–86.

[6] ———, *Optimal control of semilinear multistate systems with state constraints*, SIAM J. Control Optim., 27 (1989), pp. 446–455.

[7] ———, *Un principe de Pontryagine pour le contrôle des systèmes elliptiques*, J. Differential Equations, 90 (1991), pp. 288–303.

[8] ———, *A boundary Pontryagin's principle for the optimal control of state constrained elliptic systems*, in Proc. French-Romanian Conference on Optimization, Optimal Control and Partial Differential Equations, Internat. Ser. Numer. Math. 107, V. Barbu, J. F. Bonnans, and D. Tiba, eds., Birkhäuser, Basel, 1992, pp. 241–249.

[9] J. F. BONNANS AND D. TIBA, *Pontryagin's principle in the control of semilinear elliptic variational equations*, J. Appl. Math. Optim., 23 (1991), pp. 299–312.

[10] H. BREZIS, *Problèmes unilatéraux*, J. Math. Pures Appl. (g), 51 (1972), pp. 1–68.

[11] J. BURKE, *Calmness and exact penalization*, SIAM J. Control Optim., 29 (1991), pp. 493–497.

[12] E. CASAS, *Control of an elliptic problem with pointwise state constraints*, SIAM J. Control Optim., 24 (1986), pp. 1309–1318.

[13] ———, *Finite element approximations for some optimal control problems with pointwise state constraints*, in IMACS '91 13th World Congress on Computation and Applied Mathematics, R. Vichnevetsky and J. Miller, eds., vol. 3, Dublin, 1991, Criterion Press, pp. 1165–1166.

[14] ———, *Análisis de la convergencia en la aproximación numérica de problemas de control con restricciones sobre el estado*, in Actas del XII C.E.D.Y.A./II Congreso de Matemática Aplicada, Oviedo, 1991, pp. 301–306.

[15] F. CLARKE, *A new approach to Lagrange multipliers*, Math. Oper. Res., 1 (1976), pp. 165–174.

[16] I. EKELAND, *Nonconvex minimization problems*, Bull. Amer. Math. Soc. (N.S.), 1 (1979), pp. 76–91.

[17] D. GILBARG AND N. TRUDINGER, *Elliptic Partial Differential Equations of Second Order*, Springer-Verlag, Berlin, Heidelberg, New York, 1977.

[18] Z.-X. HE, *State constrained control problems governed by variational inequalities*, SIAM J. Control Optim., 25 (1987), pp. 1119–1143.

[19] J. L. LIONS, *Contrôle Optimal de Systèmes Gouvernés par des Equations aux Dérivées Partielles*, Dunod, Paris, 1968.

[20] F. MIGNOT, *Contrôle dans les inéquations variationelles elliptiques*, Functional Anal., 22 (1974), pp. 130–185.

[21] F. MIGNOT AND J. PUEL, *Optimal control in some variational inequalities*, SIAM J. Control Optim., 22 (1984), pp. 466–476.

[22] G. STAMPACCHIA, *Le problème de Dirichlet pour les équations elliptiques du second ordre à coefficients discontinus*, Ann. Inst. Fourier (Grenoble), 15 (1965), pp. 189–258.

© 1995 Society for Industrial and Applied Mathematics

014

A NONCONVEX VARIATIONAL PROBLEM WITH CONSTRAINTS*

MICÒL AMAR† AND CARLO MARICONDA‡

Abstract. A multidimensional version of Liapunov-type theorems is proven. As an application, it is proven that, under proper hypothesis on the possibly nonconvex function f, the problem $\min\int_0^T f(u'(t))\,dt$ on the subset of $W^{1,p}([0,T],\mathbb{R}^n)$ of those functions u satisfying the prescribed boundary conditions and whose trajectory lies out of a prescribed open subset of $\mathbb{R}^n$ admits at least one solution.

Key words. relaxed problem, bipolar, Liapunov, simplex, convex, extremal point

AMS subject classification. 49A05

1. Introduction. The most general scalar problem that has been investigated without the classical Tonelli convexity condition on the function $\xi \to h(t,s,\xi)$ is that of minimizing

$$\text{(P}'\text{)}\qquad \begin{gathered}\min I(u) = \min\left\{\int_0^T h(t,u(t),u'(t))\,dt\right\},\\ u\in W^{1,p}([0,T],\mathbb{R}^n),\quad u(0)=a,\quad u(T)=b.\end{gathered}$$

Under differentiability assumptions on the integrand, this problem was studied by Aubert and Tahraoui in [2] and [3], Raymond in [11], and Tahraoui in [13].

In the case $h(t,s,\xi) = g(t,s) + f(t,\xi)$, this problem was studied by Olech (see [10]), Marcellini (see [8]), Cellina and Colombo (see [4]), and Raymond (see [12]), under weaker assumptions on the regularity of g and f.

In particular, in [4] the main tool is a Liapunov-type theorem, which allows the modification of a solution to the convexified problem in order to obtain a solution of the original one. The same technique has also been used in [12] and [9].

For $n = 1$, i.e., for functions with values in $\mathbb{R}$, a more precise version of Liapunov's theorem has recently been given in [1].

THEOREM 1.1. *Let $\Phi : [0,T] \to 2^{\mathbb{R}}$ be a measurable multifunction with values in the closed intervals of $\mathbb{R}$. Then for each integrable selection $\tilde{u}'$ of $\Phi(t)$, there exists a measurable selection $\bar{u}'$ with values in the extreme points of $\Phi(t)$ such that $\int_0^T \bar{u}'(t)\,dt = \int_0^T \tilde{u}'(t)\,dt$ and for each $t\in[0,T]$, $\bar{u}(t)\le\tilde{u}(t)$.*

This result has been successfully applied in [5] in order to prove that there exists a dense subset D of $C([0,T],\mathbb{R})$ such that, for g in it, the problem of minimizing $\int_0^T g(u(t))\,dt + \int_0^T f(u'(t))\,dt$ does always admit at least one solution for each f satisfying growth conditions.

* Received by the editors July 27, 1992; accepted for publication (in revised form) September 28, 1993. This research was partially supported by a Consiglio Nazionale della Ricerche fellowship during the authors' visit to the University of Paris VI Pierre et Marie Curie.

† Dipartimento di Matematica, Università degli Studi di Pavia, Via Abbiategrasso 209, 27100 Pavia, Italy (AMAR%IPVIAN.BITNET@VM.CNUCE.CNR.IT).

‡ Dipartimento di Matematica Pura e Applicata, Università degli Studi di Padova, Via Belzoni 7, 35131 Padova, Italy (MARICONDA@PDMAT1.MATH.UNIPD.IT).

A more than one-dimensional version of the above Liapunov type lemma does not hold, in general.

Example. Let $n = 2$, $T = 1$, $\Phi(t) = \{\lambda(1,t) \ : \lambda \in [0,1]\}$, $\tilde{u}'(t) = (\tilde{u}_1'(t), \tilde{u}_2'(t)) = (\frac{1}{2}, (\frac{1}{2})t) \in \Phi(t)$ a.e. in $[0,1]$. Assume, by contradiction, that there exists $\bar{u}'(t) = (\bar{u_1}'(t), \bar{u_2}'(t)) \in \{(0,0),(1,t)\}$ a.e. such that

$$\int_0^1 \bar{u}'(t)\,dt = \int_0^1 \tilde{u}'(t)\,dt, \tag{1.1}$$

$$\bar{u}_1(t) \geq \tilde{u}_1(t) \quad \text{for a.e. } t \in [0,1], \tag{1.2}$$

$$\bar{u}(0) = \tilde{u}(0). \tag{1.3}$$

Then there exists a measurable subset E of $[0,1]$ such that

$$\bar{u}'(t) = (0,0)\chi_{[0,1]\setminus E} + (1,t)\chi_E;$$

whence, $\bar{u}_2'(t) = t\bar{u}_1'(t)$. Conditions (1.1) and (1.3) and integration by parts of the second component give

$$\int_0^1 \bar{u}_1(t)\,dt = \int_0^1 \tilde{u}_1(t)\,dt$$

so that, by (1.2), $\bar{u}_1(t) = \tilde{u}_1(t)$, i.e., $\chi_E = \frac{1}{2}$. This is a contradiction.

Neverthless, we prove here that a multidimensional version of the above theorem holds if the measurable function Φ is identically equal to a convex bounded subset of $\mathbb{R}^n$. As an application, we study the problem of minimizing

$$\int_0^T f(u'(t))\,dt$$

on the subset of $W^{1,p}([0,T],\mathbb{R}^n)$ of those functions u satisfying prescribed boundary conditions and whose trajectory lies out of a prescribed open subset of $\mathbb{R}^n$.

2. Notation and preliminary results. In the following, Γ will denote an open convex poligone contained in $\mathbb{R}^n$ and, given $a, b \in \mathbb{R}^n \setminus \Gamma$, K will be the set of those functions $u : [0,T] \to \mathbb{R}^n$ that are in the Sobolev space $W^{1,p}((0,T),\mathbb{R}^n)$ $(p \geq 1)$ and such that $u(0) = a, u(T) = b$.

Given a set A, we denote by ∂A the boundary of A, by $\text{extr}A$ the extremal points of A, and by $\text{meas}(A)$ the Lebesgue measure of A. Finally, given two vectors v_1 and v_2 of $\mathbb{R}^n$, we denote by $v_1 \cdot v_2$ the usual scalar product in $\mathbb{R}^n$ and by $|v_1|$ the euclidean norm of v_1 in $\mathbb{R}^n$.

Let $f : \mathbb{R}^n \to \mathbb{R}$ be a nonnecessarily convex and lower semicontinuous function that satisfies the following growth conditions:

$$\text{(F)} \qquad \begin{aligned} c_1|\xi|^p - c_2 \leq f(\xi) \quad & \forall \xi \in \mathbb{R}^n \quad \text{if } p > 1, \\ \psi(|\xi|) - c_2 \leq f(\xi) \quad & \forall \xi \in \mathbb{R}^n \quad \text{if } p = 1, \end{aligned}$$

where c_1 and c_2 are real positive constants and $\psi : [0,+\infty) \to [0,+\infty)$ is a convex and lower semicontinuous function such that $\lim_{r\to+\infty} \frac{\psi(r)}{r} = +\infty$.

Given a function f, we denote by f^{**} its bipolar function.

LEMMA 2.1 (see, for instance, [6, Prop. I.4.1]). *Let $f : \mathbb{R}^n \to \mathbb{R}$ be a lower semicontinuous function. Then f^{**} is the greatest lower semicontinuous and convex function not greater than f.*

Let us consider the set

$$E = \{\xi \in \mathbb{R}^n \ : \ f^{**}(\xi) < f(\xi)\}. \tag{2.1}$$

In the following, we shall assume that $E = \bigcup_{i\in\mathbb{N}} A_i$ and f^{**} is affine on every A_i, where A_i is a convex open and bounded subset of $\mathbb{R}^n$.

Notice that, while the hypothesis on the structure of E is quite natural, the hypothesis on the form of f^{**} on every A_i is a technical hypothesis, which is automatically satisfied only in the scalar case (i.e., when $n = 1$). On the contrary, when n is strictly greater than one, this hypothesis is not always fulfilled.

LEMMA 2.2. *Let A be a simplex in $\mathbb{R}^n$, $I \subseteq [0,T]$ be a measurable set, $u' : [0,T] \to A$ be a measurable function, and η be a fixed vector in $\mathbb{R}^n$. Then there exists a measurable function, $\omega : [0,T] \to \operatorname{extr} A$ depending on u', A, and η, such that*

$$\int_I \omega(s)\,ds = \int_I u'(s)\,ds, \tag{2.2}$$

$$\forall t \in [0,T] \quad \int_0^t [\omega(s)\cdot\eta\chi_I(s)]\,ds \geq \int_0^t [u'(s)\cdot\eta\chi_I(s)]\,ds. \tag{2.3}$$

Proof. Let $v_0,\dots,v_n$ be the $n+1$ vertices of the simplex A; then $u'(s) = p_0(s)v_0 + \cdots + p_n(s)v_n$ for a proper choice of $p_0,\dots,p_n : [0,T] \to [0,1]$ with $p_0(s) + \cdots + p_n(s) \equiv 1$. Moreover, for every $i = 0,\dots,\ n$, let us define $a_i := v_i\cdot\eta$. Without loss of generality, we may assume that $a_0 > \cdots > a_n$.

We shall prove that there exists a measurable partition $E_0,\dots,E_n$ of I such that

$$\operatorname{meas}(E_i) = \int_I p_i(s)\,ds \quad \forall\ i = 0,\dots,n, \tag{2.4}$$

$$\int_0^t \sum_{i=0}^n a_i p_i(s)\chi_I(s)\,ds \leq \int_0^t \sum_{i=0}^n a_i\chi_{E_i}(s)\,ds \quad \forall t \in [0,T]. \tag{2.5}$$

It is clear that, setting $\omega(s) = \sum_{i=0}^n v_i\chi_{E_i}(s)$, (2.2) and (2.3) follow from (2.4) and (2.5). In order to prove (2.4) and (2.5), we proceed by induction. When $n = 0$, we have that $p_0(s) \equiv 1$; and if we set $E_0 = I$, the thesis is trivially satisfied. Let us assume now that $n > 0$. Let $0 = t_0 < t_1 < \cdots < t_{n+1} = T$ be a partition of $[0,T]$ such that

$$\int_{t_i}^{t_{i+1}} \chi_I(s)\,ds = \int_I p_i(s)\,ds \quad \forall\ i = 0,\dots,n.$$

Such partition exists since $p_0(s) + \cdots + p_n(s) \equiv 1$. Let us define

$$E_0 = [t_0,t_1] \cap I, \quad \text{and} \quad E_i = (t_i,t_{i+1}] \cap I \quad \forall\ i = 0,\dots,n.$$

First, by the very definition of E_i, (2.4) trivially holds. In order to prove (2.5), we proceed as follows.

Let us define

$$\tilde{E}_i = E_i, \quad \tilde{p}_i = p_i \quad \forall i = 0,\dots,n-2,$$
$$\tilde{E}_{n-1} = E_{n-1} \cup E_n, \qquad \tilde{p}_{n-1} = p_{n-1} + p_n.$$

Clearly, $\bigcup_{i=0}^{n-1} \tilde{E}_i = I$, $\sum_{i=0}^{n-1} \tilde{p}_i = 1$, and (2.4) is satisfied by $\tilde{E}_i$ and $\tilde{p}_i$ for $i =$

$0, \ldots, n-1$. Moreover, the hypothesis of induction assures that

$$\int_0^t \sum_{i=0}^{n-1} a_i \chi_{\tilde{E}_i}(s)\, ds \geq \int_0^t \sum_{i=0}^{n-1} a_i \tilde{p}_i(s) \chi_I(s)\, ds. \tag{2.6}$$

Assume that $t \leq t_n$. We observe that, in this case, $E_n \cap [0,t] = \emptyset$; hence for every $i = 0, \ldots, n-1$ we have that $E_i \cap [0,t] = \tilde{E}_i \cap [0,t]$. Then, by (2.6), it follows that

$$\begin{aligned}\int_0^t \sum_{i=0}^{n} a_i \chi_{E_i}(s)\, ds &= \int_0^t \sum_{i=0}^{n-1} a_i \chi_{E_i}(s)\, ds \\ &= \int_0^t \sum_{i=0}^{n-1} a_i \chi_{\tilde{E}_i}(s)\, ds \geq \int_0^t \sum_{i=0}^{n-1} a_i \tilde{p}_i(s) \chi_I(s)\, ds \\ &\geq \int_0^t \sum_{i=0}^{n} a_i p_i(s) \chi_I(s)\, ds.\end{aligned}$$

Assume now that $t_n < t \leq T$. Then

$$\begin{aligned}\int_0^t \sum_{i=0}^{n} a_i \chi_{E_i}(s)\, ds &= \sum_{i=0}^{n-1} a_i \operatorname{meas}(E_i) + \int_{t_n}^t a_n \chi_{E_n}(s)\, ds \\ &= \sum_{i=0}^{n-1} a_i \int_0^T p_i(s) \chi_I(s)\, ds + \int_{t_n}^t a_n \chi_{E_n}(s)\, ds \\ &\geq \int_0^t \sum_{i=0}^{n-1} a_i p_i(s) \chi_I(s)\, ds + a_n \left[\int_t^T \sum_{i=0}^{n-1} p_i(s) \chi_I(s)\, ds + \int_0^t \chi_{E_n}(s)\, ds \right] \\ &= \int_0^t \sum_{i=0}^{n-1} a_i p_i(s) \chi_I(s)\, ds + a_n \left[\int_{t_n}^T \sum_{i=0}^{n-1} p_i(s) \chi_I(s)\, ds + \int_{t_n}^t p_n(s) \chi_I(s)\, ds \right] \\ &= \int_0^t \sum_{i=0}^{n-1} a_i p_i(s) \chi_I(s)\, ds + a_n \operatorname{meas}(E_n) - a_n \int_t^T p_n(s) \chi_I(s)\, ds \\ &= \int_0^t \sum_{i=0}^{n-1} a_i p_i(s) \chi_I(s)\, ds + \int_0^t a_n p_n(s) \chi_I(s)\, ds = \int_0^t \sum_{i=0}^{n} a_i p_i(s) \chi_I(s)\, ds.\end{aligned}$$

Hence, also (2.5) holds and the lemma is proved. □

LEMMA 2.3. *Let A be an open convex bounded subset of $\mathbb{R}^n$. Then A can be covered by a countable family of simplexes whose vertices are contained in the boundary of A.*

Proof. Let $x_1, \ldots, x_{n+1}$ be $n+1$ points of the boundary ∂A of A, such that they generate a closed $(n+1)$-dimensional simplex denoted by S_1. We denote by F_i (for $i = 1, \ldots, n+1$) the face generated by

$$\{x_1, \ldots, \check{x}_i, \ldots, x_{n+1}\}$$

and let q_i be a point of F_i such that

$$d(q_i, \partial A) = \max\{d(x, \partial A) : x \in F_i\}.$$

Moreover, let ν_i be the external half-line normal to F_i at the point q_i and let $x_{1,i}$ be its intersection with ∂A. Let $T_{1,i}$ $(i = 1, \ldots, n+1)$ be the closed $(n+1)$-dimensional simplex generated by

$$\{x_1, \dots, x_{i-1}, x_{1,i}, x_{i+1}, \dots, x_{n+1}\},$$

and set

$$S_2 = S_1 \cup \bigcup_{i=1}^{n+1} T_{1,i}.$$

Recursevely, one obtains an increasing family of convex polygons whose vertices lie in ∂A; moreover, each S_{j+1} is obtained by adding to S_j a finite union of $(n+1)$-simplexes $T_{j,k}$ $(1 \leq k \leq k(j) < +\infty)$ with vertices in ∂A. We claim that

$$A \subset \bigcup_j S_j.$$

Clearly, it is enough to prove that

$$\lim_j \max_{x \in \partial S_j} d(x, \partial A) = 0. \tag{2.7}$$

In order to prove (2.7), let us remark that, if it does not hold, there exists $d_0 > 0$ such that

$$\max_{x \in \partial S_j} d(x, \partial A) \geq d_0$$

for each $j \in \mathbb{N}$. It follows that, by construction,

$$\max_{x \in \partial T_{j,k}} d(x, \partial A) \geq d_0$$

for each $j \in \mathbb{N}$ and $k \leq k(j)$, so that the "heights" of the simplexes $T_{j,k}$ (and hence their volumes) are bounded below by a positive constant, a contradiction, the set A being bounded. □

3. Main results.

THEOREM 3.1. *Let A be an open convex and bounded subset of $\mathbb{R}^n$, I a measurable subset of $[0,T]$, $u' : I \longrightarrow A$ a measurable function, and η an arbitrary vector in $\mathbb{R}^n$. Then there exists a function $\omega : I \longrightarrow \partial A$, depending on u', A, and η, such that*

$$\int_I \omega(s)\, ds = \int_I u'(s)\, ds, \tag{3.1}$$

$$\forall t \in [0,T] \qquad \int_0^t [\omega(s) \cdot \eta\chi_I(s)]\, ds \geq \int_0^t [u'(s) \cdot \eta\chi_I(s)]\, ds. \tag{3.2}$$

COROLLARY 3.2. *Assume that A, I, u', ω, and η are as in the previous theorem. Assume that $f : \mathbb{R}^n \longrightarrow \mathbb{R}$ is lower semicontinuous, f^{**} is affine on A, and $f(\xi) = f^{**}(\xi)$ when $\xi \in \partial A$. Then*

$$\int_I f^{**}(u'(s))\, ds = \int_I f(\omega(s))\, ds. \tag{3.3}$$

Proof. Since f^{**} is affine on A, there exist two vectors v_1 and v_2 such that

$$f^{**}(\xi) = v_1 \cdot \xi + v_2 \quad \forall \xi \in A;$$

hence, by (2.2) of Lemma 2.2, it easily follows that

$$\int_I f^{**}(u'(s))\, ds = \int_I f^{**}(\omega(s))\, ds.$$

Finally, recalling that ω takes values in ∂A and f^{**} coincides with f on ∂A, (2.4) follows. □

Proof of Theorem 3.1. By Lemma 2.3 $A = \bigcup_{j\in\mathbb{N}} S_j$, where S_j is a simplex contained in $\mathbb{R}^n$ whose vertices belongs to ∂A. Let us set, for every $j \in \mathbb{N}$, $I_j := (u')^{-1}(S_j) \cap I$ and $u'_j := u'\chi_{I_j} : I_j \to S_j$. Applying Lemma 2.2 to u'_j in I_j, we obtain a function $\omega_j : I_j \to \operatorname{extr} S_j \subset \partial A$, which satisfies (2.2) and (2.3). Hence, defining $\omega(s) := \sum_{j\in\mathbb{N}} \omega_j(s)$, it is clear that ω takes values in ∂A and satisfies (3.1) and (3.2). □

4. Applications. We consider the following minimum problem with obstacle

$$\min_{\substack{u\in K\\ u(t)\notin\Gamma}} \int_0^T f(u'(t))\,dt.$$

As we have already announced in the introduction, our main goal is to prove the existence of a solution for this minimum problem.

THEOREM 4.1. *Let $\Gamma \subset \mathbb{R}^n$, $K \subset W^{1,p}([0,T],\mathbb{R}^n)$, $f : \mathbb{R}^n \to \mathbb{R}$ be as in* §2. *Assume further that $f^{**}(0) = f(0)$. Then the problem*

$$\text{(P)} \qquad \min\left\{\int_0^T f(u'(s))\,ds \;:\; u\in K,\, u(t)\notin\Gamma\right\}$$

admits at least one solution.

Proof. Assume, for the sake of simplicity, that the set defined by

$$\{\xi\in\mathbb{R}^n : f^{**}(\xi) < f(\xi)\}$$

coincides with a simplex E on which f^{**} is affine; by the remark following Lemma 2.1 and by Lemmas 2.3 and 2.2, this is not restrictive. Assume further that $n = 2$, the general case being similar. Let p_i $(i = 1,\dots,m)$ be the vertices of Γ; by G_i we denote the relative interior of the side $\overline{p_ip_{i+1}}$ and by ν_i their external normal vector. The set Γ being open, there exists a solution $\tilde u$ to the associated relaxed problem

$$\min\left\{\int_0^T f^{**}(u'(s))\,ds \quad : \quad u\in K, u(t)\notin\Gamma\right\}.$$

Since the measure of the interval $[0,T]$ is finite, for each vertex p_i there exists an external half-line L_i containing p_i such that, setting $N = \{t : \tilde u(t) \in L_i \setminus \{p_i\}\}$, we have $\operatorname{meas}(N) = 0$.

Fix G_i and consider the "external" unbounded set O_i defined by the interior of the region delimited by the half-lines L_i, L_{i+1} and the side G_i, jointly with the side G_i itself.

Clearly, each O_i is open in the relative topology of $\mathbb{R}^2\setminus\Gamma$; moreover, the solution $\tilde u$ does not belong to Γ. Hence, the inverse image of $\mathcal{O} = \bigcup_i O_i$ under $\tilde u$ is a countable union of relative open intervals (α_j,β_j) of $[0,T]$: for every $j\in\mathbb{N}$ let $i(j)$ be such that

$$\tilde u(\alpha_j,\beta_j) \subset O_{i(j)}.$$

Let us define by $\mathcal{K}$ the subset of $[0,T]$ where $f\circ\tilde u'$ does not coincide with $f^{**}\circ\tilde u'$, i.e.,

$$\mathcal{K} = (\tilde u')^{-1}(E) = \{t : f^{**}(\tilde u'(t)) \neq f(\tilde u'(t))\};$$

and set, for each $j\in\mathbb{N}$,

$$\mathcal{S}_j = (\alpha_j,\beta_j)\cap\mathcal{K}.$$

Now, for each j, $\tilde{u}'(\mathcal{S}_j) \subset E$, on which f^{**} is affine; by Corollary 3.2 there exists a measurable function

$$\sigma_j : \mathcal{S}_j \to \mathbb{R}^2$$

with values in ∂E (on which $f^{**} = f$) satisfying

$$\int_{\mathcal{S}_j} \sigma_j(t)\,dt = \int_{\mathcal{S}_j} \tilde{u}'(t)\,dt, \tag{4.1}$$

$$\forall t \in [\alpha_j, \beta_j] \qquad \int_0^t \sigma_j(s) \cdot \nu_{i(j)} \chi_{\mathcal{S}_j}(s)\,ds \geq \int_0^t \tilde{u}'(s) \cdot \nu_{i(j)} \chi_{\mathcal{S}_j}(s)\,ds, \tag{4.2}$$

$$\int_{\mathcal{S}_j} f(\sigma_j(t))\,dt = \int_{\mathcal{S}_j} f^{**}(\tilde{u}'(t))\,dt. \tag{4.3}$$

Let $\bar{u}' : [0,T] \to \mathbb{R}^2$ be the measurable function defined by

$$\bar{u}' = \tilde{u}' \chi_{[0,T] \setminus \bigcup_j \mathcal{S}_j} + \sum_{j \in \mathbb{N}} \sigma_j \chi_{\mathcal{S}_j}.$$

The growth conditions on f and relation (4.1) show, together with Vitali's convergence theorem, that $\bar{u}' \in L^p$. Let $\bar{u}$ be the function defined by

$$\bar{u}(t) = a + \int_0^t \bar{u}'(s)\,ds.$$

We claim that $\bar{u}$ is a solution to (P).

Clearly, by (4.1) and the definition of $\bar{u}$ we have

$$\bar{u}(0) = \tilde{u}(0), \qquad \bar{u}(T) = \tilde{u}(T).$$

In order to prove that

$$\int_0^T f(\bar{u}'(t))\,dt = \int_0^T f^{**}(\tilde{u}'(t))\,dt = \min_{\substack{u \in K \\ u(t) \notin \Gamma}} \int_0^T f^{**}(u'(t))\,dt \tag{4.4}$$

we first remark that $[0,T]$ can be partitioned as a disjoint union of four measurable subsets N, D_1, D_2, D_3 where

$$D_1 = \bigcup_j \mathcal{S}_j = \tilde{u}^{-1}(\mathcal{O}) \cap \mathcal{K}, \qquad D_2 = \tilde{u}^{-1}(\mathcal{O}) \cap ([0,T] \setminus \mathcal{K}), \qquad D_3 = \tilde{u}^{-1}(\{p_1, \ldots, p_m\}).$$

By (4.3) we have

$$\int_{D_1} f(\bar{u}'(t))\,dt = \int_{D_1} f^{**}(\tilde{u}'(t))\,dt; \tag{4.5}$$

by the very definitions of $\bar{u}$ and $\mathcal{K}$ we have

$$\text{for a.e. t} \in [0,T] \setminus \mathcal{K} \qquad f(\bar{u}'(t)) = f(\tilde{u}'(t)) = f^{**}(\tilde{u}'(t))$$

so that in particular

$$\int_{D_2} f(\bar{u}'(t))\,dt = \int_{D_2} f^{**}(\tilde{u}'(t))\,dt. \tag{4.6}$$

Finally, by [7, Lemma 7.7], on $\tilde{u}^{-1}(\{p_1,\dots,p_m\}) = D_3$ we have $\tilde{u}' = 0$ a.e.; since by the very definition $\bar{u}' = 0$ on D_3 and by our assumption $f^{**}(0) = f(0)$

$$\text{for a.e. t} \in D_3 \qquad f(\bar{u}'(t)) = f(0) = f^{**}(0) = f^{**}(\tilde{u}'(t))$$

so that

$$\int_{D_3} f(\bar{u}'(t))\,dt = \int_{D_3} f^{**}(\tilde{u}'(t))\,dt. \tag{4.7}$$

Taking into account that N has measure zero, equalities (4.5), (4.6), and (4.7) together give (4.4).

At this stage we only need to show that

$$\forall t \in [0,T]: \quad \bar{u}(t) \notin \Gamma.$$

Fix t in $[0,T]$: either there exists $j_0 \in \mathbb{N}$ such that $t \in (\alpha_{j_0}, \beta_{j_0})$ or t does not belong to $\tilde{u}^{-1}(\mathcal{O})$. In the first case let $i \in \mathbb{N}$ be such that $\tilde{u}(\alpha_{j_0}, \beta_{j_0}) \subset O_i$; in order to prove that $\bar{u}(t) \notin \Gamma$ it is enough to show that

$$\bar{u}(t) \cdot \nu_i \geq \tilde{u}(t) \cdot \nu_i. \tag{4.8}$$

Since $\tilde{u}' = \bar{u}'$ on $[0,T] \setminus \bigcup_j \mathcal{S}_j$ then by (4.1) and (4.2) we have

$$\begin{aligned}
(\bar{u}(t) - \tilde{u}(t)) \cdot \nu_i &= \int_0^t (\bar{u}'(s) - \tilde{u}'(s)) \cdot \nu_i\,ds \\
&= \int_0^t (\bar{u}'(s) - \tilde{u}'(s)) \cdot \nu_i \chi_{\cup_j \mathcal{S}_j}(s)\,ds \\
&= \sum_{\{j:\mathcal{S}_j \subset [0,t]\}} \int_{\mathcal{S}_j} (\sigma_j(s) - \tilde{u}'(s)) \cdot \nu_i\,ds \\
&\quad + \int_0^t (\sigma_{j_0}(s) - \tilde{u}'(s)) \cdot \nu_i \chi_{\mathcal{S}_{j_0}}(s)\,ds \\
&= \int_0^t (\sigma_{j_0}(s) - \tilde{u}'(s)) \cdot \nu_i \chi_{\mathcal{S}_{j_0}}(s)\,ds \geq 0,
\end{aligned}$$

which proves (4.8).

In the second case ($t \notin \tilde{u}^{-1}(\mathcal{O})$) there is no interval (α_j, β_j) containing t. It follows that for each j in $\mathbb{N}$ either $\mathcal{S}_j \subset [0,t]$ or $\mathcal{S}_j \cap [0,t] = \emptyset$. As a consequence we have

$$\begin{aligned}
\bar{u}(t) - \tilde{u}(t) &= \int_0^t \bar{u}'(s) - \tilde{u}'(s)\,ds \\
&= \int_0^t (\bar{u}'(s) - \tilde{u}'(s))\chi_{\cup_j \mathcal{S}_j}(s)\,ds \\
&= \sum_{\{j:\mathcal{S}_j \subset [0,t]\}} \int_{\mathcal{S}_j} \sigma_j(s) - \tilde{u}'(s)\,ds.
\end{aligned}$$

Equality (4.1) yields $\bar{u}(t) = \tilde{u}(t)$; in particular $\bar{u}(t) \notin \Gamma$, the conclusion follows. □

As a consequence of the proof of the above theorem, we have the following result, with no assumption on the bipolar of f in 0.

THEOREM 4.2. *Let $\Gamma \subset \mathbb{R}^n$ be an open half-space, $K \subset W^{1,p}([0,T],\mathbb{R}^n)$, $f : \mathbb{R}^n \to \mathbb{R}$ be as in §2. Then the problem*

$$\min\left\{\int_0^T f(u'(s))\,ds \;:\; u \in K,\, u(t) \notin \Gamma\right\}$$

admits at least one solution.

Acknowledgments. We wish to thank Professor Arrigo Cellina for suggesting this problem and for his helpful advice. The authors are grateful to the Laboratoire d'Analyse Numérique of Paris VI for its hospitality and to the Consiglio Nazionale delle Ricerche, who gave them the opportunity to stay in Paris. We are indebted to the unknown referees for kind remarks that improved the presentation of the paper.

REFERENCES

[1] M. AMAR AND A. CELLINA, *On passing to the limit for non convex variational problems*, Asymptotic Anal., to appear.

[2] G. AUBERT AND R. TAHRAOUI, *Théorèmes d'existence pour des problèmes du calcul des variations*, J. Differential Equations, 33 (1979), pp. 1–15.

[3] ———, *Théorèmes d'existence en optimisation non-convexe*, Appl. Anal., 18 (1984), pp. 75–100.

[4] A. CELLINA AND G. COLOMBO, *On a classical problem of the calculus of variations without convexity assumptions*, Ann. Inst. H. Poincaré Anal. Non linéaire, 7 (1990), pp. 97–106.

[5] A. CELLINA AND C. MARICONDA, *The existence question in the calculus of variations: a density result*, Proc. Amer. Math. Soc., 120 (1994), pp. 1145–1150.

[6] I. EKELAND AND R. TEMAM, *Convex analysis and variational problems*, North-Holland, Amsterdam, 1976.

[7] D. GILBARG AND N. S. TRUDINGER, *Elliptic partial differential equations of second order*, Springer-Verlag, New York, 1977.

[8] P. MARCELLINI, *Non-convex integrals of the calculus of variations*, Notes of the CIME Course on Methods of Nonconvex Analysis, 1989.

[9] C. MARICONDA, *A generalization of the Cellina–Colombo Theorem for a class of non-convex variational problems*, J. Math. Anal. Appl., 175 (1993), pp. 514–522.

[10] C. OLECH, *Integrals of set-valued functions and linear optimal control problems*, Colloque sur la Théorie Mathématique du contrôle optimal, C. B. R. M., Vander, Louvain, 1970, pp. 109–125.

[11] J. P. RAYMOND, *Conditions nécessaires et suffisantes d'existence de solutions en calcul des variations*, Ann. Inst. H. Poincaré Anal. Non linéaire, 4 (1987), pp. 169–202.

[12] ———, *Existence theorems in optimal control problems without convexity assumptions*, J. Optim. Theory Appl., 67.1 (1990), pp. 109–132.

[13] R. TAHRAOUI, *Sur une classe de fonctionnelles non-convexes et applications*, SIAM J. Math. Anal., 21 (1990), pp. 37–52.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 33, No. 1, pp. 308–322, January 1995

©1995 Society for Industrial and Applied Mathematics
015

THE GLOBAL CONTROL OF NONLINEAR DIFFUSION EQUATIONS*

J. E. RUBIO[†]

Abstract. The boundary control of a nonlinear difussion equation with an integral performance criterion and a fixed final state is considered. By means of a process of embedding used by the author and others for finite-dimensional systems, this problem is replaced by one in which a linear form is minimized over a set of pairs of positive measures satisfying linear constraints. The advantages of this formulation are: (i) There is an automatic existence theory. (ii) There exists the possibility of using linear functional analysis to develop the theory. (iii) The minimization is global. The final state is only reached, however, in an asymptotic fashion, as the number of constraints being considered tends to infinity. A theory of controllability and reachability is developed, as well as a computational method using an infinite-dimensional simplex method. An example is given.

Key words. nonlinear difussion, optimization, measure theory, simplex, reachability, global optimization

AMS subject classifications. 49J20, 49M30

1. Introduction.

In this paper we study the control of nonlinear diffusion equations by means of an approach that has proved useful in the analysis of the control of nonlinear ordinary differential equations [1] and linear partial differential equations [1], [2], [3]; lately we have extended it to the control of nonlinear diffusion equations with a small nonlinearity [4]. No such assumption, of "small" nonlinearities, will be made in this paper; we deal here with *fully nonlinear* diffusion.

The main approach that is used here is based on an idea of Young [5], consisting of the replacement of classical variational problems by problems in measure spaces; he mentions the possibility of treating these problems using the tools of linear analysis. The extension of these ideas to optimal control problems and the realization that one is dealing with fully linear problems—even if the original problems were nonlinear in the usual sense—are due to us ([1], [6],[7]); it appears we were the first, in [6], to develop linear programs whose solution would give rise to solutions of nonlinear optimal control problems. The reader can consult the review paper [8] for a full bibliography and a historical analysis of these matters.

The present work is not significantly related to our earlier ones [6], [7], [1, Chap. 7], which proved to give rise to unsuitable finite-dimensional approximations and which could not be used to study controllability in any depth; see [1, Chap. 6, §1], for an analysis of this problem.

We should mention also that independently of our work there has been much research on dual methods, especially by the Leipzig school [9]–[12], and by the group at Imperial College [13]. There has also been some exciting work, mainly on the dual approach, by workers involved in stochastic matters; see, for instance, [14]–[16].

In this paper we consider a nonlinear diffusion equation with a boundary control, through which we wish to minimize an integral performance criterion, given that the terminal state should be fixed. We first write some well-known integral relationships satisfied by the solution to this equation and then proceed to transform the problem; instead of minimizing over a set of admissible trajectory-control pairs, we find that it

*Received by the editors December 2, 1991; accepted for publication (in revised form) September 29, 1993.

[†]University of Leeds, Leeds LS2 9JT, England, United Kingdom.

is possible to minimize over a product of two measure spaces. The advantages of the new formulation are:

(i) An automatic existence theory—there always is a minimizer for our measure-theoretical problem.

(ii) The new problem is linear, and then one can use the whole paraphernalia of linear analysis for dealing with such a problem.

(iii) Also, our minimization is *global*—the value reached, say, numerically is close to what one could reasonably call the global infimum of the problem.

The price to pay for these advantages is that the final state is reached only asymptotically—that is, as the number of (linear) constraints associated with the measure-theoretical problem tends to infinity; the situation is similar to our results in the finite-dimensional case; see [1, Chap. 4].

Our main achievement in the first part of the paper—that is, §§2–4—is the construction of a framework for the treatment of nonlinear partial differential equations; in §§2 and 3 by developing the embedding process in Sobolev spaces; and in §4 by developing our main result on approximation, Theorem 1. Our intention here is to repeat in this context what we have achieved with our work in [1] in the case of finite-dimensional systems: to develop a frame of reference on which applications could be based, as has happened in [11], [12].

We present in the rest of the paper two applications of our theory: a reachability and controllability theory for these equations in which we introduce the concept of *asymptotic reachability* and develop a sufficient condition for a state to be asymptotically reachable and a computational method in which we treat the semi-infinite linear-programming problem developed in the paper by means of an infinite-dimensional simplex method; nearly optimal controls can be constructed in this manner.

2. The equations. Consider a boundary control problem. We follow the notation in [17] and [18] and take D, a bounded domain in $\mathbf{R}^n$ with smooth boundary ∂D and T, a positive real number, and define

$$\begin{aligned} &Q_T := D \times (0,T),\\ &\Gamma_T := \partial D \times (0,T),\\ &D_0 := D \times \{0\},\\ &D_T := D \times \{T\}. \end{aligned}$$

We also choose some functions

$$\begin{aligned} &k : \bar{Q}_T \to \mathbf{R}, k \in C^1(\bar{Q}_T);\\ &f : \mathbf{R} \times \mathbf{R}^n \times \bar{Q}_T \to \mathbf{R}, f \in C(\mathbf{R} \times \mathbf{R}^n \times \bar{Q}_T); \end{aligned}$$

a function is said to be differentiable in the closure of a domain if it is uniformly differentiable in the domain itself. Consider the nonlinear diffusion equation

$$u_t(x,t) - \operatorname{div}(k(x,t)\nabla u(x,t)) = f(u(x,t), \nabla u(x,t), x, t), \tag{2.1}$$

for $(x,t) \in Q_T$, with the initial condition

$$u(x,0) = 0, \qquad x \in D, \tag{2.2}$$

and the boundary condition

$$\nabla u \cdot n|_{\Gamma_T} = v; \tag{2.3}$$

here n is the outward normal, and the function $(s,t) \in \Gamma_T \to v(s,t) \in V \subset \mathbf{R}$ is the *control function*, taking values in a *bounded* control set V. This form appears to be as general as it can be made while retaining the possibility of weak solutions and of boundary control by means of heat flow [19], [20].

A pair (u, v) of trajectory function u and control function v is said to be *admissible* if

(i) The function $(x,t) \to u(x,t)$ is a classical solution of (2.1), that is, it is in $C^{2,1}(Q_T) \cap C(Q_T \cup \Gamma_T \cup \bar{D}_0)$ and satisfies (2.1), (2.2), and (2.3).

(ii) The control function is continuous in Γ_T.

(iii) The terminal relationship

$$u(\cdot, T) = g \tag{2.4}$$

is satisfied; g is a given continuous function on D_T.

Conditions for the existence of a classical solution of (2.1)–(2.3) can be found in the literature [17, Chap. VI], [18, Chap. V, VI].

The set of admissible pairs for this problem will be denoted by $\mathcal{F}$ and assumed to be nonempty, at least until §5, where we will be concerned with controllability.

We make the further point that, since the control set V is bounded, that is, there is a constant M_V so that $|v(s,t)| \leq M_V, (s,t) \in \Gamma_T$, there are *bounded* sets $A \subset \mathbf{R}$ and $B \subset \mathbf{R^n}$ so that (see [18])

$$u(x,t) \in A, \quad \nabla u(x,t) \in B \qquad \forall (x,t) \in \bar{Q}_T. \tag{2.5}$$

We must be *very careful* here, since there are many sets A and B which satisfy (2.5); we must choose from those the *minimum sets*, that is, either the intersections $\cap A$ and $\cap B$ of all sets satisfying (2.5) or subsets of them. Thus, every point in our set A (respectively B) will be a state (respectively, a gradient of a state) that can be reached by an admissible control inside the time interval $[0, T]$. This property will, of course, be needed in our proofs on approximation later.

The optimization problem associated with this equation is as follows. Let f_0, f_1 be continuous, nonnegative, real-valued functions on $\mathbf{R}^{2n+2}$, $\mathbf{R}^{n+1}$, respectively; further, we assume that there are constants $h', h'' > 0$ so that

$$f_0(u,w,x,t) \leq h'|u| + h''\|w\|_E, \qquad (u,w,x,t) \in A \times B \times \bar{Q}_T,$$

with the norm the euclidean norm in $\mathbf{R^n}$. Then we wish to find a minimizing pair in $\mathcal{F}$ for the functional

$$J(u,v) := \int_{Q_T} f_0(u(x,t), \nabla u(x,t), x, t)\, dxdt + \int_{\Gamma_T} f_1(v(s,t), s, t)\, dsdt. \tag{2.6}$$

We transform now this problem, with a view at generalization. Let ψ be in $C^1(\bar{Q}_T)$. Then one can show [17], [18] that the classical solution of (2.1)–(2.3), if it exists, satisfies the integral relation

$$\int_{Q_T} [u\psi_t - k\nabla u \nabla \psi + f\psi]\, dxdt = -\int_{\Gamma_T} k\psi v\, dsdt + \int_{D_T} g\psi\, dx, \tag{2.7}$$

for all $\psi \in C^1(\bar{Q}_T)$. We proceed in the next section to transform this problem.

3. Metamorphosis. In general, the minimization of the functional (2.6) over the set $\mathcal{F}$ is not possible—the infimum is not attained at any admissible pair; it is not possible, for instance, to write necessary conditions for this problem. We proceed then to transform it, realizing that a solution of (2.1)–(2.3) defines a linear, bounded, positive functional

$$u(\cdot,\cdot) : F \longrightarrow \int_{Q_T} F(u(x,t), \nabla u(x,t), x, t)\, dxdt$$

in the space $C(\Omega)$ of continuous real-valued functions F, with $\Omega := A \times B \times Q_T$. Also, a control v defines a linear, bounded, positive functional

$$v(\cdot,\cdot) : G \longrightarrow \int_{\Gamma_T} G(v(s,t), s, t)\, dsdt$$

in the space $C(\omega)$ of continuous functions G, $\omega := V \times \Gamma_T$.

By Riesz's theorem, an admissible pair (u, v) defines two Radon measures μ and ν, the first on Ω, the second on ω, so that (2.7) becomes

$$\int_\Omega F_\psi\, d\mu + \int_\omega G_\psi\, d\nu = \int_{D_T} g\psi\, dx := \alpha_\psi, \quad \forall \psi \in C^1(\bar{Q}_T), \tag{3.1}$$

where

$$F_\psi(u, w, x, t) := u\psi_t(x,t) - k(x,t) w \nabla\psi(x,t) + f(u, w, x, t)\psi(x,t), \tag{3.2a}$$

$$G_\psi(v, s, t) := k\psi(x|_{\partial D}, t)v. \tag{3.2b}$$

Thus, the minimization of the functional (2.6) over $\mathcal{F}$ is equivalent to the minimization of

$$I(\mu, \nu) = \mu(f_0) + \nu(f_1), \tag{3.3}$$

where we have written $\mu(f)$ for $\int_\Omega f\, d\mu$ and $\nu(g)$ for $\int_\omega f\, d\nu$, over the set of measures (μ, ν) corresponding to admissible pairs, which satisfy

$$\mu(F_\psi) + \nu(G_\psi) = \alpha_\psi \quad \forall \psi \in C^1(\bar{Q}_T). \tag{3.4}$$

So far, we have not achieved anything new. We consider the extension of our problem; we shall consider the minimization of (3.3) over the set S of all pairs of measures (μ, ν) in $\mathcal{M}^+(\Omega) \times \mathcal{M}^+(\omega)$ satisfying (3.4) plus the extra condition, satisfied of course by the admissible pairs, that these measures project on the (x, t) or (s, t) spaces as the respective Lebesgue measures. (See [1, Chap. 1], for a further discussion of this point.) Thus, if a function $\xi : \Omega \longrightarrow \mathbf{R}$ depends only on (x, t),

$$\mu(\xi) = a_\xi, \tag{3.5a}$$

the Lebesgue integral of ξ over Q_T. Also, if a function $\zeta : \omega \longrightarrow \mathbf{R}$ depends only on (s, t),

$$\nu(\zeta) = b_\zeta, \tag{3.5b}$$

the Lebesgue integral of ζ over Γ_T. Note that (3.5) imply that, writing 1_Ω and 1_ω for the characteristic functions of Ω and ω and L' and L'' for the Lebesgue measures of D and ∂D, respectively,

$$\mu(1_\Omega) = TL', \qquad \nu(1_\omega) = TL''. \tag{3.6}$$

Even if these equalities are a consequence of the equalities (3.5), they may be used to advantage when only a finite number of these are employed, as will be the case in our main approximation scheme.

In the next section we analyze this new optimization problem. We show that it always has at least one optimizer and that—this is most important—admissible trajectory-control pairs that are nearly minimizing for the original problem can be obtained from this construction.

4. Existence and approximation. The proof of the existence of an optimal measure for the functional I defined in (3.3) on the set S of measures in $\mathcal{M}^+(\Omega) \times \mathcal{M}^+(\omega)$ satisfying (3.4)–(3.5) is quite simple, being based on simple compactness properties of the weak*-topology. The proof of the following Proposition is much like that of Theorem II.1 in [1, Chap. 2], and is therefore omitted.

PROPOSITION 1. *There exists an optimal pair* $(\mu^*, \nu^*) \in S$ *that minimizes the functional* I.

The proof of the following proposition, which will be necessary later to prove our main theorem of approximation, can be found in detail in [7, Prop. 2], and will also be omitted; see also [2]. Note that the function $u \to \nabla u$ as a mapping from $H_1(Q_T) \to \{L_2(Q_T)\}^n$ is continuous.

PROPOSITION 2. *The set* $S_1 \subset S$ *of measures* (u, v) *that are piecewise-constant functions on* Ω *and* ω*, respectively and satisfy* (3.4)–(3.5) *is weakly*-dense in* S.

We start now the arduous process of approximation—that is, of building a framework, based on approximation to our main weak problem, so as to construct admissible pairs that nearly minimize the functional (2.6). The first of such constructions involves the possibility of approximating the set S. Let $\{\psi_i, i = 1, 2, \ldots\}$ be a set of functions that is *total* in $C^1(\bar{Q}_T)$, that is, so that, given $\psi \in C^1(\bar{Q}_T)$ and $\epsilon > 0$, there is an integer $N > 0$ and scalars $\alpha_i, i = 1, \ldots, N$, so that

$$\max_{\bar{Q}_T} \left| \psi - \sum_{i=1}^{N} \alpha_i \psi_i \right| < \epsilon,$$

$$\max_{\bar{Q}_T} \left| \psi_t - \sum_{i=1}^{N} \alpha_i \psi_{it} \right| < \epsilon,$$

$$\max_{\bar{Q}_T} \left\| \nabla\psi - \sum_{i=1}^{N} \alpha_i \nabla\psi_i \right\|_E < \epsilon;$$

we shall write

$$F_i := F_{\psi_i}, \qquad G_i := G_{\psi_i}, \qquad \alpha_i := \alpha_{\psi_i}, \qquad \forall i.$$

Further, we shall also take sets of functions $\{\xi_j, j = 1, 2, \ldots\}$ and $\{\zeta_k, k = 1, 2, \ldots\}$ that are total in the respective subspaces of $C(\Omega)$ and $C(\omega)$, writing a_j for a_{ξ_j} and b_k for b_{ζ_k}. We have then our first result of approximation; the proof is much like that of Proposition III.1 in [1, Chap. 3], and will be omitted.

PROPOSITION 3. *Let* M_1*,* M_2*, and* M_3 *be positive integers. Consider the problem of minimizing*

$$(\mu, \nu) \to \mu(f_0) + \nu(f_1) \tag{4.1a}$$

over the set $S(M_1, M_2, M_3)$ of measures in $\mathcal{M}^+(\Omega) \times \mathcal{M}^+(\omega)$ satisfying $\mu(1_\Omega) = TL'$, $\nu(1_\omega) = TL''$, and

$$\begin{aligned} \mu(F_i) + \nu(G_i) &= \alpha_i, & i &= 1, 2, \dots, M_1; \\ \mu(\xi_j) &= a_j, & j &= 1, 2, \dots, M_2; \\ \nu(\zeta_k) &= b_k, & k &= 1, 2, \dots, M_3. \end{aligned} \tag{4.1b}$$

Then, as $M_1, M_2, M_3 \to \infty$,

$$\inf_{S(M_1, M_2, M_3)} [\mu(f_0) + \nu(f_1)] \to \inf_S [\mu(f_0) + \nu(f_1)]. \tag{4.2}$$

We note that

$$\inf_S I \leq \inf_{\mathcal{F}} J$$

and that this may be a strict inequality; see [1, Chaps. 1 and 4], for a discussion of this point.

We have reached the main point of this section. How do we construct suboptimal pairs of trajectories and controls for the functional (2.6)? We shall proceed in several steps:

(i) First we shall obtain optimal measures (μ^*, ν^*) for a problem such as the one in Proposition 3. The existence of such a minimizer follows from the same simple considerations as the existence theorem given in Proposition 1.

(ii) We then obtain a (weak*) approximation to (μ^*, ν^*) by a set of two piecewise-constant functions (u, v) by means of the results given in Proposition 2.

(iii) The control function v obtained above is in $L_2(\Gamma_T)$, that is, for each $t \in (0, T)$, $v(\cdot, t) \in L_2(\partial D)$, since it is piecewise constant and the set Q_T is bounded. It can serve then as boundary function for a *weak solution* of the system (2.1)–(2.3), to be denoted by u_v. This solution will be in $H^1(Q_T)$. Conditions as to the existence of such weak solutions are given in [17], [18].

(iv) Borrowing the term from Rudolph [12], we shall call the pair (u_v, v) of trajectory and control functions *asymptotically admissible* if:

  (a) The control function $v \in L_2(Q_T)$ and $v(s, t) \in V$.

  (b) The trajectory u_v is the weak solution of (2.1)–(2.3) corresponding to the admissible control $v \in L_2(Q_T)$.

  (c) The trajectory function satisfies the constraint (2.7).

  (d) The final value $u_v(\cdot, T)$ of the trajectory function tends in $L_2(D_T)$ to the prescribed function g in (2.4) as $M_1, M_2, M_3 \to \infty$.

(v) We shall prove below that if the numbers M_1, M_2, M_3, are sufficiently large and the weak*-approximation in step (ii) above sufficiently good, then the value at the pair (u_v, v) of the functional J defined by (2.6), $J(u_v, v)$, is close to the $\inf_S I$ and thus is a good suboptimal pair. Note that no use is made of the trajectory u, obtained in step (ii) together with the control v.

We proceed to prove all this.

THEOREM 1. *Let (u_v, v) be the pair constructed as explained above. Then, under the appropriate conditions on the approximations involved,*

(i) *The pair (u_v, v) is asymptotically admissible.*

(ii) *As M_1, M_2, M_3 tend to ∞,*

$$J(u_v, v) \to \inf_S I(\mu, \nu).$$

Proof. (i) Fix $M_1 > 0$ and write $\epsilon := 1/M_1$. Fix also the values of M_2 and M_3. Let (μ^*, ν^*) be the minimizer for the functional (4.1a) over the set $S(M_1, M_2, M_3)$; its existence can be proven by the same arguments used in proving Proposition 1 (see [1, Chap. 3 and 4]).

By Proposition 2, we can find a pair $(u, v) \in S_1$ so that

$$|\{\mu_u(f_0) + \nu_v(f_1)\} - \{\mu^*(f_0) + \nu^*(f_1)\}| < \epsilon; \tag{4.3a}$$

$$|\mu_u(F_i) + \nu_v(G_i) - \alpha_i| < \epsilon, \qquad i = 1, \ldots, M_1. \tag{4.3b}$$

Here we have written (μ_u, ν_v) for the measure in $\mathcal{M}^+(\Omega) \times \mathcal{M}^+(\omega)$ generated by the pair trajectory-control (u, v). We note, further, that these measures satisfy automatically the rest of the relations in (4.1b).

(ii) We proceed to prove now that the pair (u_v, v) defined above is asymptotically admissible; we only have to prove the contention in (iv)(d) above, on the final value $u_v(\cdot, T)$, since the rest of the requirements in (iv) are trivially satisfied. Choose the functions $\psi_i \in C^1(\bar{Q}_T), i = 1, \ldots, M_1$, of the form $\psi_i = \varphi_i + \chi_i$, so that, for $i = 1, \ldots, M_1$

(a) The function $\chi_i \in C^1(\bar{Q}_T)$ is chosen so that the Lebegue measure of the support of χ_i in $\bar{Q}_T$ is not higher than a number $\epsilon' > 0$ to be chosen below and so that the maximum of the numbers

$$|\chi_i(x,t)|, \qquad |\chi_{it}(x,t)|, \qquad \|\nabla\chi_i(x,t)\|_E, \qquad (t,x) \in \bar{Q}_T,$$

is not larger than a number $\varsigma > 0$, to be defined. Then, since $F_{\chi_i} = u\chi_{it} - k\nabla u \nabla \chi_i + f\chi_i$ and putting $k_{\max}$ and $f_{\max}$ for the maximum values of k and f over their respective domains, we can choose ς so that

$$|F_{\chi_i}(x,t)| \leq \varsigma[\operatorname{diam} A + k_{\max} \operatorname{diam} B + +f_{\max}] < 1.$$

We shall call $\theta_i := F_{\chi_i}$, so that $|\theta_i| < 1$ on Ω, and the Lebesgue measure of the support of θ_i is less than $\epsilon' \cdot \operatorname{meas} A \cdot \operatorname{meas} B$, which can be made less than ϵ by choosing ϵ' accordingly.

(b) The function $\varphi_i \in C^1(\bar{Q}_T)$ is chosen so that $\varphi_i(x, T) = 0$ on D and so that

$$\mu_u(F_{\varphi_i}) + \nu_v(G_{\varphi_i}) < \epsilon. \tag{4.4}$$

This is possible since the pair (μ^*, ν^*) assigns to the pair $(F_{\varphi_i}, G_{\varphi_i})$ a value near $\int_{D_T} g\varphi_i \, dx = 0$, by (4.3b); note that the functions F_ψ, G_ψ and the integral $\int_{D_T} g\psi \, dx$ are linear in the function ψ and that the set $\{\psi_j\}$ is total, so that the functions φ_i are linear combinations of these functions. The relation (4.4) follows then from Proposition 2.

Thus, we have

$$F_i = F_{\varphi_i} + \theta_i,$$

with $\varphi_i \in C^1(\bar{Q}_T)$, $\varphi_i(x,T) = 0$, and $|\theta_i| < 1$ on Ω, with the Lebesgue measure of $\operatorname{supp} \theta_i < \epsilon, i = 1, \ldots, M_1$, and (4.4) is satisfied.

Now we obtain the weak trajectory u_v corresponding to the control v, as explained above. Then, writing μ_{u_v} for the measure corresponding to this trajectory function u_v,

$$\mu_{u_v}(F_i) + \nu_v(G_i) = \int_{D_T} u_v(\cdot, T)\psi_i \, dx, \qquad i = 1, \ldots, M_1, \tag{4.5}$$

by definition of a weak solution. But the pair (μ^*, ν^*) satisfies (4.1):

$$\mu^*(F_i) + \nu^*(G_i) = \int_{D_T} g\psi_i \, dx, \qquad i = 1, \ldots, M_1. \tag{4.6}$$

Then

$$\begin{aligned} |\int_{D_T} (u_v(\cdot, T) - g)\psi_i \, dx| &= |(\mu^* - \mu_{u_v})(F_i) + (\nu^* - \nu_v)(G_i)| \\ &\le |\mu_u(F_i) + \nu_v(G_i) - \alpha_i)| + |(\mu_u - \mu_{u_v})(F_i)| \\ &\le 3\epsilon, \qquad i = 1, \ldots, M_1, \end{aligned} \tag{4.7}$$

by (4.3b) and

$$|(\mu_u - \mu_{u_v})(F_i)| \le |(\mu_u - \mu_{u_v})(F_{\varphi_i})| + |(\mu_u - \mu_{u_v})(\theta_i)| \le 2\epsilon, \tag{4.8}$$

because of (4.4) and the fact that

$$\mu_{u_v}(F_{\varphi_i}) + \nu_v(G_{\varphi_i}) = 0.$$

It follows from (4.7) that as $M_1 \to \infty$,

$$\|u_v(\cdot, T) - g\|_{L_2} \to 0;$$

the proof is as in [1, p. 47]. Note that we need also $M_2, M_3 \to \infty$, a requirement of the proof of Proposition 2 in [2].

(iii) The second contention of the theorem follows from the fact that

$$\begin{aligned} |J(u_v, v) - (\mu^*(f_0) + \nu^*(f_1))| &\le |J(u_v, v) - (\mu_u(f_0) + \nu_v(f_1))| \\ &+ |(\mu_u(f_0) + \nu_v(f_1)) - (\mu^*(f_0) + \nu^*(f_1))| \le 2\epsilon, \end{aligned}$$

by (4.3a) and the fact that

$$J(u_v, v) - (\mu_u(f_0) + \nu_v(f_1)) = (\mu_{u_v} - \mu_u)(f_0), \tag{4.9}$$

since

(a) The function f_0 satisfies the condition $f_0(u, w, x, t) \le h'|u| + h''\|w\|_E$ on Ω, so that

$$|(\mu_{u_v} - \mu_u)(f_0)| \le |(\mu_{u_v} - \mu_u)|(h'\vartheta + +h''\|\sigma\|_E), \tag{4.10}$$

where $\vartheta(u, w, x, t) := u, \sigma(u, w, x, t) := w, (u, w, x, t) \in \Omega$. We are assuming, without loss of generality, that $u \ge 0$ on Ω.

(b) Further, choose $\psi \in C^1(\bar{Q}_T)$ of the form

$$\psi = \sum_{i=1}^{K} \beta_i \psi_i, \qquad \sum_{i=1}^{K} |\beta_i| \le \lambda, \tag{4.11}$$

with K a fixed integer not higher than M_1 and

$$\lambda \le \max\left(\frac{1}{16h'}, \frac{1}{16h''}\right).$$

Then, by (4.8),

$$|(\mu_{u_v} - \mu_u)(F_\psi)| \leq 2\lambda\epsilon.$$

Further, choose the coefficients $\beta_i, i = 1, \ldots, K$, so that

$$\vartheta(|\psi_t| - 1) \leq \epsilon'/3 \quad |k|\,\|\nabla\psi\|_E \leq \epsilon'/3 \quad |f\psi| \leq \epsilon'/3,$$

which implies

$$|\theta(\psi_t - 1) + k\sigma\nabla\psi - f\psi| \leq \epsilon',$$

on Ω; the number ϵ' will be determined below. Then

$$\begin{aligned} |(\mu_{u_v} - \mu_u)\vartheta| &\leq |(\mu_{u_v} - \mu_u)(\vartheta\psi_t - k\sigma\nabla\psi)| \\ &\quad + |(\mu_{u_v} - \mu_u)(\vartheta(\psi_t - 1) + k\sigma\nabla\psi - f\psi)| \\ &\leq 2\lambda\epsilon + \epsilon' L(Q_T) \leq \epsilon/4h', \end{aligned} \tag{4.12}$$

if we take $\epsilon' \leq \epsilon/8L(Q_T)h'$; as above, $L(Q_T)$ is the Lebesgue measure of the domain Q_T.

Note that (4.12) could have been proved for the negative $\mu_u - \mu_{u_v}$ of the measure for which it was actually proved, so that, finally,

$$|(\mu_u - \mu_{u_v})|(\vartheta) \leq \epsilon/2h'.$$

Similarly, we can prove that

$$|(\mu_{u_v} - \mu_u)|\,\|\sigma\|_E \leq \epsilon/2h'',$$

so that, by (4.10),

$$|(\mu_{u_v} - \mu_u)(f_0)| \leq \epsilon.$$

The second contention of the theorem follows. □

Remark. Note that it is not necessary to include in our set of equations (4.1) relationships reflecting equations of the form $\int_{Q_T}(u\Delta\psi + w\nabla\psi)\,dxdt = 0$, for $\psi_n = 0$ on Γ_T, relating w to ∇u; one can show, by methods similar to those used in the proof above, that the function $(x,t) \to w(x,t)$ tends in $\{\mathrm{L}_2(Q_T)^n\}$ to $(x,t) \to \nabla u_v(x,t)$, as M_1, M_2, M_3 tend to ∞.

In the next two sections we present applications of this theory, first to matters on controllablity and reachability and then to matters computational.

5. Controllability and reachability. The—apparently—nonlinear control and dynamical problem (2.1)–(2.4) has many facets that one could explore using the *linear* theory that we have developed. In this Section we examine reachability and controllability. A good review paper on the reachability and controllability of distributed systems is [21], where the concepts of approximate and exact reachability and controllability are defined and illustrated. The most interesting development on nonlinear controllability has been by Henry [22]; his approach has been developed further by Zhou in [23] and [24]. This method has been applied to equations simpler than ours and is based on the development of an associated linear system, whose controllability can be studied in the usual way. It is not related to our approach, even if some results are not dissimilar.

According to our results in the previous section, we are entitled to make the following definition.

DEFINITION. *The state g is said to be* asymptotically reachable *(from the origin, by the system* (2.1)–(2.3)*) if there is an integer $N > 0$ so that the system* (4.1b) *has a solution in $\mathcal{M}^+(\Omega) \times \mathcal{M}^+(\omega)$ for all integers M_1, M_2, M_3 that are larger than N.*

In other words, according to our results on approximation, if a state g is asymptotically reachable one can, given $\epsilon > 0$, find a control v_ϵ that generates a *weak solution* u_{v_ϵ} so that the $L_2(D_T)$-norm of the difference $u_{v_\epsilon}(T) - g$ is not higher than ϵ. This of course does not mean that there is a control by means of which the state g can be reached exactly, so this is a weaker concept than that of exact reachability. It is just as strong as approximate reachability—a concept based on the classical solutions—for systems for which these and weak solutions coincide.

It is an open problem whether one can find a control to reach exactly a state g based on the construction of the family of controls $\{v_\epsilon, \epsilon > 0\}$. Our philosophy is that it is sufficient for any foreseeable application to have the possiblity of reaching a final state as closely as necessary. The mathematical problem, essentially that of exact controllability, remains. It does not appear that our methods can throw any new light on this problem.

We shall follow the mathematical structure developed in [1, Chap. 6], and applied there to finite-dimensional systems. Consider then the real vector space $\mathcal{L}$ of pairs (f, h) of continuous functions $f : \Omega \to \mathbf{R}$ and $h : \omega \to \mathbf{R}$. The linear operations are defined as follows:

$$\lambda(f,h) := (\lambda f, \lambda h), \qquad (f,h) + (v,w) := (f+v, h+w),$$

for all real λ and appropriate functions v and w.

Further, we define an *ordering* on the space $\mathcal{L}$ by defining its positive cone, $\mathcal{C}$:

$$\mathcal{C} = \{(f,h) \in \mathcal{L} : (f,h) \geq 0\} = \{(f,h) \in \mathcal{L} : f \geq 0 \,\text{on}\, \Omega, h \geq 0 \,\text{on}\, \omega\}.$$

It is trivial to show that this defines an ordering compatible with the linear structure. It is also compatible with the topological structure generated by the uniform norm

$$||(f,h)|| = \sup_{\Omega} |f| + \sup_{\omega} |h|.$$

Thus, the space $\mathcal{L}$ is an *ordered topological vector space*, and we can apply to our problem of existence a particular theorem, due to Schaefer [25, Thm. 5.4]. To do this, consider the functional Υ defined by the equalities (4.1b) on the subspace $\mathcal{E}$ of $\mathcal{L}$ spanned by the functions appearing in (4.1b), that is,

$$\text{(5.1)} \quad \mathcal{E} := \left\{ (f,h) \in \mathcal{L} : (f,h) = \sum_{i=1}^{M_1} A_i(F_i, G_i) + \sum_{j=1}^{M_2} B_j(\xi_j, 0) + \sum_{k=1}^{M_3} C_k(0, \zeta_k) \right\},$$

for real numbers A_i, B_j, C_k. Then, according to our results in [1, Chap. 6]—a trivial development of Schaefer's results [25]—we have the following proposition.

PROPOSITION 4. *The functional Υ defined by the equalities* (4.1b) *on the subspace $\mathcal{E}$ can be extended to the whole of the linear space $\mathcal{L}$ as a positive functional if and only if the functional Υ is bounded above on $\mathcal{E} \cap (\mathcal{V} - \mathcal{C})$, where $\mathcal{V}$ is a suitable neighborhood of zero in $\mathcal{L}$. Further, this is true if and only if the functional Υ is bounded above on the subset U of $\mathbf{R}^{M_1+M_2+M_3}$ defined by*

$$U := \Big\{ \{A_i, i = 1, \ldots, M_1, B_j, j = 1, \ldots, M_2, C_k, k = 1, \ldots, M_3\} :$$

$$\left.\sum_{i=1}^{M_1} A_i(F_i,G_i)+\sum_{j=1}^{M_2} B_j(\xi_j,0)+\sum_{k=1}^{M_3} C_k(0,\zeta_k)<1\right\}$$

$$=\left\{\{A_i, i=1,\dots,M_1, B_j, j=1,\dots,M_2, C_k, k=1,\dots,M_3\}:\right.$$

$$\left.\sum_{i=1}^{M_1} A_i(F_i+G_i)+\sum_{j=1}^{M_2} B_j\xi_j+\sum_{k=1}^{M_3} C_k\zeta_k<1 \quad \text{on} \quad \Omega\times\omega\right\}; \tag{5.2}$$

that is, the state g is asymptotically reachable by the system (2.1)–(2.3) *if and only if there is a constant D so that*

$$\{A_i, i=1,\dots,M_1, B_j, j=1,\dots,M_2, C_k, k=1,\dots,M_3\}\in U$$

$$\Rightarrow \sum_{i=1}^{M_1} A_i\alpha_i+\sum_{j=1}^{M_2} B_j a_j+\sum_{k=1}^{M_3} C_k b_k<D. \tag{5.3}$$

Proof. The firt part follows directly from [25, Thm. 5.4]. The second part follows from simply writing down the form of a neighbourhood $\mathcal{V}$; take $\epsilon>0$ and then define the neighborhood $\mathcal{V}_\epsilon$ by

$$(f,h)\in\mathcal{V}_\epsilon \Rightarrow \sup_{\Omega}|f|+\sup_{\omega}|h|<\epsilon,$$

so that

$$\mathcal{V}_\epsilon-\mathcal{C}=\{(f,h)\in\mathcal{L}: f+h<\alpha \quad \text{on} \quad \omega\times\Omega\}, \tag{5.4}$$

for some scalar α. The expression (5.2) follows by simply normalizing as in Proposition VI.3 in [1], and (5.3) follows fom the definition of the functional Υ.

Finally, we can develop our main result of this section.

(i) The set $U\subset\mathbf{R}^{\mathbf{M_1+M_2+M_3}}$ is a convex set that contains the origin. It will normally be unbounded.

(ii) The expression (5.3) defines a hyperplane Π in $\mathbf{R}^{\mathbf{M_1+M_2+M_3}}$ by

$$\sum_{i=1}^{M_1} A_i\alpha_i+\sum_{j=1}^{M_2} B_j a_j+\sum_{k=1}^{M_3} C_k b_k=D. \tag{5.5}$$

Proposition 4 says that (4.1b) has a solution in $\mathcal{M}^+(\Omega)\times\mathcal{M}^+(\omega)$ if and only if the set U is wholly at one side of the hyperplane Π.

(iii) We shall make some assumptions now as to the nature of the functions $\xi_j, j=1,\dots,M_2; \zeta_k, k=1,\dots,M_3$. First, we shall partition the (x,t) and (s,t) spaces and define them as the characteristic functions of the sets in the partitions. Of course, they will not be continuous but can be chosen as lower-semicontinuous, and our arguments above follow without much change; see [1, Chap. 5] for a discussion of this and related problems. Then we shall assume, without loss of generality, that these partitions have been chosen so that $M_1=M_2=M_3\equiv M$.

(iv) Put now $A_i=B_i=C_i\equiv 0 \ \ \forall i\neq\ell$. What does the set U and the hyperplane Π look like, in this subspace with variables A_ℓ, B_ℓ, C_ℓ? Let

$$\begin{aligned}
\mathcal{G}'_\ell &:= \{((x,t),(s,t))\in\Omega\times\omega:\xi_\ell=1,(F_\ell+G_\ell)>0)\},\\
\mathcal{H}'_\ell &:= \{((x,t),(s,t))\in\Omega\times\omega:\xi_\ell=1,(F_\ell+G_\ell)<0)\},\\
\mathcal{G}''_\ell &:= \{((x,t),(s,t))\in\Omega\times\omega:\zeta_\ell=1,(F_\ell+G_\ell)>0\},\\
\mathcal{H}''_\ell &:= \{((x,t),(s,t))\in\Omega\times\omega:\zeta_\ell=1,(F_\ell+G_\ell)<0)\}
\end{aligned} \tag{5.6}$$

and

$$\begin{aligned}
r'_\ell &:= \inf_{\mathcal{G}'_\ell}(1/(F_\ell + G_\ell)),\\
s'_\ell &:= \sup_{\mathcal{H}'_\ell}(1/(F_\ell + G_\ell)),\\
r''_\ell &:= \inf_{\mathcal{G}''_\ell}(1/(F_\ell + G_\ell)),\\
s''_\ell &:= \sup_{\mathcal{H}''_\ell}(1/(F_\ell + G_\ell)).
\end{aligned}\tag{5.7}$$

In the (A_ℓ, B_ℓ) plane, U projects as a set bounded by two straight lines, one from the point P_ℓ of coordinates $A_\ell = 0, B_\ell = 1$ to the point with coordinates $A_\ell = r'_\ell, B_\ell = 0$ and another from the same point P_ℓ to the point with coordinates $A_\ell = s'_\ell, B_\ell = 0$.

In the (A_ℓ, C_ℓ) plane, U projects as a set bounded by two straight lines, one from the point Q_ℓ of coordinates $A_\ell = 0, C_\ell = 1$ to the point with coordinates $A_\ell = r''_\ell, B_\ell = 0$ and another from the same point Q_ℓ to the point with coordinates $A_\ell = s''_\ell, B_\ell = 0$.

The hyperplane Π projects as lines in the (A_ℓ, B_ℓ) and (A_ℓ, C_ℓ) planes. Consider lines parallel to these but passing through the points P_ℓ and Q_ℓ respectively. They cut the A_ℓ axis at the points

$$A'_\ell = \frac{a_\ell}{\alpha_\ell}, \qquad A''_\ell = \frac{b_\ell}{\alpha_\ell},\tag{5.8}$$

respectively. We have thus our main theorem.

Theorem 2. *The system* (4.1b) *has a solution in* $\mathcal{M}^+(\Omega) \times \mathcal{M}^+(\omega)$ *if*

$$\begin{gathered}
A'_\ell > r'_\ell \quad \text{if} \quad A'_\ell > 0, \quad A'_\ell < s'_\ell, \quad \text{if} \quad A'_\ell < 0\\
A''_\ell > r''_\ell \quad \text{if} \quad A''_\ell > 0, \quad A''_\ell < s''_\ell, \quad \text{if} \quad A''_\ell < 0,\\
\forall \ell = 1, \dots, M.
\end{gathered}\tag{5.9}$$

Proof. It follows from the remarks above and the convexity of the set U. □

Corollary 1. *The state g is asymptotically reachable if and only if there is an integer $N > 0$ so that the conditions* (5.9) *of Theorem* 2 *are true for all $M > N$.*

What can go wrong? We can always satisfy (5.9) by making α_ℓ sufficiently small, unless $r'_\ell, r''_\ell, |s'_\ell|, |s_\ell|''$ are actually ∞ for some values of ℓ—because the function $F_\ell + G_\ell$ does not take the appropriate sign on the appropriate subset of $\Omega \times \omega$—or because they tend to ∞ as $M \to \infty$. We shall examine these possibilities further below.

We introduce now a concept of *controllability.* Since the control set is bounded, the trajectories will be in a bounded set A; see (2.5). Thus we cannot hope to reach the whole of the state space and should be content with the following.

Definition. *A system such as* (2.1)–(2.3) *is* asymptotically controllable *if there is a ball centered at the origin in $L_2(D_T)$ so that every state $g \in C(D_T)$ in this ball can be reached asymptotically.*

We have then the following corollary.

Corollary 2. *There is a control set $V \subset \mathbf{R}$ so that the system* (2.1)–(2.3) *is asymptotically controllable. This control set is defined in general in the paragraph after* (2.3).

Proof. If the quantities (5.7) stay bounded as $M \to \infty$, we can choose the moments α_ℓ sufficiently small so that the inequalities of Theorem 2 are satisfied. Consider the first quantity in (5.7). It follows from (3.1) and (3.2) that

$$F_\ell + G_\ell = u\psi_\ell - kw\nabla\psi_\ell + f\psi_\ell + k\psi_\ell|_{\Gamma_T} v;$$

provided that ψ_ℓ is nonzero on Γ_T, we can choose V so that this function takes positive values on $\mathcal{G}'_\ell$, bounded away from zero, for all values of the index ℓ; similarly for the other functions and sets in (5.6), (5.7). For such a control set V the system (2.1)–(2.3) is asymptotically controllable. □

6. Numerical results. The system (4.1) is a semi-infinite linear-programming problem, since the unknown is in $\mathcal{M}^+(\Omega) \times \mathcal{M}^+(\omega)$ but there are only a finite number of constraints. There are several methods for treating numerically such problems [26], [27], even though it appears that only Rudolph's [11], [12], [28], [29] and Hoffman-Klostermair's [30] methods can be assured to converge to the global minimizer; these are simplex methods working in infinite dimensions. We have chosen here Rudolph's method, partly because its author has been very successful in using our finite-dimensional framework for the estimation of optimal controls [11], [12].

To solve these problems we could work in $\mathcal{M}^+(\Omega \cup \omega)$, and indeed this is the most convenient way for work involving discretized approximations; we have done just that in a paper involving the control of variational inequalities, to be published elsewhere. Here, however, it is more convenient to work in $\mathcal{M}^+(\Omega \times \omega)$, identifying $(\mu, \nu) \to \mu \times \nu$, so that

$$\mu \times \nu(f+g) = \mu(f) + \nu(g), \qquad f \in C(\Omega), g \in C(\omega);$$

note that the Lebesgue measures L', L'' are equal to unity and $T = 1$. Also, the conditions (3.6) take the form

$$\mu \times \nu(1_{\Omega\times\omega}) = T = 1.$$

We have added this equality to our constraints because it improved the accuracy of the computation.

The system we have chosen is

$$\begin{aligned} u_t(x,t) - \operatorname{div}\left(\frac{x^2}{1+t^2}\nabla u(x,t)\right) &= \frac{\|\nabla u(x,t)\|_E^2 + 1}{1 + x^2 + u(x,t)^2}, \\ u_x(0,t) = 0, \quad u_x(1,t) &= v(t), \\ t \in (0,1), \quad x &\in (0,1); \end{aligned} \tag{6.1}$$

that is, $k(x,t) := x^2/(1+t^2)$, $f(u,w,x,t) := (\|w\|_E^2 + 1)/(1+x^2+u^2)$. We have taken $V = [-10, 10]$, $g(x) = 0.075$, $x \in [0,1]$; and we wish to minimize

$$J(u,v) = \int_{Q_T} \frac{u(x,t)^2 + \|\nabla u(x,t)\|_E^2}{1 + \sin^2(tu(t,x))}\, dt;$$

that is, $f_0(x,w,x,t) := (u^2 + \|w\|_E^2)/(1 + \sin^2(tu))$, and $f_1 \equiv 0$. The boundary ∂D is composed of two points only, of which only one, the one at $x = 1$, plays an active role, the control being the heat flow at that point.

The functions ψ in (3.2) were chosen of the form $\psi(x,t) = t^p \cos(\ell\pi x) + q(t)$ or $\psi(x,t) = t^p \sin(h\pi x) + q(t)$; the functions q are test functions introduced to improve the behaviour at $x = 1$ for the determination of an initial solution, as explained below. Ten such functions ψ were chosen, with values of $p = 1, 2$ and $\ell = 1, 2, 3$, $h = 1, 2$. The 16 functions ξ were chosen by dividing the square $[0,1] \times [0,1]$ into 16 equal squares, the functions ξ being the characteristic functions of the individual squares. The condition (3.5b) for the functions ζ is satisfied automatically—thus the advantage of this formulation based on product measures. Thus the total number of constraints m equals $m = 1 + 10 + 16 = 27$. The computational method consists of three steps:

(i) The most difficult problem encountered was that of finding an initial solution from which, in part (ii) of the method, one can iterate toward the minimum. This was done here by means of a finite-dimensional linear program, obtained by discretizing all the variables of the problem. It was necessary to find an initial solution by first solving for some of the parameters, thus the need of the functions q introduced above, because the (rather rough) discretization tended to make the problem unfeasible. Then a finite-dimensional simplex program was run, of rather small size, with 677 variables and, of course, 27 constraints. Only the first phase—the one that produces a feasible solution—was run. It is usual in these problems to use a discretized solution as an initial one; see [11], [12], [26, Chap. 5, 6].

(ii) Then the simplex algorithm of Rudolph was run using the output of step (i) as initial solution, and after 87 iterations it converged to a value of 0.202247; it had started with a value of 1.93919. This is a numerical estimation of the *global minimum*.

(iii) Once the optimization is performed, a nearly-optimal control v can be constructed; the method is shown in detail in [1, Chap. 5]. The graph of the resulting control is shown in Fig. 1.

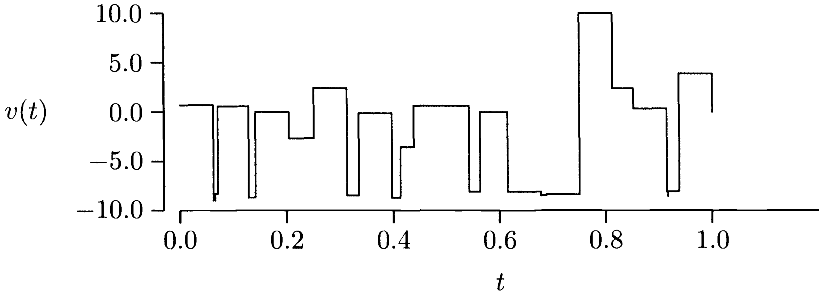


FIG. 1. *Graph of the nearly-optimal control v. Note that the control set is $V = [-10, 10]$.*

We mention, finally, that our approach can be applied to other nonlinear partial differential equations and to variational inequalities.

REFERENCES

[1] J. E. RUBIO, *Control and Optimization: the Linear Treatment of Nonlinear Problems*, Manchester University Press, Manchester, and John Wiley, New York and London, 1986.

[2] A. KAMYAD, J. E. RUBIO AND D. A. WILSON, *Optimal control of the multidimensional diffusion equation*, J. Optim. Theory Appl., 70 (1991), pp. 191–209.

[3] A. KAMYAD, J. E. RUBIO AND D. A. WILSON, *An optimal control problem for the multidimensional diffusion equation with a general control variable*, J. Optim. Theory Appl., 75 (1992), pp. 211–230.

[4] J. E. RUBIO AND A. V. HOLDEN, *The optimal control of an excitable neural fiber*, in Nonlinear Wave Processes in Excitable Media, A. V. Holden, M. Markus, and H. G. Othmev, eds., Plenum Press, New York, 1991, pp. 115–122.

[5] L. C. YOUNG, *Calculus of Variations and Optimal Control Theory*, Saunders, Philadelphia, 1969.

[6] J. E. RUBIO, *Existence and approximation for control problems in Hilbert spaces*, J. Optim. Theory Appl., 69 (1979), pp. 419–427.

[7] ———, *An existence theorem for control problems in Hilbert spaces*, Bull. London Math. Soc., 9 (1977), pp. 70–78.

[8] ———, *Modern trends in the calculus of variations and optimal control theory*, in Vortragsauszüge, Mathematiker Kongress 1990, R. Geise et al, eds., Mathematische Gesellschaft, Berlin, 1990, pp. 128–133.

[9] R. KLÖTZLER, *Dualität bei Steuerungsproblemen and zugeordneten Flußproblemen* I, II, Z. Anal. Anwendungen (1) 4 (1982), pp. 45–47, and 2 (1) (1983), pp. 57–74.

[10] ———, *Flow problems and duality*, in Mathematical Control Theory, C. Olech, B. Jacubczyk, and J. Zabzyk, eds, Polish Scientific Publishers, Warsaw, 1985.

[11] H. RUDOLPH, *Global solution in optimal control via SILP*, in Lecture Notes in Control and Inform. Sci. 143, Springer-Verlag, New York, 1990, pp. 394–402.

[12] ———, *The SILP relaxation method in optimal control theory: general boundary conditions*, I, II. Z. Anal. Anwendungen, 11 (1992), 143–151, 431–436.

[13] R. B. VINTER AND R. M. LEWIS, *The equivalence of the strong and weak formulations for certain problems in optimal control*, SIAM J. Control Optim., 16 (1978), pp. 57–74.

[14] W. H. FLEMING, *Generalized solutions in optimal control theory*, in Differential Games and Control Theory, E. O. Roxin, P. Liu, and R. L. Sternberg, eds., Dekker, New York, pp. 147-165.

[15] ZL. H. ZHU, *Convex duality and generalized solutions in an optimal control problem for a stopped process: the deterministic model*, SIAM J. Control Optim., 30 (1992), pp. 465–476.

[16] O. HERNÁNDEZ-LERMA AND D. HERNÁNDEZ-HERNÁNDEZ, *Linear programming and discounted dynamic programs with unbounded costs*, preprint.

[17] V. P. MIKHAILOV, *Partial Differential Equations*, MIR, Moscow, 1978.

[18] O. A. LADYŽENSKAYA, *Linear and quasilinear equations of parabolic type*, Amer. Math. Soc., Providence, 1968.

[19] J. L. LIONS, *Quelques Méthodes de Résolution des Problèmes aux Limites non Linéaires*, Dunod, Paris, 1969.

[20] W. E. FITZGIBON III AND H. F. WALKER, EDS, *Nonlinear Diffusion*, Pitman, London, 1977.

[21] D. L. RUSSELL, *Controllability and stabilizability theory for linear partial differential equations: recent progress and open questions*, SIAM Rev., 20 (1978), pp. 639–739.

[22] J. HENRY, *Étude de la controlabilité de certain équations paraboliques non-linéaires*, Thèse d'état, Paris, 1978.

[23] H. X. ZHOU, *Approximate controllability for a class of semilinear abstract equations*, SIAM J. Control Optim., 21 (1983), pp. 551-565.

[24] ———, *Controllability properties of linear and semilinear abstract control systems*, SIAM J. Control Optim., 22 (1984), pp. 405–422.

[25] H. H. SCHAEFER, *Topological Vector Spaces*, Macmillan, New York, 1966.

[26] K. GLASHOF AND S. GUSTAFSON, *Linear Optimization and Approximation*, Springer-Verlag, Berlin, 1983.

[27] E. J. ANDERSON AND A. B. PHILPOT, *Infinite Programming*, Springer-Verlag, Berlin, 1985.

[28] H. RUDOLPH, *Zur Approximation semiinfiniter Programme*, Wiss. Z. Karl-Marx-Univ. Leipzig, Math-Natur. Reine, 5 (1978), pp. 501–508.

[29] ———, *Simplexalgorithmus der semiinfiniten linearen Optimierung*, Wiss. Z. TH Leuna-Merseburg Tech. Hochsch., 29 (1987), pp. 782-806.

[30] K.-H HOFFMANN AND A. KLOSTERMAIR, *A semiinfinite linear programming procedure and applications to approximation problems in optimal control*, in Approximation Theory II, Academic Press, New York, 1976.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 33, No. 1, pp. 323–345, January 1995

© 1995 Society for Industrial and Applied Mathematics

016

ASYMPTOTICAL STUDY OF PARAMETER TRACKING ALGORITHMS*

BERNARD DELYON† AND ANATOLI JUDITSKY†

Abstract. This paper addresses the problem of tracking random drifting parameters of a linear regression system. The asymptotic properties of several estimation algorithms in the limit of slow drift are studied. The basic tool is the central limit theorem for a class of stochastic difference equations established under weak conditions on disturbances and observations. The estimates of the rate of convergence obtained in the paper allow the asymptotically optimal algorithms to be developed.

Key words. tracking algorithms, stochastic averaging, normal approximation

AMS subject classifications. 93E10, 93E11, 62L20, 60G35

1. Introduction. In this paper we consider the system of the form

$$y_t = \theta_{t-1}^T \varphi_t + e_t,$$

where e_t is a random noise, $\varphi_t \in \mathbf{R}^N, y_t \in \mathbf{R}$ are observed input and output, and θ_t is an unknown time-varying parameter vector. We assume that θ_t is random and varies slowly, i.e.,

$$\theta_t = \theta_{t-1} + \gamma w_t, \tag{1}$$

where γ is a small parameter.

To estimate the time-varying parameter θ_t in many cases the following general-form identification algorithm can be used (see [11] for a recent review):

$$\hat{\theta}_t = \hat{\theta}_{t-1} + \mu L_t(y_t - \varphi_t^T \hat{\theta}_{t-1}), \qquad \hat{\theta}_t \in \mathbf{R}^N. \tag{2}$$

The way in which the vector gain L_t is updated is characteristic for each particular method. Here μ is a small positive real number. Our paper is devoted essentially to the asymptotical (as $\mu \to 0$ and $\gamma \to 0$) study of the stationary law of the error $\hat{\theta}_t - \theta_t$ of the algorithm. More precisely, our objectives are to show that for a variety of methods the distribution of the error converges to the normal law and to develop asymptotically optimal versions of the algorithms.

Let $\Gamma = \Gamma^T > 0$. In what follows we consider the following gain constructions:

1. Widrow's least mean squares algorithm (LMS):

$$L_t = \Gamma \varphi_t \quad \text{or} \quad L_t = \frac{\Gamma \varphi_t}{1 + \mu \varphi_t^T \Gamma \varphi_t}.$$

2. Normalized least mean squares algorithm (NLMS):

$$L_t = \Gamma \varphi_t / (1 + |\varphi_t|^2).$$

3. Recursive least squares "forgetting factor" algorithm (RLS):

$$\begin{aligned} R_t &= (1-\mu) R_{t-1} + \mu \varphi_t \varphi_t^T, \quad R_0 = \varrho I, \ \varrho > 0, \\ L_t &= R_t^{-1} \varphi_t. \end{aligned}$$

* Received by the editors October 21, 1992; accepted for publication (in revised form) October 15, 1993.

† Institute de Recherche en Informatique et Systèmes Aleatoires, Campus de Beaulieu, 35042 Rennes, France (Delyon@irisa.fr and Anatoli.Iouditski@irisa.fr).

4. "Stabilized" least squares (LS) algorithm:

$$\begin{aligned} L_t &= \Gamma_t \varphi_t, \\ \Gamma_t &= (\mu \rho^{-1} R_t + \Gamma^{-1})^{-1}, \\ R_t &= (1-\rho) R_{t-1} + \rho \varphi_t \varphi_t^T \,, \qquad R_0 = 0, \quad \text{where } 0 < \rho < 1. \end{aligned}$$

Comment. The algorithms above are well known (cf. [11]) except, probably, the stabilized LS. This method was designed as a regularized version of both LMS and forgetting factor algorithms (see [7] for details).

The main theoretical issue of the paper is the central limit theorem (CLT) for the stochastic difference equations developed in §2. It provides an *infinite horizon* approximation for the asymptotic distribution of the error $\hat{\theta}_t - \theta_t$ of equation (2) assuming the stationarity of the sequence (φ_t, e_t, w_t). Next we provide a guide to the application of this result to estimation algorithms. In §3 we consider the algorithms described above and develop conditions of the CLT and expressions for the asymptotic error covariance.

2. Main Result.

2.1. CLT for the stationary case. As usual, we suppose we have a probability space $\{\Omega, \mathcal{F}, P\}$. Let us consider a process (Δ_t), $\Delta_t \in \mathbf{R}^N$, which is generated with the following equation:

$$\Delta_t = (I - \mu P_t)\Delta_{t-1} + \mu \zeta_t \qquad \Delta_0 \in \mathbf{R}^N, \tag{3}$$

where $\mu \leq \mu_0$ is a scalar coefficient, and (P_t) and (ζ_t) are random processes valued in $\mathbf{R}^{N\times N}$ and $\mathbf{R}^N$, respectively. We are going to prove the CLT for the solution Δ_t of (3) which will be quite useful in the study of the estimation algorithms. Let us allow the processes P_t and ζ_t to depend on μ, and P_t can be decomposed in the following way:

$$P_t = P_t^0 + P_t', \tag{4}$$

$$\zeta_t = \zeta_t^0 + \zeta_t', \tag{5}$$

where P_t^0 and ζ_t^0 do not depend on μ. Suppose that there exists a strictly stationary process (X_t) and $\mathcal{F}_t = \sigma\{\dots, X_t\}$ such that $P_t^0, \zeta_t^0, P_t', \zeta_t' \in \mathcal{F}_t$ (for example, in the state-space model X_t contains all the past "innovations"). Consider the following assumptions:

[A1] $P_t^0 = (P_t^0)^T \geq 0$, $EP_1^0 > 0$, $E|P_1^0|^2 < \infty$.

[A2] (P_t^0) and (ζ_t^0) are strictly stationary ergodic processes.

[A3] $\sum_{s=0}^{\infty} (E|E(\zeta_s^0|\mathcal{F}_0)|^2)^{1/2} < \infty$.

[A4] $\lim_{\mu\to 0} \mu^{-1/2} \sup_t E(|\zeta_t'|) = 0$.

[A5] There exists a stationary ergodic process p_t' which satisfies

$$\begin{aligned} &|P_t'| \leq p_t', \\ &\lim_{\mu_0 \to 0} E\left(\sup_{\mu\leq\mu_0} [p_t']^2\right) = 0, \quad t \geq 0, \\ &E([p_t']^2) < \infty, \quad \mu \leq \mu_0. \end{aligned}$$

THEOREM 1. *Let assumptions* [A1]–[A5] *hold. Then for system* (3) *with any initial condition* Δ_0

$$\lim_{\mu\to 0} \lim_{t\to\infty} \mu^{-1/2} \Delta_t = \mathcal{N}(0, V) \quad \textit{in distribution,} \tag{6}$$

i.e., $\mu^{-1/2}\Delta_t$ *converges in distribution to the Gaussian random variable with zero mean and covariance* V. *Here* V *is a solution of the Lyapunov equation*

$$BV + VB = S_0, \tag{7}$$

where $B = EP_1^0$ *and*

$$S_0 = E\zeta_0^0\zeta_0^{0T} + \sum_{i=1}^{\infty} E\zeta_0^0\zeta_i^{0T} + E\zeta_i^0\zeta_0^{0T}. \tag{8}$$

Note that assumptions [A2] and [A3] guarantee the existence of $S_0 < \infty$. The precise sense of the limit above is

$$\lim_{\mu\to 0}\limsup_{t\to\infty} |E[f(\mu^{-1/2}\Delta_t)] - E[f(\mathcal{N}(0, V))]| = 0$$

for any bounded continuous function f.

The proof of Theorem 1 is postponed until §4 below.

Comment. We should note that the assertions of the theorem above provide an infinite horizon approximation. We emphasize the difference with respect to the standard results on weak approximation, when the weak convergence on the finite interval of order γ^{-1} is usually shown. Let us compare Theorem 1 with the analogous results of Theorem 4.15, Part 2 of [1] or the corresponding result in [9]. In fact, those results implicitly require that the trajectories (Δ_t) of equation (3) are bounded. Furthermore, the condition on the dependence in the input sequence (P_t) in Theorem 1 is more explicit and less restrictive than those, for instance, of Theorem 4.15, Part 2 in [1] since only ergodicity is required.

Compared with results obtained in [4], [18], [12]–[14] this paper does not consider moments of $\hat{\theta}_t - \theta_t$ (which may as well be infinite) but investigates the limit in law of this process. The notable advantage of this approach is the weakness of the assumptions made in the theorem. It should be noted that the assertion of the theorem does not imply the convergence of the distribution of $\mu^{-1/2}\Delta_t$ to any limit as $t \to \infty$ for fixed μ. It only means that the distance between the distributions of $\mu^{-1/2}\Delta_t$ and the Gaussian distribution converges to 0 as $\mu \to 0$.

Discussion. Let us consider assumptions of the result above. They are quite simple, and only assumption [A3] needs to be clarified. Note that this condition is usual when proving the CLT for the stationary processes. It is satisfied if, for instance, ζ_t^0 is a martingale-difference process with $E|\zeta_1^0|^2 < \infty$. Let ζ_t^0 be defined by the equation

$$\zeta_t^0 = G\zeta_{t-1}^0 + \xi_t, \quad \zeta_0^0 \in \mathbf{R}^N,$$

where $G \in \mathbf{R}^{N\times N}$. If G is stable (i.e., $|\lambda_{\max}(G)| \le \rho < 1$) and ξ_t is a martingale-difference with $E|\xi|^2 \le \infty$, then $E|E(\zeta_t^0|\mathcal{F}_0)|^2 \le K\rho^t$ and assumption [A3] holds.

We can also translate assumption [A3] into usual mixing conditions on the sequences (ζ_t^0) and (w_t).

Let us recall the definitions of the three mixing coefficients $\phi(k)$, $\alpha(k)$, and $\rho(k)$ (cf. [6]). Consider the sequence (x_t) with $Ex_i = 0$.

1. The uniformly strong mixing coefficient

$$\phi(k) = \sup |P(B|A) - P(B)|, \quad t, k \ge 1, \quad P(A) > 0$$

where the supremum is taken over all sets $A \in \mathcal{M}_t = \sigma\{\ldots, x_{t-1}, x_t\}$ and $B \in \mathcal{M}^{t+k} = \sigma\{x_{t+k}, x_{t+k+1}, \ldots\}$.

2. The strong mixing coefficient

$$\alpha(k) = \sup_{A \in \mathcal{M}_t, B \in \mathcal{M}^{t+k}} |P(AB) - P(A)P(B)|, \quad t, k \geq 1.$$

3. The coefficient of the maximum correlation

$$\rho(k) = \sup \operatorname{cov}(z', z''), \quad t, k \geq 1,$$

where the supremum is taken over all random variables z' and z'' with the bounded second moment such that z' is $\mathcal{M}_t$-measurable and z'' is $\mathcal{M}^{t+k}$-measurable.

LEMMA 1. *The following inequalities hold:*

$$||E(x_{t+k}|\mathcal{M}_t)||_p \leq 2\phi(k)^{1/p} \sup_t ||x_t||_p, \quad p \geq 2;$$

$$||E(x_{t+k}|\mathcal{M}_t)||_p \leq 8\alpha(k)^{1/p-1/q} \sup_t ||x_t||_p, \quad p > 1, \ q \geq p;$$

$$||E(x_{t+k}|\mathcal{M}_t)||_2 \leq \rho(k)^{1/2}.$$

Proof. The proof follows immediately from the classical mixing inequalities (see, for instance, appendix III of [5]). □

2.2. Extension to the Markov case. We use the following propositions to obtain some further generalizations for the the case of Markov processes (P_t) and (ζ_t). By using Proposition 1 below, we can identify the process (P_t, ζ_t) with the stationary sequence $(\bar{P}_t, \bar{\zeta}_t)$ with probability $1 - \epsilon$ for arbitrary $\epsilon > 0$ and $t \geq n(\epsilon)$. The sequence $\bar{\Delta}_t$ obtained from $(\bar{P}_t, \bar{\zeta}_t)_{t \geq n}$ with $\bar{\Delta}_{n(\epsilon)} = \Delta_{n(\epsilon)}$ satisfies (6) and

$$P(\bar{\Delta}_t = \Delta_t, t \geq n(\epsilon)) \geq 1 - \epsilon.$$

The result is now immediate since ϵ is arbitrary (the dependence of $n(\epsilon)$ on μ does not interfere due to the order of the limits in the statement of Theorem 1).

The following proposition shows how trajectories of certain Markov chains may be identified, up to a change of probability space, with those of a stationary Markov chain with same transition probability.

PROPOSITION 1. *Let $\Pi(x, dy)$ be a transition probability on some measurable space S, with invariant measure π, such that for any initial probability measure ν on S,*

$$\lim_{t \to \infty} ||\nu \Pi^t - \pi|| = 0$$

($||\nu||$ denotes the total variation of measure ν). Then for any distribution ν and any $\epsilon > 0$, there exist a (deterministic) time T and two Markov chains X_t and Y_t with transition probability Π, and initial distributions ν and π, respectively, such that

$$P(X_t = Y_t, \ t \geq T) > 1 - \epsilon.$$

Proof. We first recall a useful result [3]: Let X be a Polish space and $M(X)$ be the set of probability measures on X. There exists a Borel function

$$\rho : M(X) \times [0, 1] \longrightarrow X$$

such that if U is a uniform right variable (r.v.) on $[0, 1]$ the distribution of $\rho(m, U)$ is m for any probability measure $m \in M(X)$.

Using this result, we can build a realization of any Markov chain on the probability space $[0,1]^N$ by defining $X_n = \rho(P_n(X_{n-1}, dy), U_n)$ where $P_n(x, dy)$ is the transition probability at time n and U_n is an independently and identically distributed (i.i.d.) sequence of r.v. uniformly distributed on $[0,1]$. This is the mechanism we use in the upcoming sections in order to avoid the construction of trajectory spaces (we will need two functions ρ: one corresponding to $X = S$ and the other to $X = S \times S$).

Consider a realization Z_t of the Markov chain with initial distribution ν. From the assumptions, we can find a time T such that

$$||\nu\Pi^T - \pi|| < 1 - \epsilon,$$

and this implies the existence of a random variable Y_T with distribution π such that

$$P(Z_T = Y_T) > 1 - \epsilon.$$

By running (as explained before) the chain on $S \times S$ with transition probability

$$\Psi(x, x', dy, dy') = I(x = x')\Pi(x, dy)\delta_y(y') + I(x \neq x')\Pi(x, dy)\Pi(x', dy')$$

and starting with (Z_T, Y_T), we obtain a process (X_t, Y_t) defined for $t \geq T$. Then setting $X_t = Z_t$ for $t < T$, we get the desired assertion. □

Consider the following assumption:

[A6] There exist a measure ϕ and a measurable set A such that for any ϕ-positive set B

$$\inf_{x \in A} P_x(X_n \in B \text{ for some } n \leq n_0) > 0 \quad \text{for some } n_0 = n_0(B)$$

and the chain X_n is aperiodic. Furthermore, there exist a measurable non-negative function g and a positive real number ϵ such that

$$\begin{aligned} &\textstyle\int g(y)\Pi(x, dy) \leq g(x) - \epsilon \quad \text{for any } x \in A^c, \\ &\textstyle\sup_{x \in A} \int g(y)\Pi(x, dy) < \infty. \end{aligned} \tag{9}$$

The condition required in the settings of Proposition 1 may be easily checked through the following result (cf. Theorems 4.6 and 5.2 in [16]).

PROPOSITION 2. *Let X_n be a Markov chain with transition probability* $\Pi(x, B)$ ($x \in \mathbf{R}^N$, $B \subset \mathbf{R}^N$). *If assumption* [A6] *holds, then the chain is positive Harris-recurrent* (*ergodic*), *it has a unique invariant probability measure π. In particular,*

$$\lim_{n \to \infty} ||\lambda\Pi^n - \pi|| = 0$$

for any initial distribution λ.

Example. Consider the process (φ_t), $\varphi_t \in \mathbf{R}^N$ defined by the equation

$$\varphi_t = G\varphi_{t-1} + F\xi_t, \quad \varphi_0 \in \mathbf{R}^N,$$

where $G \in \mathbf{R}^{N \times N}$, and (ξ_t) is a sequence of independent and identically distributed random variables with, say, $E\xi_1 = 0$, $E|\xi_1|^2 < \infty$. If the matrix G is stable then by the Lyapunov theorem we conclude that there are matrices $W = W^T > 0$ and $U > 0$ such that $W = G^T W G + U$. Thus we obtain for the function $v_t = \varphi_t^T W \varphi_t$

$$\begin{aligned} E(v_t|\mathcal{F}_{t-1}) &= v_{t-1} - \varphi_t^T U \varphi_t + \xi_t^T F^T W F \xi_t \\ &\leq (1 - \alpha)v_{t-1} + \operatorname{trace}(F^T W F E\xi_t\xi_t^T) \end{aligned} \tag{10}$$

for some $\alpha > 0$. Set $A = \{x \in \mathbf{R}^N : |x| \leq R\}$. From (10) we conclude that there is some $R < \infty$ that the inequalities in (9) hold. Suppose that the distribution of ξ_t has an absolutely continuous on $\mathbf{R}^N$ component with respect to the Lebesgue measure. If the pair of matrices (G, F) is controllable, then it can be easily verified that $\Pi^N(x, dy)$ has a positive continuous component on A, thus assumption [A6] holds. More results may be found in [17, Chap. 6] for the convergence of $\Pi^n(x, \cdot)$ to π and in [19] about the second assertion of the proposition.

When summing up the above arguments we obtain the following result.

THEOREM 2. *Suppose that a Markov chain* (ξ_t) *satisfies assumption* [A6]. *This implies the existence of the invariant probability* $\pi(\cdot)$. *Let* Δ_t *be defined by equation* (3), *where* $P_t = P_t(\xi)$ *and* $\zeta_t = \zeta_t(\xi)$. *Suppose that* P_t *can be decomposed as in* (5). *Furthermore, assumptions* [A1], [A3]–[A5] *hold for the invariant measure* π. *Then* $\lim_{\mu\to 0}\lim_{t\to\infty}\mu^{-1/2}\Delta_t = \mathcal{N}(0, V)$ *in distribution, i.e.,* $\mu^{-1/2}\Delta_t$ *converges in distribution to the Gaussian random variable with zero mean and covariance* V. *Here* V *is a solution of the Lyapunov equation* (7).

2.3. Application guide for Theorem 1. Equation (2) leads to the following recursion on the estimation error $\delta_t = \theta_t - \hat{\theta}_t$:

$$\delta_t = (I - \mu L_t \varphi_t^T)\delta_{t-1} + \mu L_t e_t - \gamma w_t. \tag{11}$$

Set

$$\begin{aligned} \delta_t^{(1)} &= (I - \mu L_t \varphi_t^T)\delta_{t-1}^{(1)} + \sqrt{\mu} L_t e_t, \quad \delta_0^{(1)} = \delta_0, \\ \delta_t^{(2)} &= (I - \mu L_t \varphi_t^T)\delta_{t-1}^{(2)} - \sqrt{\mu} w_t, \quad \delta_0^{(2)} = 0, \\ \Delta_t^T &= \sqrt{\mu}(\delta_t^{(1)}, \delta_t^{(2)})^T. \end{aligned}$$

Then

$$\delta_t = \sqrt{\mu}\delta_t^{(1)} + \frac{\gamma}{\sqrt{\mu}}\delta_t^{(2)}, \tag{12}$$

and equation (3) is equivalent to (11) if we put

$$P_t = \begin{pmatrix} L_t \varphi_t^T & 0 \\ 0 & L_t \varphi_t^T \end{pmatrix}, \tag{13}$$

$$\zeta_t = \begin{pmatrix} L_t e_t \\ -w_t \end{pmatrix}. \tag{14}$$

Thus, as μ tends to zero, δ is supplied by equation (12) where the vector $(\delta^{(1)}, \delta^{(2)})$ is approximated (independently of γ) by a Gaussian vector with known covariance matrix of order one. These arguments are well known when dealing with the moment bounds for δ_t (cf. [15], [13], etc.), but to the best of our knowledge, are somewhat new when considering the distributions. Note that the optimal choice of the gain coefficient is $\mu = \gamma$ (we can attribute the coefficient of order of magnitude one to the gain matrix Γ). Yet, the degenerate case ($\mu = o(\gamma)$ or $\gamma = o(\mu)$) is interesting from the practical point of view, since the scale parameter γ of the disturbance is often unknown and is extremely hard to extract from data. We introduce the following useful notation.

DEFINITION 1. *Let*

$$\delta_t = \sqrt{\mu}\delta_t^{(1)} + \frac{\gamma}{\sqrt{\mu}}\delta_t^{(2)},$$

where $\delta_t^{(1)}$ *and* $\delta_t^{(2)}$ *are two random processes which depend on parameter* μ *only. If* $(\delta_t^{(1)}, \delta_t^{(2)})$ *converges in distribution to the Gaussian random variable* $N(0, V)$, *where*

$$V = \begin{pmatrix} V_1 & 0 \\ 0 & V_2 \end{pmatrix},$$

then we will write $\delta_t \asymp N(0, \mu V_1 + \frac{\gamma^2}{\mu} V_2)$.

In particular

(i) if $\gamma = o(\mu)$ then $\lim_{\mu\to 0}\lim_{t\to\infty} \mu^{-1/2}\delta_t = \mathcal{N}(0, V_1)$ in distribution;

(ii) if $\mu = o(\gamma)$ then $\lim_{\mu\to 0}\lim_{t\to\infty} \frac{\sqrt{\mu}}{\gamma}\delta_t = \mathcal{N}(0, V_2)$ in distribution;

(iii) if $\mu = \gamma$ then $\lim_{\mu\to 0}\lim_{t\to\infty} \gamma^{-1/2}\delta_t = \mathcal{N}(0, V_1 + V_2)$ in distribution.

3. Algorithms study. To simplify the presentation we will require the disturbances e_t and w_t to be martingale-difference processes. This condition is more restrictive than assumption [A3] on ζ_t in (14). Note that in a variety of cases some more general assumptions on the dependence in (e_t), (w_t), and (φ_t) would be sufficient for ζ_t in (14) to verify assumption [A3] of Theorems 1 and 2. We will consider only the stationary case. The implication for the Markov case is then straightforward through Theorem 2 if assumption [A6] holds for the Markov process (φ_t, e_t, w_t).

Denote $\mathcal{F}'_t = \sigma\{\dots, \varphi_{t+1}, e_t, w_t\}$ and by $\mathcal{G}'$ the σ-algebra of invariant sets of the sequence (φ_t, e_t, w_t). Consider the following assumptions:

[B1] (φ_t, e_t, w_t) is a strictly stationary ergodic sequence; $E(e_t|\mathcal{F}'_{t-1}) = 0$, $E(w_t|\mathcal{F}'_{t-1}) = 0$, $Ee_t^4 < \infty$, $E(e^2|\mathcal{F}'_{t-1}) = \sigma_e^2$, $E(w_t w_t^T|\mathcal{F}'_{t-1}) = R_w$, and $E(w_t e_t|\mathcal{F}'_{t-1}) = 0$.

[B2] $E|\varphi_t|^4 < \infty$ and $E\varphi_t\varphi_t^T > 0$.

[B3] $\Gamma = \Gamma^T > 0$.

3.1. Widrow's algorithm (LMS). Let us consider the following algorithm [20]:

$$\hat\theta_t = \hat\theta_{t-1} + \mu\Gamma\varphi_t(y_t - \varphi_t^T\hat\theta_{t-1}). \tag{15}$$

We will also consider a slightly different version of this method:

$$\hat\theta_t = \hat\theta_{t-1} + \mu\Gamma\frac{\varphi_t}{1 + \mu\varphi_t^T\Gamma\varphi_t}(y_t - \varphi_t^T\hat\theta_{t-1}). \tag{16}$$

Let $D = E\varphi_1\varphi_1^T$. Consider the following Lyapunov equations with respect to the matrix V_i:

$$\Gamma D V_1 + V_1 D \Gamma = \sigma_e^2 \Gamma D \Gamma, \tag{17}$$

and

$$\Gamma D V_2 + V_2 D \Gamma = R_w. \tag{18}$$

THEOREM 3. *Let assumptions* [B1]–[B3] *hold true. Consider algorithm* (15) *or* (16) (*in this case we require, in addition,* $E|\varphi_t|^8 < \infty$) *with arbitrary initial condition* $\hat\theta_0$. *Then using the notation of Definition* 1 *we have*

$$\hat{\theta}_t - \theta_t \asymp \mathcal{N}(0, \mu V_1 + V_2\gamma^2/\mu).$$

The correspondent result on L_2 convergence of the Widrow algorithm has been obtained by Macchi and Eweda [15] under condition of m-dependence of regressors φ_t.

Proof. Denote

$$\delta_t = \Gamma^{-1/2}(\hat{\theta}_t - \theta_t),\ \bar{\varphi}_t = \Gamma^{1/2}\varphi_t,\ \bar{w}_t = \Gamma^{-1/2}w_t, \tag{19}$$

$$P_t^0 = \bar{\varphi}_t\bar{\varphi}_t^T, \qquad \zeta_t^0 = \begin{pmatrix} \bar{\varphi}_t e_t \\ \bar{w}_t \end{pmatrix},$$
$$P_t^1 = \frac{\bar{\varphi}_t\bar{\varphi}_t^T}{1+\mu|\bar{\varphi}_t|^2}, \qquad \zeta_t^1 = \begin{pmatrix} \frac{\bar{\varphi}_t e_t}{1+\mu|\bar{\varphi}_t|^2} \\ \bar{w}_t \end{pmatrix}.$$

Using these notations we obtain the transformed error δ_t of algorithms (15) or (16):

$$\delta_t = \sqrt{\mu}\delta_t^{(1)} + \delta_t^{(2)}\gamma/\sqrt{\mu},$$

where the vector $\Delta_t = \sqrt{\mu}(\delta_t^{(1)}, \delta_t^{(2)})$ is defined by the following equation:

$$\Delta_t = (I - \mu P_t)\Delta_{t-1} + \mu\zeta_t \tag{20}$$

with

$$P_t = \begin{pmatrix} P_t^i & 0 \\ 0 & P_t^i \end{pmatrix},$$
$$\zeta_t = \zeta_t^i.$$

We choose i=0 or i=1 for algorithms (15) or (16), respectively. The processes P_t^0, ζ_t^0 verify [A1]–[A5] of Theorem 1, thus the assertion of the theorem is immediate for algorithm (15). Next note that P_t^1 can be decomposed in the following way: $P_t^1 = P_t^0 + P_t'$, where

$$P_t' = -\mu\frac{|\bar{\varphi}_t|^2\bar{\varphi}_t\bar{\varphi}_t^T}{1+\mu|\bar{\varphi}_t|^2}.$$

One can easily verify that assumption [A5] holds true. Furthermore,

$$E|\zeta_t^1 - \zeta_t^0| \leq \mu E(|\bar{\varphi}_t|^3)\sigma_e,$$

which implies the limit in [A4]. Then Theorem 1 implies the assertion of the theorem for algorithm (16). □

To accomplish the asymptotic study of the Widrow algorithm, we develop its optimal version. Let $\mu = \gamma$. Then

$$\gamma^{-1/2}(\hat{\theta}_t - \theta_t) \xrightarrow{D} \mathcal{N}(0, V_3),$$

where $V_3 = V_1 + V_2$ is the solution of the following equation:

$$\Gamma D V_3 + V_3 D\Gamma = \sigma_e^2\Gamma D\Gamma + R_w. \tag{21}$$

We readily get the following corollary.

COROLLARY 1. *Consider a gain matrix* Γ^* *satisfying*

$$\Gamma^* D \Gamma^* = \frac{R_w}{\sigma_e^2}.$$

Then for any choice of Γ *admissible with respect to Theorem* 1

$$V(\Gamma) \geq V(\Gamma^*) = \sigma_e^2 \Gamma^*,$$

where $V(\Gamma)$ *is the solution of equation* (21) *with the fixed matrix* Γ.

The proof of the corollary can be found in [8]. We just remark here that $V(\Gamma^*)$ is given by

$$V(\Gamma^*) = \sigma_e D^{-1/2}(D^{1/2} R_w D^{1/2})^{1/2} D^{-1/2}. \tag{22}$$

Note that if the disturbances e_t and w_t have conditional Gaussian distribution, then this value of the error covariance is the least achievable by any estimation algorithm (see the lower bound in [8]).

3.2. Normalized algorithm (NLMS). Let us consider the following algorithm:

$$\hat{\theta}_t = \hat{\theta}_{t-1} + \mu \Gamma \frac{\varphi_t}{1 + |\varphi_t|^2}(y_t - \varphi_t^T \hat{\theta}_{t-1}). \tag{23}$$

It was introduced in the literature (see, for instance, [13]) as a stable modification of the usual LMS algorithm. The L_p stability of the method has been shown in [13] and some estimates of the rate of convergence have been established.

With a slight abuse of notation we denote $D = E\varphi_t \varphi_t^T (1 + |\varphi_t|^2)^{-1}$ and $G = E\varphi_t \varphi_t^T (1 + |\varphi_t|^2)^{-2}$. Note that if assumption [B2] holds then $D > 0$ and $G > 0$. Consider the following Lyapunov equations with respect to the matrix V_i:

$$\Gamma D V_1 + V_1 D \Gamma = \sigma_e^2 \Gamma G \Gamma,$$
$$\Gamma D V_2 + V_2 D \Gamma = R_w.$$

THEOREM 4. *Let assumptions* [B1]–[B3] *hold true. Consider algorithm* (23) *with arbitrary initial condition* $\hat{\theta}_0$. *Then*

$$\hat{\theta}_t - \theta_t \asymp \mathcal{N}(0, \mu V_1 + V_2 \gamma^2 / \mu).$$

The proof of the theorem is analogous to that of Theorem 3.

3.3. Forgetting factor algorithm (RSL). Consider the following equation for the estimate $\hat{\theta}_t$:

$$\begin{aligned} \hat{\theta}_t &= \hat{\theta}_{t-1} + \mu R_t^{-1} \varphi_t (y_t - \hat{\theta}_{t-1}^T \varphi_t), \\ R_t &= (1 - \mu) R_{t-1} + \mu \varphi_t \varphi_t^T, \quad R_0 = \varrho I \end{aligned} \tag{24}$$

with $\varrho > 0$. Denote $D = E\varphi_t \varphi_t^T$. Define

$$V_1 = \sigma_e^2 D^{-1}/2,$$
$$V_2 = R_w/2$$

THEOREM 5. *Let assumptions* [B1] *and* [B2] *hold true. Consider algorithm* (24) *with arbitrary initial condition* $\hat{\theta}_0$. *Then*

$$\hat{\theta}_t - \theta_t \asymp \mathcal{N}(0, \mu V_1 + V_2\gamma^2/\mu).$$

The sufficient conditions of L_p stability of the method were provided in [4] and [18] along with some estimates of the rate of convergence. We need the following simple lemma.

LEMMA 2. *Suppose that* (φ_t) *is an ergodic stationary process* $\mathbf{R}^N$*-valued,* $E\varphi_t\varphi_t^T = D$, *and* $E|\varphi_t|^4 < \infty$. *Let* R_t *be defined by the equation*

$$R_t = (1-\mu)R_{t-1} + \mu\varphi_t\varphi_t^T, \quad R_0 = \varrho I > 0.$$

Then $\sup_{0<\mu<1}\sup_t E|R_t - D|^2 < \infty$, *and* $\lim_{\mu\to 0}\lim_{t\to\infty} E|R_t - D|^2 = 0$.

Proof. Denote $\Delta_t = R_t - D$. Then we have for Δ_t the following equation:

$$\Delta_t = (1-\mu)\Delta_{t-1} + \mu(\varphi_t\varphi_t^T - D). \tag{25}$$

Note that for any $x, y \in \mathbf{R}$ $2|xy| \le \alpha x^2 + \alpha^{-1}y^2$. If we take $\alpha = 1-\mu$ then we obtain from (25)

$$\begin{aligned} E|\Delta_t|^2 &\le (1-\mu)^2(1+\mu)E|\Delta_{t-1}|^2 + \mu^2(1+\mu^{-1})E|\varphi_t\varphi_t^T - D|^2 \\ &\le (1-\mu)E|\Delta_{t-1}|^2 + K\mu(E|\varphi|^4 + 1), \end{aligned}$$

which implies the first assertion of the lemma.

Let us fix some integer $l > 0$. For $t \ge l$ we have

$$\begin{aligned} |\Delta_t|^2 &= \mathrm{trace}\Delta_t\Delta_t^T \\ &= (1-\mu)^2|\Delta_{t-1}|^2 + \mu^2|\varphi_t\varphi_t^T - D|^2 + 2\mu(1-\mu)\mathrm{trace}(\Delta_{t-1}(\varphi_t\varphi_t^T - D)) \\ &\le (1-\mu)^2|\Delta_{t-1}|^2 + 2\mu^2(|\varphi_t|^4 + |D|^2) \\ &\quad +2\mu(1-\mu)\mathrm{trace}(\Delta_{t-l}(\varphi_t\varphi_t^T - D)) + 2\mu(1-\mu)|\Delta_t - \Delta_{t-l}||\varphi_t\varphi_t^T - D|. \end{aligned} \tag{26}$$

It follows from the stationarity and ergodicity of (φ_t) (and the Levy theorem) that there is $\epsilon(t)$ ($\epsilon(t) \to 0$ as $t \to \infty$) such that $E|E(\varphi_t\varphi_t^T - D|\mathcal{F}_0)|^2 \le \epsilon(t)$. Here, as usual, $\mathcal{F}_t = \sigma\{\dots, \varphi_t\}$. Therefore,

$$|E\Delta_{t-l}(\varphi_t\varphi_t^T - D)|^2 \le E|\Delta_{t-l}|^2 E|E(\varphi_t\varphi_t^T - D|\mathcal{F}_{t-l})|^2 \le K\epsilon(l).$$

On the other hand, we easily derive that $E|\Delta_t - \Delta_{t-l}|^2 \le Kl\mu$. Summing up, we have for $t \ge l$ from (26)

$$E|\Delta_t|^2 \le (1-\mu)^2 E|\Delta_{t-1}|^2 + K\mu(\epsilon(l)^{1/2} + (l+1)\mu).$$

Choosing $l = \mu^{-1/2}$, we get for any $t \ge l$

$$E|\Delta_t|^2 \le K(1-\mu)^t + o(1),$$

which implies the second assertion of the lemma. □

Proof of Theorem 5. For the sake of conciseness we consider only the case $\mu = \gamma$. Denote $x_t = R_t(\hat{\theta}_t - \theta_t)$. Then we obtain for x_t the following equation (cf. [2]):

$$x_t = (1-\gamma)x_{t-1} + \gamma\varphi_t e_t + \gamma R_t w_t.$$

This equation can be decomposed in the following way:

$$x_t = \bar{x}_t + \bar{\delta}_t,$$

where

$$\bar{\delta}_t = (1-\gamma)\bar{\delta}_{t-1} + \gamma(R_t - D)w_t,\ \bar{\delta}_0 = 0,$$
$$\bar{x}_t = (1-\gamma)\bar{x}_{t-1} + \gamma\varphi_t e_t + \gamma D w_t,\ \bar{x}_0 = x_0.$$

It follows from Lemma 2 that all conditions of Theorem 1 are satisfied for the process $(\bar{x}_t)$, thus

$$\lim_{\gamma\to 0}\lim_{t\to\infty} \gamma^{-1/2}\bar{x}_t = \mathcal{N}(0,\bar{V}) \text{ in distribution},$$

where the covariance matrix $\bar{V} = \sigma_e^2 D + DR_w D$. Since $(R_t - D)w_t$ is a martingale-difference, we get for $\bar{\delta}_t$

$$E|\bar{\delta}_t|^2 \leq (1-\gamma)^2 E|\bar{\delta}_{t-1}|^2 + \gamma^2|R_w|E|R_t - D|^2 = o(\gamma).$$

Thus $\lim_{\gamma\to 0}\lim_{t\to\infty}\gamma^{-1/2}|\bar{\delta}_t| = 0$. From Lemma 2 we conclude that for any $0 < \varepsilon \leq 1/2$ there is $\mu(\varepsilon)$ such that $P(|R_t - D| > \varepsilon) < \varepsilon$ as soon as $\mu \leq \mu(\varepsilon)$. Thus $P(|R_t^{-1} - D^{-1}| > 2\lambda_{\min}^{-1}(D)\varepsilon) < \varepsilon$. This implies that $\lim_{\gamma\to 0}\lim_{t\to\infty}\gamma^{-1/2}(R_t^{-1}x_t - D^{-1}x_t) = 0$ and $\gamma^{-1/2}(\hat{\theta}_t - \theta_t) \sim D^{-1}x_t$. □

Note that if $\mu = \gamma$ then the normalized error covariance matrix $V = (\sigma_e^2 D^{-1} + R_w)/2$ differs from the optimal value in (22). This reflects the fact that the forgetting factor method usually used is not optimal.

3.4. Stabilized LS. Consider the following equations:

$$\hat{\theta}_t = \hat{\theta}_{t-1} + \mu\Gamma_t\varphi_t(y_t - \varphi_t^T\hat{\theta}_{t-1}), \tag{27}$$

$$\Gamma_t = (\mu\rho^{-1}R_t + \Gamma^{-1})^{-1},\ \Gamma = \Gamma^T > 0, \tag{28}$$

$$R_t = (1-\rho)R_{t-1} + \rho\varphi_t\varphi_t^T, \quad R_0 = 0. \tag{29}$$

Here $0 < \rho < 1$ and $\mu > 0$ are the algorithm parameters. As $\mu \to 0$ equations (27)–(29) can be seen as a regularized version of the LMS algorithm (15) with the constant gain Γ, as well as a regularized version of the forgetting factor method. The amazing property of this method is its L_p-stability for any choice of parameters μ and ρ under some excitation condition (cf. [7]).

THEOREM 6. *Let assumptions* [B1]–[B3] *hold true. Consider algorithm* (27)–(29) *with arbitrary initial condition* $\hat{\theta}_0$. *Then*

$$\hat{\theta}_t - \theta_t \asymp \mathcal{N}(0, \mu V_1 + V_2\gamma^2/\mu)$$

where the matrices V_1 *and* V_2 *are defined in* (17) *and* (18).

Proof. If we set $P_t^0 = \Gamma$, then $\Gamma_t = P_t^0 + P_t'$, where

$$P_t' = -\Gamma(\Gamma + \mu^{-1}\rho R_t^{-1})^{-1}\Gamma,$$

and

$$|P_t'| \leq |\Gamma|^2|(\Gamma + \mu^{-1}\rho R_t^{-1})^{-1}| \leq |\Gamma|^2\mu\rho^{-1}|R_t|.$$

Let us introduce the "stationary version" of the sequence (R_t):

$$S_0 = \rho \sum_{j=-\infty}^{0} (1-\rho)^{-j} \varphi_j \varphi_j^T,$$
$$S_t = (1-\rho) S_{t-1} + \rho \varphi_t \varphi_t^T.$$

Note that

$$S_t - R_t = (1-\rho)(S_{t-1} - R_{t-1})$$

and $R_0 = 0$ implies $R_t \leq S_t$; thus the choice

$$p_t' = |\Gamma|^2 \mu \rho^{-1} |S_t|$$

satisfies assumption [A5]. The rest of the proof can be carried out along the same lines as the proof of Theorem 3. □

Since the expression for the asymptotic error covariance is the same as in Theorem 3, Corollary 1 remains valid for algorithm (27)–(29).

4. Proof of Theorem 1. The proof of the theorem consists of the following parts. First we show that the matrix product $\prod_{i=k}^{t}(I - \mu P_i)$ is exponentially stable as $t \to \infty$ (Lemma 4). Using this fact, we prove that the process (Δ_t) is asymptotically equivalent to $(\bar{\Delta}_t)$ defined by equation (3) with ζ_t substituted with some martingale-difference process ξ_t (Proposition 3). Next we show that the distribution of $\mu^{-1/2}\bar{\Delta}_t$ is tight (Proposition 4). Next we provide the linear approximation Δ_t' of Δ_t (Lemma 6), which makes it possible to accomplish the proof by using the classical CLT for martingales (Propositions 5 and 6).

LEMMA 3. *For any sequence of symmetric matrices $(Q_i)_{i=1,\dots,n}$, the following inequality holds*:

$$\log\left(\left|\prod_{i=1}^{n}(I - Q_i)\right|\right) \leq -\min\left(\lambda_{\min}\left(\sum_{i=1}^{n} Q_i\right), \frac{1}{2}\right) + \frac{n}{2}\log\left(1 + K(n)\max_i |Q_i|^2\right),$$

where $\lambda_{\min}(P)$ stands for the smallest eigenvalue of P and $K(n)$ is a constant depending only on n.

Proof. For any unit vector X, one has

$$X^T\left(\prod_{i=1}^{n}(I - Q_i)\right)^T \prod_{i=1}^{n}(I - Q_i)X = 1 - 2X^T\left(\sum_{i=1}^{n} Q_i\right)X + X^T \sum_l \Pi_l X,$$

where each Π_l is a product of matrices Q_i with more than 2 and less than $2n$ terms. Putting $q = \max|Q_i|$ and $\lambda = \lambda_{\min}(\sum Q_i)$, we then have

$$\begin{aligned}
\left|\prod_{i=1}^{n}(I - Q_i)\right|^2 &\leq 1 - 2\lambda + K(n)(q^2 + q^{2n}) \\
&\leq 1 - 2\min(\lambda, 1/2) + K(n)(q^2 + q^{2n}) \\
&\leq (1 - \min(\lambda, 1/2))^2 + K(n)(q^2 + q^{2n}) \\
&\leq (1 - \min(\lambda, 1/2))^2(1 + 4K(n)(q^2 + q^{2n})) \\
&\leq (1 - \min(\lambda, 1/2))^2(1 + K'(n)q^2)^n,
\end{aligned}$$

and finally

$$\log \left| \prod_{i=1}^{n} (I - Q_i) \right| \leq -\min(\lambda, 1/2) + \frac{n}{2} \log(1 + K'(n)q^2). \qquad \square$$

Denote

$$\pi_{j,t} = \prod_{i=j}^{t} (I - \mu P_i).$$

LEMMA 4. *Let conditions* [A1], [A2], *and* [A5] *be satisfied, then there exist* μ_0, α_0, *and a stationary random process* $C_s(\omega)$ *such that for any* T

$$\sup_{t>T, \mu\leq\mu_0} (1 - \alpha_0\mu)^{T-t} \pi_{T,t} < C_T(\omega) < \infty \ \ a.s.,$$
$$\sup_{t<T, \mu\leq\mu_0} (1 - \alpha_0\mu)^{t-T} \pi_{t,T} < C_T(\omega) < \infty \ \ a.s..$$

Proof. Note that the second assertion is a consequence of the first used backwards in time (we define the constant $C_T(\omega)$ as the maximum of the backward-forward constants). Let us prove the first one. Consider the case $T = 1$. Note that we have the estimate

$$\log(1 - \alpha\mu)^{1-t} \left| \prod_{i=1}^{t} (I - \mu P_i) \right| \leq t\alpha\mu + \sum_{k=0}^{t/n} \log \left| \prod_{i=kn+1}^{(k+1)n} (I - \mu P_i) \right|.$$

On the other hand, by Lemma 3 we get

$$\begin{aligned}
&(1/\mu) \log \left(\left| \prod_{i=s+1}^{s+n} (I - \mu P_i) \right| \right) \\
&\leq -(1/\mu) \min \left(\lambda_{\min} \left(\mu \sum_{i=s+1}^{s+n} P_i \right), 1/2 \right) + (n/2\mu) \log \left(1 + K(n)\mu^2 \sum_{i=s+1}^{s+n} |P_i|^2 \right) \\
&\leq -\min \left(\lambda_{\min} \left(\sum_{i=s+1}^{s+n} P_i \right), 1/(2\mu) \right) + nK\mu/2 \sum_{i=s+1}^{s+n} |P_i|^2 \\
&\leq -\min \left(\min_{\mu<\mu_0} \lambda_{\min} \left(\sum_{i=s+1}^{s+n} P_i \right), 1/(2\mu_0) \right) + nK\mu_0 \max_{\mu<\mu_0} \left(\sum_{i=s+1}^{s+n} |P_i|^2 \right) /2.
\end{aligned}$$

The bound

$$\begin{aligned}
\lambda_{\min} \left(\sum P_i \right) &\geq \lambda_{\min} \left(\sum P_i^0 + P_i' \right) \\
&\geq \lambda_{\min} \left(\sum P_i^0 \right) - \sum p_i'
\end{aligned}$$
$$\min \left(\lambda_{\min} \left(\sum P_i \right), 1/(2\mu_0) \right) \geq \min \left(\lambda_{\min} \left(\sum P_i^0 \right), 1/(2\mu_0) \right) - \sum p_i'$$

leads to

$$\frac{1}{\mu} \log \left(\left| \prod_{i=s+1}^{s+n} (I - \mu P_i) \right| \right)$$

$$\begin{aligned}&\leq -\min\left(\lambda_{\min}\left(\sum_{i=s+1}^{s+n} P_i^0\right), 1/(2\mu_0)\right) + \sum_{i=s+1}^{s+n} p_i' + (n/2)K\mu_0 \max_{\mu<\mu_0}\left(\sum_{i=s+1}^{s+n} |P_i|^2\right)/2\\ &\leq -\min\left(\lambda_{\min}\left(\sum_{i=s+1}^{s+n} P_i^0\right), 1/(2\mu_0)\right) + \sum_{i=s+1}^{s+n} p_i' + nK\mu_0 \max_{\mu<\mu_0}\left(\sum_{i=s+1}^{s+n} (|P_i^0|^2 + [p_i']^2)\right)\\ &\triangleq K_s(\omega)\end{aligned}$$

where K_s is a sum of stationary ergodic processes. If n has been chosen large enough and μ_0 small enough, then $E[K_s] < 0$. Indeed,

$$\begin{aligned}E[K_s] \leq -E\left[\min\left(\lambda_{\min}\left(\sum P_i^0\right), 1/(2\mu_0)\right)\right] + nE\left[\max_{\mu<\mu_0} p_i'\right]\\ + K\mu_0 n^2\left(E|P_1^0|^2 + E\left[\max_{\mu<\mu_0} |p_1'|\right]^2\right),\end{aligned}$$

where the right-hand side is negative for μ_0 small enough due to the uniform (in μ) bound of p_i' in assumption [A5]. Finally we have

$$\begin{aligned}\log(1-\alpha\mu)^{1-t}\left|\prod_{i=1}^{t}(I-\mu P_i)\right| &\leq t\alpha\mu + \sum_{k=0}^{t/n}\mu K_{kn}\\ &\leq \mu_0 \max_t\left(t\left(\alpha + 1/t\sum_{k=0}^{t/n} K_{kn}\right)\right) \triangleq C_1(\omega).\end{aligned}$$

By the ergodic theorem we conclude that as α is small enough the expression inside the max tends to $-\infty$ as t tends to ∞, which implies the lemma in the case $T = 1$. A shift of indices immediately gives the result for other values of T. □

Let us now make the reduction.

PROPOSITION 3. *The ζ_t^0 process may be written as*

$$\zeta_t^0 = \xi_t - \nu_t + \nu_{t-1} \tag{30}$$

where ξ_t is a stationary martingale-difference process with $E\xi_0\xi_0^T = S_0$, and the process $\bar{\Delta}_t$ defined by

$$\bar{\Delta}_t = (I - \mu P_t)\bar{\Delta}_{t-1} + \mu\xi_t\ , \quad \bar{\Delta}_0 = \Delta_0 \tag{31}$$

satisfies

$$\lim_{\mu\to 0}\lim_{t\to\infty} \mu^{-1/2}(\Delta_t - \bar{\Delta}_t) = 0 \quad \textit{in probability.} \tag{32}$$

This limit is to be taken in the sense explained after Theorem 1.

Proof. Let $C = \sum_{s=0}^{\infty}(E|E(\zeta_s^0|\mathcal{F}_0)|^2)^{1/2}$. If

$$\xi_t = \sum_{s\geq 0} E(\zeta_{t+s}^0|\mathcal{F}_t) - E(\zeta_{t+s}^0|\mathcal{F}_{t-1}), \quad \nu_t = \sum_{s\geq 1} E(\zeta_{t+s}^0|\mathcal{F}_t),$$

then $\zeta_t^0 = \xi_t - \nu_t + \nu_{t-1}$, ξ_t is a martingale-difference and

$$E|\xi_t|^2 \leq 4C^2, \quad E|\nu_t|^2 \leq C^2$$

(this well-known result (see, for example, [5]) is an immediate consequence of the Minkowski inequality), and one easily checks that

$$E\zeta_0^0\zeta_0^{0T} + \sum_{i=1}^{\infty} E\zeta_0^0\zeta_i^{0T} + E\zeta_i^0\zeta_0^{0T} = E\xi_0\xi_0^T.$$

We consider the process $\Delta_t' = \Delta_t - \bar{\Delta}_t$ as follows:

$$\begin{aligned}
\Delta_t' &= (I - \mu P_t)\Delta_{t-1}' + \mu(-\nu_t + \nu_{t-1} + \zeta_j') \\
&= \sum_{j=1}^{t} \pi_{j+1,t}\mu(-\nu_j + \nu_{j-1} + \zeta_j') \qquad (\pi_{t+1,t} = I) \\
&= \mu \sum_{j=2}^{t} (\pi_{j+1,t} - \pi_{j,t})\nu_{j-1} + \pi_{2,t}\mu\nu_0 - \mu\nu_t + \mu \sum_{j=1}^{t} \pi_{j+1,t}\zeta_j' \\
&= \mu \sum_{j=2}^{t} (\mu P_j)\pi_{j+1,t}\nu_{j-1} + \pi_{2,t}\mu\nu_0 - \mu\nu_t + \mu \sum_{j=1}^{t} \pi_{j+1,t}\zeta_j' \\
|\Delta_t'| &\le C_t\mu \sum_{j=1}^{t+1} (1 - \alpha_0\mu)^{t-j-1}(\mu(1 + |P_j|)|\nu_{j-1}| + |\zeta_j'|) \\
&= \mu^{1/2} C_t(\omega) X_t(\mu)
\end{aligned}$$

with $E[X_t(\mu)] < \eta(\mu)$, $\eta(\mu) \to 0$ as $\mu \to 0$. Finally, for any $\epsilon > 0$,

$$\begin{aligned}
P(\mu^{-1/2}|\Delta_t'| > \epsilon) &= P(C_t(\omega)X_t(\mu) > \epsilon) \\
&= P(C_t(\omega) > A) + P(X_t(\mu) > \epsilon/A) \quad \text{for any } A \\
&= P(C_t(\omega) > \eta(\mu)^{-1/2}) + P(X_t(\mu) > \epsilon\eta(\mu)^{1/2}) \\
&= P(C_1(\omega) > \eta(\mu)^{-1/2}) + \eta(\mu)^{1/2}/\epsilon.
\end{aligned}$$

This last quantity is now independent of t and tends to 0 as $\mu \to 0$. □

It remains to prove that

$$\lim_{\mu \to 0} \lim_{t \to \infty} \mu^{-1/2}\bar{\Delta}_t = \mathcal{N}(0, V) \quad \text{in distribution.}$$

We recall that the process $\bar{\Delta}_t$ is defined by

$$\bar{\Delta}_t = (I - \mu P_t)\bar{\Delta}_{t-1} + \mu\xi_t, \quad \bar{\Delta}_0 = \Delta_0,$$

and ξ_t is a stationary martingale-difference process independent of μ with $E\xi_0\xi_0^T = S_0$.

PROPOSITION 4. *The process* $(\mu^{-1/2}\bar{\Delta}_t)$ *is tight in the sense that for some* $\mu' > 0$ *the following is true: for any* $\epsilon > 0$ *there exists* R *such that*

$$P(\mu^{-1/2}|\bar{\Delta}_t| > R(1 + |\bar{\Delta}_0|) < \epsilon$$

for all $\mu \le \mu'$ *and* $t \ge (\alpha_0\mu)^{-1}\ln\mu^{-1}$ (*here* α_0 *was defined in Lemma* 4).

Proof. From equation (3) we get

$$\bar{\Delta}_t = \pi_{1,t}\bar{\Delta}_0 + \sum_{j=1}^{t} \pi_{j+1,t}\mu\xi_j = I_t^{(1)} + I_t^{(2)}$$

(we set $\pi_{t+1,t} = I$). From Lemma 4 we derive that

$$|\pi_{1,t}\bar{\Delta}_0| \le C_1(\omega)(1-\alpha_0\mu)^t|\bar{\Delta}_0| \le C_1(\omega)e^{-\alpha_0\mu t}|\bar{\Delta}_0|$$

with $C_1(\omega) < \infty$ almost surely. This implies that for any $\epsilon > 0$ there is R such that

$$P(C_1(\omega)e^{-\alpha_0\mu t}|\bar{\Delta}_0| > R\mu^{1/2}) \le \epsilon$$

as soon as $t > (\alpha_0\mu)^{-1}\ln\mu^{-1}$; thus $\mu^{-1/2}I_t^{(1)}$ is tight. Let us take $\alpha = \alpha_0/2$. Summing by parts, we have

$$\begin{aligned} I_t^{(2)} &= \sum_{j=1}^{t} \pi_{j+1,t}(1-\alpha\mu)^{-(t-j)}\mu\left((1-\alpha\mu)^{t-j}\xi_j\right) \\ &= \sum_{j=1}^{t-1} \mu(P_{j+1}-\alpha I)\pi_{j+2,t}(1-\alpha\mu)^{-(t-j)}\mu \sum_{i=j+1}^{t}(1-\alpha\mu)^{t-i}\xi_i \\ &\quad + \mu\pi_{2,t}(1-\alpha\mu)^{1-t}\sum_{j=1}^{t}(1-\alpha\mu)^{t-j}\xi_j = r_t^{(1)} + r_t^{(2)}. \end{aligned}$$

The preceding lemma implies that for $\alpha = \alpha_0/2$ the random process

$$\max_{s>0}\{|\pi_{t-s,t}|(1-\alpha\mu)^{-2s}\} \tag{33}$$

may be bounded by a stationary process $C_t'(\omega)$ a.s. finite. Note that

$$E\left|\mu\sum_{i=1}^{t}(1-\alpha\mu)^{t-i}\xi_i\right| \le \mu E\left[\left|\sum_{i=1}^{t}(1-\alpha\mu)^{t-i}\xi_i\right|^2\right]^{1/2} \le K\mu^{1/2},$$

thus

$$|r_t^{(2)}| \le C_t'(\omega)(1-\alpha\mu)^t\mu^{1/2}C_1(\omega)$$

with $EC_1(\omega) < \infty$; $\mu^{-1/2}r_t^{(2)}$ is tight. $r_t^{(1)}$ satisfies the bound

$$|r_t^{(1)}| \le C_t'(\omega)\mu^{3/2}\sum_{j=1}^{t-1}(1-\alpha\mu)^{t-j}C_{j,t}(\omega),$$

where

$$C_{j,t}(\omega) = (|P_{j+1}| + \alpha)\mu^{1/2}\left|\sum_{i=j+1}^{t}(1-\alpha\mu)^{t-i}\xi_i\right|$$

and

$$EC_{j,t}(\omega) \le 4E(|P_{j+1}|^2 + \alpha^2) + 2\mu E\left|\sum_{i=j+1}^{t}(1-\alpha\mu)^{t-i}\xi_i\right|^2 \le K.$$

Hence, we get for $r_t^{(1)}$

$$|r_t^{(1)}| \le \alpha^{-1}\mu^{1/2}C_t'(\omega)C_t''(\omega)$$

where C_t' is a finite stationary process independent of μ, and $E(C_t'')$ is bounded independently of μ. Therefore $\mu^{-1/2}r_t^{(1)}$ is tight. □

LEMMA 5. *There is $\mu' > 0$ and K such that for any $\mu \le \mu'$ and $n \ge 1$*

$$E\left(\sup_{t\le n}\ln(|\bar{\Delta}_t|^2+1)1_{|\bar{\Delta}_0|\le R}\right) \le \ln(R^2+1) + K\mu(n^{1/2}+\mu n).$$

Proof. From equation (3) we get for the probe function $U_t = |\bar{\Delta}_t|^2$

$$\begin{aligned} U_t &\le \bar{\Delta}_{t-1}^T(I-\mu P_t)^T(I-\mu P_t)\bar{\Delta}_{t-1} + 2\mu\bar{\Delta}_{t-1}^T(I-\mu P_t)^T\xi_t + \mu^2|\xi_t|^2 \\ &\le U_{t-1} + 2\mu\xi_t^T\bar{\Delta}_{t-1} + K\mu^2(|\bar{\Delta}_{t-1}|^2|P_t|^2 + |\xi_t|^2). \end{aligned}$$

Denote $l_t = \ln(U_t+1)$. Due to the concavity of the function $\ln(x)$ we get

$$\begin{aligned} l_t &\le l_{t-1} + (U_{t-1}+1)^{-1}\left(2\mu\xi_t^T\bar{\Delta}_{t-1} + K\mu^2(U_{t-1}|P_t|^2 + |\xi_t|^2)\right) \\ &\le l_0 + 2\mu\left|\sum_{i=1}^t \frac{\xi_i^T\bar{\Delta}_{i-1}}{|\bar{\Delta}_{i-1}|^2+1}\right| + K\mu^2\sum_{i=1}^t(|P_i|^2+|\xi_i|^2) \\ &= l_0 + I_t^{(1)} + I_t^{(2)}. \end{aligned} \tag{34}$$

We conclude from the Doob inequality that

$$E\left(\sup_{t\le n} I_t^{(1)}\right) \le \mu C K n^{1/2}$$

and

$$E\left(\sup_{t\le n} I_t^{(2)}\right) = K\mu^2\sum_{i=1}^n(E|P_i|^2 + E|\xi_i|^2) \le K\mu^2 n.$$

Summing up we obtain the desired lemma. □

Denote $L_t = P_t^0 - B$. Let us define for any $t \ge 0$ and $R_0 < \infty$ a random time

$$\sigma'(s) = \inf\{t \ge s : |\bar{\Delta}_t| > \mu^{1/2}R_0\}. \tag{35}$$

We denote $\sigma'(0) = \sigma$. Set

$$m(\varepsilon) = \min\{l \ge 0 : |(I-\mu B)^l| \le \varepsilon\}. \tag{36}$$

Assumption [A1] implies that there exist μ' and $K(\varepsilon) < \infty$ such that

$$m(\varepsilon) \le K(\varepsilon)\mu^{-1} \tag{37}$$

for any $\mu \le \mu'$. Let us consider a new process (Δ_t') along with $(\bar{\Delta}_t)$. We define it by the equation

$$\begin{aligned} \Delta_t' &= (I-\mu B)\Delta_{t-1}' + \mu\xi_t, \quad t \ge 1, \\ \Delta_0' &= \bar{\Delta}_0. \end{aligned} \tag{38}$$

LEMMA 6. *For any $\delta > 0$ there exists $\mu_0(\varepsilon, \delta)$ such that*

$$P(\mu^{-1/2}|\Delta'_{m(\varepsilon)} - \bar{\Delta}_{m(\varepsilon)}| > \delta) < \delta + P\{\sigma < m(\varepsilon)\}$$

for any $\mu \leq \mu_0(\varepsilon, \delta, R_0)$.

Proof. We have the following equation for the difference:

$$\delta'_t = \bar{\Delta}_t - \Delta'_t = (I - \mu B)\delta'_{t-1} - \mu L_t \bar{\Delta}_{t-1} - \mu P'_t \bar{\Delta}_{t-1}, \quad \delta'_0 = 0.$$

Denote $v = \min(\sigma, m(\varepsilon))$. We have for δ'_v the decomposition

$$\begin{aligned} \delta'_v &= -\mu \sum_{i=1}^{v} (I - \mu B)^{v-i}(L_i \bar{\Delta}_{i-1} + \mu P'_i \bar{\Delta}_{i-1}) \\ &= -\mu \sum_{i=l}^{v} (I - \mu B)^{v-i} E(L_i|\mathcal{F}_{i-l})\bar{\Delta}_{i-l} \\ &\quad -\mu \sum_{i=l}^{v} (I - \mu B)^{v-i}(L_i - E(L_i|\mathcal{F}_{i-l}))\bar{\Delta}_{i-l} \\ &\quad -\mu \sum_{i=l}^{v} (I - \mu B)^{v-i} L_i(\bar{\Delta}_{i-1} - \bar{\Delta}_{i-l}) \\ &\quad -\mu \sum_{i=1}^{l-1} (I - \mu B)^{v-i} L_i \bar{\Delta}_{i-1} - \mu \sum_{i=1}^{v} (I - \mu B)^{v-i} P'_i \bar{\Delta}_{i-1} = \sum_{j=1}^{5} I_v^{(j)}. \end{aligned} \tag{39}$$

Let us estimate the first term in (39):

$$\begin{aligned} E|I_v^{(1)}| &\leq K\mu \sum_{i=1}^{m(\varepsilon)} E|E(L_i|\mathcal{F}_{i-l})|\mu^{1/2}R_0 \leq K\mu m(\varepsilon) E|E(L_l|\mathcal{F}_0)|\mu^{1/2}R_0 \\ &\leq K(\epsilon)\epsilon(l) R_0 \mu^{1/2}, \end{aligned}$$

where $K(\epsilon) < \infty$ is defined in (37) and $\epsilon(l)$ is a sequence which tends to 0. From assumption [A5] we derive the estimate for $I_v^{(5)}$:

$$E|I_v^{(5)}| \leq K\mu \sum_{i=1}^{m(\varepsilon)} E|P'_t|\mu^{1/2}R_0 \leq K(\epsilon) R_0 E|p'_1|\mu^{1/2} = \epsilon'(\mu)\mu^{1/2}$$

with $\epsilon'(\mu) \to 0$ as $\mu \to 0$. Next we have for $I_v^{(4)}$

$$\begin{aligned} E|I_v^{(4)}|^2 &\leq K(l+1)\mu^2 E|L_i|^2 \mu R_0^2 \\ &\leq K(l+1)\mu^2 R_0^2. \end{aligned}$$

Note that

$$\begin{aligned} &\bar{\Delta}_{i-1} 1_{i-1<\sigma} - \bar{\Delta}_{i-l} 1_{i-l<\sigma} \\ &= \sum_{k=1}^{l-1} (\bar{\Delta}_{i-k} - \bar{\Delta}_{i-k-1}) 1_{i-k-1<\sigma} - \sum_{k=1}^{l-1} \bar{\Delta}_{i-k} 1_{i-k=\sigma}. \end{aligned}$$

Then we get for $|I_v^{(3)}|$

$$|I_v^{(3)}| \le \mu \sum_{k=1}^{l-1} \sum_{i=1}^{m(\varepsilon)} |L_i||\bar{\Delta}_{i-k} - \bar{\Delta}_{i-k-1}| 1_{i-k-1<\sigma} + \mu \sum_{k=1}^{l-1} \sum_{i=1}^{m(\varepsilon)} |\bar{\Delta}_{i-k}||L_i| 1_{i-k=\sigma}$$

$$\le \mu \sum_{k=1}^{l-1} \sum_{i=1}^{m(\varepsilon)} |L_i||\bar{\Delta}_{i-k} - \bar{\Delta}_{i-k-1}| 1_{i-k-1<\sigma} + \mu|\bar{\Delta}_\sigma| \sum_{k=1}^{l-1} |L_{\sigma+k}| 1_{\sigma+k\le m(\varepsilon)}. \tag{40}$$

Note that

$$E\left(\sup_{i\le m(\varepsilon)} \mu|P_i|\right)^2 \le \mu^2 \sum_{i=1}^{m(\varepsilon)} E|P_i|^2 \le K\,K(\varepsilon)\mu,$$

and $E(\sup_{i\le m(\varepsilon)} \mu|\xi_i|)^2 \le K\,K(\varepsilon)\mu$. Thus

$$(E|\bar{\Delta}_\sigma|^2)^{1/2} \le R_0\mu^{1/2} + K\,K(\varepsilon)\mu^{1/2}(1+\mu^{1/2}).$$

Furthermore,

$$(E|\bar{\Delta}_i - \bar{\Delta}_{i-1}|^2 1_{i-1<\sigma})^{1/2} \le K\mu.$$

Summing up, we obtain from (40)

$$E|I_v^{(3)}| \le \mu K\,K(\varepsilon)(l-1).$$

For $I_v^{(2)}$ we obtain in the same way

$$E|I_v^{(2)}|^2 \le Kl\mu R_0^2.$$

If we take $l = \mu^{-1/4}$, then

$$E|\delta'_v| \le K(\mu^{1/2}(\epsilon(l) + \epsilon'(\mu)) + \mu l) = o(\mu^{1/2})$$

for $\mu \le \mu_0$. Meanwhile, for any $\delta > 0$ we have

$$\begin{aligned}
&P(|\mu^{-1/2}|\delta'_{m(\varepsilon)}| > \delta) \\
&= P(\{|\mu^{-1/2}|\delta'_{m(\varepsilon)}| > \delta\} \cap \{\sigma \ge m(\varepsilon)\}) + P(\{|\mu^{-1/2}|\delta'_{m(\varepsilon)}| > \delta\} \cap \{\sigma < m(\varepsilon)\}) \\
&\le P(\{|\mu^{-1/2}|\delta'_v| > \delta\} \cap \{\sigma \ge m(\varepsilon)\}) + P(\sigma < m(\varepsilon)) \\
&\le \delta^{-1}o(1) + P(\sigma < m(\varepsilon)) \le \delta + P(\sigma < m(\varepsilon))
\end{aligned}$$

for suficiently small μ. □

We must show that $\mu^{-1/2}\Delta'_{m(\varepsilon)}$ converges in distribution to the Gaussian random variable as $\mu \to 0$.

LEMMA 7.

$$\mu \sum_{i=1}^{m(\varepsilon)} (I-\mu B)^{m(\varepsilon)-i} \xi_i \xi_i^T (I-\mu B)^{m(\varepsilon)-i} \xrightarrow{P} V \tag{41}$$

as $\mu \to 0$. Here $V = V^T > 0$ is a unique solution of the Lyapunov equation (7).

Proof. Let us fix μ small enough. Denote $g_i = \xi_i \xi_i^T - S_0$ and

$$\epsilon(i) = \frac{1}{i} \sum_{j=1}^{i} g_{-j}.$$

The ergodicity of the sequence (ξ_i) implies that $\epsilon(i) \to 0$ a.s. The distribution of the sum in the left side of (41) coincides with that of $I_{m(\varepsilon)} + V_{m(\varepsilon)}$, where

$$V_{m(\varepsilon)} = \mu \sum_{i=1}^{m(\varepsilon)} (I - \mu B)^{m(\varepsilon)-i} S_0 (I - \mu B)^{m(\varepsilon)-i}$$

and

$$\begin{aligned} I_{m(\varepsilon)} &= \mu \sum_{i=1}^{m(\varepsilon)} (I - \mu B)^{i-1} (\xi_{-i} \xi_{-i}^T - S_0)(I - \mu B)^{i-1} \\ &= \mu \sum_{i=1}^{m(\varepsilon)} (I - \mu B)^{i-1} g_{-i} (I - \mu B)^{-1} i. \end{aligned}$$

Summing by parts we get

$$\begin{aligned} I_{m(\varepsilon)} = {} & \mu m(\varepsilon)(I - \mu B)^{m(\varepsilon)} \epsilon(m(\varepsilon))(I - \mu B)^{m(\varepsilon)} \\ & + \mu \sum_{i=1}^{m(\varepsilon)-1} i(I - \mu B)^{i-1} \epsilon(i)(I - \mu B)^{i-1} - i(I - \mu B)^{i} \epsilon(i)(I - \mu B)^{i}. \end{aligned}$$

Since $\epsilon(t) \to 0$, for any $\epsilon' > 0$ one can choose t_0 such that $\epsilon(t) \leq \epsilon'$ for all $t \geq t_0$ (note that t_0 does not depend on μ). Thus from the definition of $m(\varepsilon)$ in (36) we have

$$\begin{aligned} |I_{m(\varepsilon)}| &\leq K(\varepsilon)\varepsilon^2 |\epsilon(m(\varepsilon))| + K\mu^2 \sum_{i=0}^{m(\varepsilon)} (1 - \alpha\mu)^{2i} i |\epsilon(i)| \\ &\leq K(\varepsilon)\varepsilon^2 |\epsilon(m(\varepsilon))| + K t_0 \mu^2 \sum_{i=0}^{t_0} (1 - \alpha\mu)^{2i} |\epsilon(i)| \\ &\quad + K\epsilon' \mu^2 \sum_{i=0}^{\infty} i(1 - \alpha\mu)^{2i}. \qquad (42) \end{aligned}$$

For the first term in (42) we have immediately $\varepsilon^2 |\epsilon(m(\varepsilon))| \xrightarrow{P} 0$ as $\mu \to 0$. Since the random time t_0 is independent of μ, we get that

$$K t_0 \mu^2 \sum_{i=0}^{t_0} (1 - \alpha\mu)^{2i} |\epsilon(i)| \xrightarrow{P} 0$$

as $\mu \to 0$. For the last term in (42) we derive

$$K\epsilon\mu^2 \sum_{i=0}^{\infty} (1 - \alpha\mu)^{2i} i \leq K\epsilon \int_0^{\infty} e^{-2\alpha t} t dt \leq K\epsilon.$$

Summing up, we obtain from (42) that $|I_{m(\varepsilon)}| \xrightarrow{P} 0$. Now it suffices to prove that $V_{m(\varepsilon)} \to V$. Note that V_t adheres to the following recursive equation:

$$V_t = (I - \mu B)V_{t-1}(I - \mu B) + \mu S_\mu, \quad V_0 = 0.$$

Let V be a solution of the Lyapunov equation (7). Then for the difference $U_t = V_t - V$ we obtain

$$U_t = (I - \mu B)U_{t-1}(I - \mu B) + \mu^2 BVB.$$

From the stability of matrix $(I - \mu B)$ we conclude that

$$|U_{m(\varepsilon)}| \leq K(1 - \mu\alpha)^{2m(\varepsilon)} + Ko(\mu) = o(\mu)$$

as $\mu \to 0$. □

PROPOSITION 5. *Let the process* (Δ'_t) *be defined by the equation*

$$\Delta'_t = (I - \mu B)\Delta'_{t-1} + \mu\xi_t, \quad t \geq 1.$$

Moreover, for any $\epsilon_1 > 0$ *there exists* $R_1 < \infty$ *such that the initial condition* $\Delta'_0 = \bar{\Delta}_0(\mu)$ *satisfies*

$$P(|\Delta'_0| > R_1\mu^{1/2}) \leq \epsilon_1$$

for any $\varepsilon > 0$, *and*

$$\lim_{\mu \to 0} |Ef(\mu^{-1/2}\Delta'_{m(\varepsilon)}) - Ef(\mathcal{N}(0, V))| \leq \epsilon\|f\|_\infty$$

for any continuous f.

Proof. Note that

$$\Delta'_{m(\varepsilon)} = (I - \mu B)^{m(\varepsilon)}\Delta'_0 + \mu \sum_{i=1}^{m(\varepsilon)} (I - \mu B)^{m(\varepsilon)-i}\xi_i = I^{(1)}_{m(\varepsilon)} + I^{(2)}_{m(\varepsilon)}. \tag{43}$$

Recall that we have chosen $m(\varepsilon)$ in (36) such that $|(I - \mu B)^{m(\varepsilon)}| \leq \varepsilon$. Thus

$$|I^{(1)}_{m(\epsilon)}| \leq \varepsilon|\Delta'_0|,$$

and we have for the first term in (43)

$$P(|I^{(1)}_{m(\varepsilon)}| > \varepsilon R_1\mu^{1/2}) \leq P(|\Delta'_0| > R_1\mu^{1/2}) < \epsilon_1.$$

Next we apply the CLT for martingales to the second term. It states that $\mu^{-1/2}I^{(2)}_{m(\varepsilon)} \xrightarrow{D} \mathcal{N}(0, V)$ if condition (41) is satisfied along with the Lindeberg conditions [10, Thm. 5.4]:

$$\sum_{i=1}^{m(\varepsilon)} \mu|(I - \mu B)^{m(\varepsilon)-i}\xi_i|^2 1_{\sqrt{\mu}|(I-\mu B)^{m(\varepsilon)-i}\xi_i|>\delta} \xrightarrow{P} 0 \quad \text{for any } \delta > 0. \tag{44}$$

Condition (44) is satisfied since (ξ_i) is a strictly stationary sequence, and for some $\alpha > 0$

$$\begin{aligned} &E\left(\sum_{i=1}^{m(\varepsilon)} \mu|(I - \mu B)^{m(\varepsilon)-i}\xi_i|^2 1_{|(I-\mu B)^{m(\varepsilon)-i}\xi_i|>\delta\mu^{-1/2}}\right) \\ &\leq K \sum_{i=1}^{m(\varepsilon)} \mu(1 - \mu\alpha)^{2(m(\varepsilon)-i)} E(|\xi_i|^2 1_{K(I-\mu\alpha)^{m(\varepsilon)-i}|\xi_i|>\delta\mu^{-1/2}}) \\ &\leq KE|\xi_1|^2 1_{|\xi_1|>K\delta\mu^{-1/2}} \to 0 \end{aligned}$$

as $\mu \to 0$. Along with lemma 7 it implies the desired statement. □

The following proposition completes the proof of the theorem.

PROPOSITION 6.

$$\lim_{\mu\to 0}\lim_{t\to\infty} \mu^{-1/2}\bar\Delta_t = \mathcal{N}(0,V) \textit{ in distribution.}$$

Proof. Let us fix any $\delta > 0$. From Proposition 4 we conclude that there is $R < \infty$ such that $P(|\bar\Delta_{t_0}| \le \mu^{1/2}R) \ge 1-\delta$ for all $t_0 \ge (\alpha_0\mu)^{-1}\ln\mu^{-1}$. Let us choose $m(\varepsilon)$ to satisfy (36) with $\varepsilon = \delta/R$ (recall that the estimate (37) implies that $m(\varepsilon) \le K(\varepsilon)\mu^{-1}$ for some $K(\varepsilon) < \infty$). Denote

$$A = \left\{ \sup_{t_0\le t\le t_0+m(\varepsilon)} |\bar\Delta_t| 1_{|\bar\Delta_{t_0}|\le\mu^{1/2}R} \le \mu^{1/2}R_0 \right\}$$

and $B = \{|\bar\Delta_{t_0}| \le \mu^{1/2}R\}$. Lemma 5 implies that

$$\begin{aligned} 1-P(A) = P(A^c) &\le (\log\mu R_0^2+1)^{-1}\left[E\log(\mu R^2+1) + K\mu(m(\varepsilon))^{1/2} + \mu m(\varepsilon))\right] \\ &\le (\log\mu R_0^2+1)^{-1}(\log(\mu R^2+1) + (K(\delta/R)+1)\mu^{1/2}). \end{aligned}$$

Thus, one can choose $R_0 < \infty$ such that $P(A) \ge 1-\delta$. Note then that

$$\left\{ \sup_{t_0\le t\le t_0+m(\varepsilon)} |\bar\Delta_t| \le \mu^{1/2}R_0 \right\} \supseteq A\cap B.$$

Therefore,

$$P\left(\sup_{t_0\le t\le t_0+m(\varepsilon)} |\bar\Delta_t| \le \mu^{1/2}R_0 \right) \ge P(A\cap B) \ge 1-2\delta$$

or

$$P(\sigma'(t_0) < t_0 + m(\varepsilon)) < 2\delta \tag{45}$$

where $\sigma'(t)$ is defined in (35). Let us consider along with $\bar\Delta_t$ a random process Δ'_t defined by the equation

$$\begin{aligned} &\Delta'_t = (I-\mu B)\Delta'_{t-1} - \mu\xi_t \quad \text{for } t > t_0, \\ &\Delta'_t = \bar\Delta_t \quad \text{for } 0 \le t \le t_0. \end{aligned}$$

Then Lemma 6, along with the bound (45), implies that

$$P(\mu^{-1/2}|\Delta'_{t_0+m(\varepsilon)} - \bar\Delta_{t_0+m(\varepsilon)}| > \delta) < 3\delta$$

for all $\mu \le \mu_0$ small enough. On the other hand, all the conditions of Proposition 5 are fulfilled, thus $\Delta'_{t_0+m(\varepsilon)} \sim \mathcal{N}(0,V)$, which results, in turn, in $\bar\Delta_{t_0+m(\varepsilon)} \sim \mathcal{N}(0,V)$.

To conclude the proof, we note that the time $t_0 \ge (\alpha_0\mu)^{-1}\ln\mu^{-1}$and $\delta > 0$ have been arbitrarily chosen. □

Acknowledgment. The authors would like to acknowledge stimulating discussions with Pierre Priouret.

REFERENCES

[1] A. Benveniste, M. Metivier, and P. Priouret, *Adaptive Algorithms and Stochastic Approximations*, Springer-Verlag, Berlin, 1990.

[2] S. Bittanti and M. Campi, *Adaptive RLS algorithms under stochastic excitation - L^2 convergence analysis*, IEEE Trans. Automat. Control, AC-36 (1991), pp. 963–967.

[3] D. Blackwell and L. E. Dubins, *An extension of Skorohod's almost sure representation theorem*, Proc. Amer. Math. Soc., 89 (1983), pp. 691–692.

[4] L. Guo, L. Ljung, and P. Priouret, *Performance Analysis of Forgetting factor* RLS, Proc. 31st CDC, Tuscon, AZ, December 1992.

[5] P. Hall and C. Heyde, *Martingale Limit Theorem and Its Applications*, Academic Press, New York, 1980.

[6] I. A. Ibragimov and Yu. V. Linnik, *Independent and Stationary Sequences of Random Variables*, Wolters–Noordhoff, Groningen, 1971.

[7] A. Juditsky and P. Priouret, *A robust algorithm for random parameter tracking*, submitted for publication.

[8] A. Juditsky and A. Nazin, *Optimal and robust estimation of slowly drifting parameters of linear regression*, Avtomat. i Telemekh., 6 (1991), pp. 66–75. (In Russian; to be translated in Automat. Remote Control).

[9] H. Kushner and H. Huang, *Asymptotic properties of stochastic approximations with constant coefficients*, SIAM J. Control Optim., 19 (1981), pp. 87–105.

[10] R. S. Liptser and A. N. Shiryaev, *Martingale Theory,* Nauka, Moscow, 1986. (In Russian.)

[11] L. Ljung and S. Gunnarson, *Adaptation and tracking in system identification—a survey*, Automatica–J. IFAC, 26 (1990), pp. 7–21.

[12] L. Ljung, *An exact formula for the tracking ability of adaptive algorithms*, Tech. report LITH-ISY-I-1189, Linkoping University, Linkoping, Sweden.

[13] L. Ljung and P. Priouret, *A result on the mean square error obtained using general tracking algorithm*, Internat. J. Adapt. Control Signal Process., 5 (1992), pp. 231–250.

[14] ———, *Remarks on the mean square tracking error*, Internat. J. Adapt. Control Signal Process., 6 (1991), pp. 395–404.

[15] O. Macchi and E. Eweda, *Second order convergence analysis of stochastic adaptive linear filtering*, IEEE Trans. Automat. Control, AC-28 (1983), pp. 76–85.

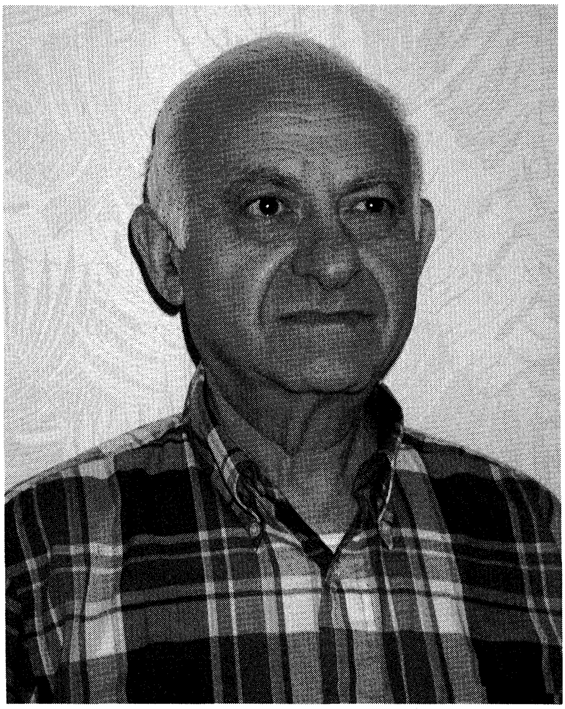
[16] S. P. Meyn and R. L. Tweedie, *Stability of Markovian processes* I: *criteria for discrete-time chains*, Adv. Appl. Prob., 24 (1992), pp. 542–574.

[17] E. Nummelin, *General Irreducible Markov Chains and Non-negative Operators*, Cambridge University Press, London, 1984.

[18] M. Niedźmiewcki and L. Guo, *Nonasymptotic results for finite-memory* WLS *filters*, IEEE Trans. Automat. Control, AC-36 (1991), pp. 198–206.

[19] R. L. Tweedie, *The existence of moments for stationary Markov chains*, J. Appl. Probab. 20 (1983), pp. 191–196.

[20] B. Widrow, J. Mc Cool, M. G. Larimore, and C. R. Johnson, *Stationary and nonstationary learning characteristics of the* LMS *adaptive filter*, Proc. IEEE, 64 (1976), pp. 1151–1161.



Jack Warga

SIAM J. CONTROL AND OPTIMIZATION
Vol. 33, No. 2, pp. 347–350, March 1995

© 1995 Society for Industrial and Applied Mathematics

001

JACK WARGA: IN APPRECIATION

The Editorial Board of the *SIAM Journal on Control and Optimization* (SICON) and members of the SIAM community take great pleasure in honoring Professor Jack Warga on the occasion of his retirement as Professor of Mathematics at Northeastern University. Jack's dedication to SICON, especially in its early formative years, was crucial to the establishment of the journal as one of the preeminent publications in the fields of control theory and optimization. In addition, his many contributions to mathematical analysis and optimal control theory have proved to be fundamental for subsequent developments in these important fields.

Jack Warga was born in Warsaw, Poland, on December 5, 1922. He received a B.A. degree in physics from Carleton College in 1944 and a Ph.D. degree in mathematics from New York University in 1950. From 1951 to 1956 he was employed as a mathematical and computing analyst by Reeves Instrument Corporation, New York, New York, Republic Aviation Corporation, Farmingdale, New York, and the ElectroData Division of Burroughs Corporation, Pasadena, California. He was a mathematical analyst and later Manager of the Mathematics Department of Avco Research and Development Division, Wilmington, Massachusetts, from 1957 to 1966. From 1966 until his retirement in July 1993, he was Professor of Mathematics at Northeastern University, Boston, Massachusetts. In addition, Professor Warga held a Weizmann Memorial Fellowship at the Weizmann Institute of Science, Rehovot, Israel, during 1956 and 1957, and spent his sabbatical leave there in 1973. During 1981 he was on sabbatical at Tel Aviv University, Tel Aviv, Israel. He has been an invited speaker at numerous conferences and workshops in the United States, Canada, Italy, and Israel.

Jack Warga joined the editorial board of SICON in 1964, the second year of publication of the journal (then called the *SIAM Journal on Control*), and served with distinction in various capacities for over 25 years. He was Co-Managing Editor with Lucien W. Neustadt from late 1967 until the end of 1969 and Associate Managing Editor from 1969 to 1972, sharing those duties with Leonard D. Berkovitz from 1970 to 1972. For six years, from 1973 through 1978, Jack served as Managing Editor of the journal.

A pioneering researcher in mathematical control theory, Professor Warga's work covers a board spectrum of problems in the field. He has contributed to the theory of necessary conditions for optimal control problems, especially the theory of relaxed controls. Much of this effort is summarized in his book [24], in which the inherent convexity of relaxed controls is exploited in order to address existence questions, as well as to derive strong-variation necessary conditions. In later work he developed a theory of generalized differentiation for nonsmooth functions as well as necessary optimality and controllability conditions for optimization and optimal control problems [26]–[29], [35], [36], [40], [43], [44], [54] , [57]. Among Jack's other significant contributions are those that address higher-order necessary conditions [32]–[34], [39], [47]–[49], [52], iterative computational techniques [2], [6], [42], [46], and related functional analytic issues [41], [61], [62].

In his service to SICON, as well as in his research, Jack is known for his high standards of scholarship, personal integrity, and helpfulness to fellow researchers. Especially noteworthy is his personal concern for young researchers; he has provided guidance and encouragement with extraordinary patience and kindness to many. Over the past 30 years he has worked extensively and anonymously for human rights for scientists throughout the world. He is truly beloved by his students and colleagues for his humble manner and exceptional generosity.

The editorial board is honored to join Jack Warga's many friends and admirers in formally recognizing his many contributions to mathematics and his outstanding service to SIAM and the mathematical community at large.

PUBLICATIONS OF JACK WARGA

[1] *On the representation of large integers as sums of primes* (with H. N. Shapiro), Comm. Pure Appl. Math., 3 (1950), pp. 153–176.

[2] *On a class of iterative procedures for solving normal systems of ordinary differential equations*, J. Math. Phys., 31 (1953), pp. 223–243.

[3] *Screened interactions of neighbor ions and symmetry properties of titration curves in polyelectrolytes* (with S. Lifson), J. Chem. Phys., 29 (1958), pp. 643–644.

[4] *Relaxed variational problems*, J. Math. Anal. Appl., 4 (1962), pp. 111–128.

[5] *Necessary conditions for minimum in relaxed variational problems*, J. Math. Anal. Appl., 4 (1962), pp. 129–145.

[6] *A convergent procedure for convex programming*, J. Soc. Indust. Appl. Math., 11 (1963), pp. 579–587.

[7] *Minimizing certain convex functions*, J. Soc. Indust. Appl. Math., 11 (1963), pp. 588–593.

[8] *A convergent procedure for solving the thermo-chemical equilibrium problem*, J. Soc. Indust. Appl. Math., 11 (1963), pp. 594–606.

[9] *Minimizing variational curves restricted to a preassigned set*, Trans. Amer. Math. Soc., 112 (1964), pp. 432–455.

[10] *The determination of extreme entry angles into a planetary atmosphere* (with W. Hailey), J. AIAA, 2 (1964), pp. 335–338.

[11] *On a class of minimax problems in the calculus of variations*, Michigan Math. J., 12 (1965), pp. 289–311.

[12] *Unilateral variational problems with several inequalities*, Michigan Math. J., 12 (1965), pp. 449–480.

[13] *Minimax problems and unilateral curves in the calculus of variations*, J. Soc. Indust. Appl. Math., Ser. A, Control, 3 (1965), pp. 91–105.

[14] *Variational problems with unbounded controls*, J. Soc. Indust. Appl. Math., Ser. A, Control, 3 (1965), pp. 424–438.

[15] *Functions of relaxed controls*, SIAM J. Control, 5 (1967), pp. 628–646.

[16] *Restricted minima of functions of controls*, SIAM J. Control, 5 (1967), pp. 642–656.

[17] *The reduction of certain control problems to an "ordinary differential" type*, SIAM Rev., 10 (1968), pp. 219–222.

[18] *Relaxed controls for functional equations*, J. Funct. Anal., 5 (1970), pp. 71–93.

[19] *Control problems with functional restrictions*, SIAM J. Control, 8 (1970), pp. 360–371.

[20] *Unilateral and minimax control problems defined by integral equations*, SIAM J. Control, 8 (1970), pp. 372–382.

[21] *On a class of pursuit and evasion problems*, J. Differential Equations, 9 (1971), pp. 155–167.

[22] *Conflicting and minimax controls*, J. Math. Anal. Appl., 33 (1971), pp. 655–673.

[23] *Normal control problems have no minimizing strictly original solutions*, Bull. Amer. Math. Soc., 77 (1971), pp. 625–628.

[24] *Optimal Control of Differential and Functional Equations*, Academic Press, New York, 1972.

[25] *Optimal controls with pseudodelays*, SIAM J. Control, 12 (1974), pp. 286–299.

[26] *Necessary conditions without differentiability assumptions in optimal control*, J. Differential Equations, 18 (1975), pp. 41–62.

[27] *Necessary conditions without differentiability assumptions in unilateral control problems*, J. Differential Equations, 21 (1976), pp. 25–38.

[28] *Controllability and necessary conditions in unilateral problems without differentiability assumptions*, SIAM J. Control Optim., 14 (1976), pp. 546–572.

[29] *Derivate containers, inverse functions, and controllability*, in Calculus of Variations and Control Theory, D. L. Russell, ed., Academic Press, New York, 1976, pp. 13–46.

[30] *Optimal Control of Differential and Functional Equations*, Chapter XI (appended to the Russian translation), Nauka, Moscow, 1977.

[31] *Steepest descent with relaxed controls*, SIAM J. Control Optim., 15 (1977), pp. 674–682.

[32] *A second order condition that strengthens Pontryagin's maximum principle*, J. Differential Equations, 28 (1978), pp. 284–307.

[33] *Necessary conditions for minimum of order zero and two*, in Optimal Control and Differential Equations, A. B. Schwarzkopf et al., eds., Academic Press, New York, 1978, pp. 131–149.

[34] *A second order Lagrangian condition for restricted control problems*, J. Optim. Theory Appl., 24 (1978), pp. 465–473.

[35] *An implicit function theorem without differentiability*, Proc. Amer. Math. Soc., 69 (1978), pp. 65–69.

[36] *Controllability and a multiplier rule for nondifferentiable optimization problems*, SIAM J. Control Optim., 16 (1978), pp. 803–812.

[37] *Controllability of nondifferentiable hereditary processes*, SIAM J. Control Optim., 16 (1978), pp. 813–831.

[38] *Review of Optimization — A Theory of Necessary Conditions by L. W. Neustadt*, IEEE Trans. Automat. Control, AC-23 (1978), pp. 514–515.

[39] *A hybrid relaxed-Lagrangian second order condition for minimum*, in Differential Games and Control Theory, Vol. 3, P. T. Liu and E. Roxin, eds., Marcel Dekker, New York, 1979, pp. 77–94.

[40] *Fat homeomorphisms and unbounded derivate containers*, J. Math. Anal. Appl., 81 (1981), pp. 545–560; and J. Math. Anal. Appl., 90 (1982), pp. 582–583.

[41] *On bounding, interior and covering functions*, J. Math. Anal. Appl., 82 (1981), pp. 255–267.

[42] *Iterative procedures for constrained and unilateral optimization problems*, SIAM J. Control Optim., 20 (1982), pp. 360–376.

[43] *Optimization and controllability without differentiability assumptions*, SIAM J. Control Optim., 21 (1983), pp. 837–855.

[44] *Controllability, extremality and abnormality in nonsmooth optimal control*, J. Optim. Theory Appl., 41 (1983), pp. 239–260.

[45] *Review of Optimization — Theory and Applications, Problems with Ordinary Differential Equations by L. Cesari*, Bull. Amer. Math. Soc., 9 (1983), pp. 396–400.

[46] *Iterative optimization with equality constraints*, Math. Oper. Res., 9 (1984), pp. 592–605.

[47] *Second order necessary conditions in optimization*, SIAM J. Control Optim., 22 (1984), pp. 524–528.

[48] *Second order controllability and optimization with ordinary controls*, SIAM J. Control Optim., 23 (1985), pp. 49–60.

[49] *Higher order conditions with and without Lagrange multipliers*, SIAM J. Control Optim., 24 (1986), pp. 715–730.

[50] *Review of Applied Nonlinear Analysis by J.-P. Aubin and I. Ekeland*, Bull. Amer. Math. Soc. (N. S.), 17 (1987), pp. 351–358.

[51] *Homeomorphisms and local C*1 *approximations*, J. Nonlinear Anal., 12 (1988), pp. 593–597.

[52] *Higher order conditions for conical controllability*, SIAM J. Control Optim., 26 (1988), pp. 1471–1480.

[53] *Global directional controllability*, SIAM J. Control Optim., 27 (1989), pp. 976–990.

[54] *Local and global directional controllability: Sufficient conditions and examples*, in Nonsmooth Optimization and Related Topics, F. H. Clarke et al., eds., Plenum Press, New York, 1989, pp. 417–435.

[55] *Some selected problems of optimal control*, in Modern Optimal Control, E. O. Roxin, ed., Marcel Dekker, New York, 1989, pp. 389–407.

[56] *An extension of the Kaskosz maximum principle*, Appl. Math. Optim., 22 (1990), pp. 61–74.

[57] *Nonsmooth problems with conflicting controls*, SIAM J. Control Optim., 29 (1991), pp. 678–701.

[58] *A proper relaxation of shifted and delayed controls* (with Q. J. Zhu), J. Math. Anal. Appl., 169 (1992), pp. 546–561.

[59] *A necessary and sufficient condition for a constrained minimum*, SIAM J. Optim., 2 (1992), pp. 665–667.

[60] *A topological index condition for conical controllability* (with G. X. Fang), Nonlinear Anal., 19 (1992), pp. 1001–1007.

[61] *Functions with unstable images* I. *Cracks of codimension* 1 (with G. X. Fang), Nonlinear Anal., 19 (1992), pp. 1047–1061.

[62] *Functions with unstable images* II. *Cracks of positive codimensions* (with G. X. Fang), Nonlinear Anal., 19 (1992), pp. 1179–1186.

[63] *The equivalence of extremals in different representations of unbounded control problems* (with Q. J. Zhu), SIAM J. Control Optim., 32 (1994), pp. 1151–1169.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 33, No. 2, pp. 351–364, March 1995

© 1995 Society for Industrial and Applied Mathematics

002

$\mathbf{H}^\infty$ OPTIMAL SENSITIVITY FOR A CLASS OF INFINITE-DIMENSIONAL SYSTEMS*

HONG YANG†

Abstract. The computation of the $\mathbf{H}^\infty$ optimal weighted pure and mixed sensitivities for a class of infinite-dimensional systems is studied by characterizing certain wandering subspaces of some related Kreĭn spaces. The new characterizations can be used to obtain explicit bases for those subspaces. Explicit formulas for the optimal sensitivity for some infinite-dimensional systems can be obtained from these bases. The new characterizations can also be used to obtain fast algorithms for computing the $\mathbf{H}^\infty$ optimal performance for pure and mixed sensitivity problems. The results are applied to obtain an explicit optimal sensitivity formula for a class of infinite-dimensional systems that generalizes a known result.

Key words. $\mathbf{H}^\infty$ optimal sensitivity, infinite-dimensional systems, Kreĭn space

AMS subject classifications. 93B28, 93B35, 93B36

1. Introduction.

We study the computation of $\mathbf{H}^\infty$ optimal weighted pure and mixed sensitivities for a class of infinite-dimensional systems. In the standard frequency domain approach for $\mathbf{H}^\infty$ control, it is has been shown that for a quite general class of systems, the original $\mathbf{H}^\infty$ sensitivity minimization problem is associated with a self-adjoint operator. For the scalar (single-input/single-output) systems, in the case when the norm of this operator is given by its largest eigenvalue, the minimal pure sensitivity is unique and can be found by means of Sarason's Theorem [16, Prop. 5.1, p. 188]. In this case Sarason's Theorem can also be used to obtain the unique minimal mixed sensitivity [4]. In the case when the norm of this operator is given by its essential spectral radius, the optimal sensitivity may not be unique. Using the Kreĭn space approach developed by Ball and Helton [1], [17] and [4] studied the parameterization of optimal pure and mixed sensitivities in the nonunique case respectively. In this approach, one needs to compute the bases for certain subspaces of some related Kreĭn spaces. Although [17] and [4] gave abstract formulas for computing such bases, for the infinite-dimensional systems, the actual computation using those formulas involves computing an inverse image of an infinite-dimensional operator, an image of an infinite-dimensional operator, and the norm of an inverse image of the square root of an infinite-dimensional operator. Explicitly carrying out these operations is difficult if not impossible. In the matrix (multi-input/multi-output) case, even when the norm of the related self-adjoint operator is given by its largest eigenvalue, the optimal pure sensitivity may not be unique [10]. Intuitively speaking, this is because in the matrix case we only minimize the largest singular value of a matrix and there is some leeway left for the smaller singular values. In the scalar case there is no such leeway.

We begin with the mixed sensitivity case; by studying the actions of various shift operators on the elements of the related subspace, we observe some "symmetry" properties and obtain a new characterization for the subspace. The pure sensitivity case can be treated as a special case. The characterizations enable us to compute explicitly

* Received by the editors October 26, 1992; accepted for publication (in revised form) October 13, 1993. This research was partially supported by Air Force Office of Scientific Research grants AFOSR-89-0205 and AFOSR-91-0231.

† Program in Applied and Computational Mathematics and Department of Electrical Engineering, Princeton University, Princeton, New Jersey 08544. Current address: Department of Chemical Engineering, Princeton University, Princeton, New Jersey 08544 (hyang@titan.princeton.edu).

the bases for the related subspaces. These bases in turn give us parameterizations of all optimal pure and mixed sensitivities respectively.

The paper is organized as follows: Notation is introduced in the next section. Computation of optimal mixed sensitivity is considered in §3, and the special case of optimal pure sensitivity is considered in §3.1. In §4 the results are applied to a class of infinite-dimensional systems to obtain an explicit optimal sensitivity formula that generalizes some known results. Some concluding remarks are made in §5.

2. Notation. In this paper we shall work on the Hardy spaces on the unit disc in order to apply more directly the mathematical literature, except for §4, where examples are computed on the Hardy spaces on the right half plane. We shall use s and z to denote the complex variables in the right half plane and the unit disc $\mathbf{D}$ respectively. Basic definitions and facts used in this paper about Hardy spaces and Kreĭn spaces can be found in [12] and [2] respectively.

$\mathbf{H}^2$, $\mathbf{H}^\infty$, $\mathbf{L}^2$, $\mathbf{L}^\infty$: Hardy spaces and Lebesgue spaces.

$\mathbf{H}^2_-$: the orthogonal complement of $\mathbf{H}^2$ in $\mathbf{L}^2$, i.e., $\mathbf{L}^2 = \mathbf{H}^2 \oplus \mathbf{H}^2_-$.

$B_{\mathbf{H}^\infty} := \{\phi \in \mathbf{H}^\infty, \|\phi\|_\infty \leq 1\}$, the unit ball of $\mathbf{H}^\infty$.

$\mathbf{H}^2(0) := \{h \in \mathbf{H}^2, h(0) = 0\}$.

Π_+: the orthogonal projection $\mathbf{L}^2 \to \mathbf{H}^2$.

Π_-: the orthogonal projection $\mathbf{L}^2 \to \mathbf{H}^2_-$.

$x^*(z)$, $x^*(s)$: the involutions of $x(z)$ and $x(s)$ respectively, i.e., $x^*(z) := \bar{x}(\frac{1}{\bar{z}})$ and $x^*(s) := \bar{x}(-\bar{s})$. Here the bar denotes the complex conjugate. Note: when "*" is applied to an operator, it denotes the adjoint.

$F^{(*)} := F^*F$, where F can be an operator or a function.

$\mathbf{K}_1$: the Kreĭn space $\mathbf{L}^2 \oplus \mathbf{H}^2$ with indefinite inner product

$$\left[\begin{pmatrix} u_1 \\ v_1 \end{pmatrix}, \begin{pmatrix} u_2 \\ v_2 \end{pmatrix}\right] = \langle u_1, u_2\rangle_{\mathbf{L}^2} - \langle v_1, v_2\rangle_{\mathbf{H}^2}.$$

$\mathbf{K}_2$: the Kreĭn space $\mathbf{L}^2 \oplus \mathbf{H}^2 \oplus \mathbf{H}^2$ with indefinite inner product

$$\left[\begin{pmatrix} u_1 \\ v_1 \\ w_1 \end{pmatrix}, \begin{pmatrix} u_2 \\ v_2 \\ w_2 \end{pmatrix}\right] = \langle u_1, u_2\rangle_{\mathbf{L}^2} + \langle v_1, v_2\rangle_{\mathbf{H}^2} - \langle w_1, w_2\rangle_{\mathbf{H}^2}.$$

$M^{[\perp]}$: orthogonal companion of M with respect to the indefinite inner product $[\cdot,\cdot]$.

s_1: the bilateral shift operator on $\mathbf{L}^2$.

s_2: the unilateral shift operator on $\mathbf{H}^2$.

S_1: shift operator on $\mathbf{K}_1$ defined as $S_1 \begin{pmatrix} u \\ v \end{pmatrix} := \begin{pmatrix} s_1 u \\ s_2 v \end{pmatrix}$.

S_2: shift operator on $\mathbf{K}_2$ defined as $S_2 \begin{pmatrix} u \\ v \\ w \end{pmatrix} := \begin{pmatrix} s_1 u \\ s_2 v \\ s_2 w \end{pmatrix}$.

$\mathcal{G}(\mathcal{X})$: the graph of $\mathcal{X}$, i.e., if $\mathcal{X} : H_1 \to H_2$, then

$$\mathcal{G}(\mathcal{X}) = \left\{\begin{pmatrix} h_1 \\ \mathcal{X}(h_1) \end{pmatrix} : h_1 \in H_1\right\} \subset H_1 \oplus H_2.$$

3. Optimal sensitivities. We begin with the mixed sensitivity case. It is well known [14], [6], [4] that for a very general class of infinite-dimensional systems (includes systems with infinite-dimensional inner part and finite-dimensional outer part and finite-dimensional weights), minimizing the mixed sensitivity in $\mathbf{H}^\infty$ norm is equivalent to the following problem:

Given matrices $G \in \mathbf{L}^\infty$ and $F \in \mathbf{H}^\infty$, find all $Z_o \in \mathbf{H}^\infty$ such that

$$\left\| \begin{pmatrix} G - Z_o \\ F \end{pmatrix} \right\|_\infty = \mu_o = \min_{Z \in \mathbf{H}^\infty} \left\| \begin{pmatrix} G - Z \\ F \end{pmatrix} \right\|_\infty . \tag{1}$$

It has been shown (see [8], [7], [11], [6], and the references therein) that the *optimal performance* μ_o can be computed in the case when $G = m^* W$ and F rational, where m is arbitrary inner and W is rational. In this paper, we shall assume that the optimal performance μ_o is known. Define $\tilde{G} = \frac{1}{\mu_o} G, \tilde{F} = \frac{1}{\mu_o} F$; then the above problem is equivalent to:

Given matrices $\tilde{G} \in \mathbf{L}^\infty$ and $\tilde{F} \in \mathbf{H}^\infty$, find all $\tilde{Z}_o \in \mathbf{H}^\infty$ such that

$$\left\| \begin{pmatrix} \tilde{G} - \tilde{Z}_o \\ \tilde{F} \end{pmatrix} \right\|_\infty = 1 = \min_{\tilde{Z} \in \mathbf{H}^\infty} \left\| \begin{pmatrix} \tilde{G} - \tilde{Z} \\ \tilde{F} \end{pmatrix} \right\|_\infty ,$$

where $\tilde{Z}$ is related to Z by $Z = \mu_o \tilde{Z}$. In the following, we shall assume $\mu_o = 1$.

For $\binom{G}{F} \in \mathbf{L}^\infty \oplus \mathbf{H}^\infty$, define an operator

$$\mathcal{A} = \mathcal{A}(G, F) : \mathbf{H}^2 \to \mathbf{H}^2_- \oplus \mathbf{H}^2 \quad (\mathcal{A}h = (\Pi_- Gh \quad \Pi_+ Fh)) .$$

Since we assume $\mu_o^2 = 1$, we have [13]

$$\|\mathcal{A}^{(*)}\| = \mu_o^2 = 1.$$

Note that modifying G by adding an $\mathbf{H}^\infty$ function does not change the operator $\mathcal{A}$; and conversely, for any $\Phi \in \mathbf{L}^\infty$ such that $\mathcal{A}(G, F) = \mathcal{A}(\Phi, F)$, we have

$$\min_{Z \in \mathbf{H}^\infty} \|G - \Phi - Z\|_\infty = \|\mathcal{A}(G - \Phi, 0)\| = \|\mathcal{A}(G, F) - \mathcal{A}(\Phi, F)\| = 0,$$

which means that $G - \Phi \in \mathbf{H}^\infty$. We see that the problem (1) is equivalent to finding all $\Phi \in \mathbf{L}^\infty$ such that $\mathcal{A}(G, F) = \mathcal{A}(\Phi, F)$ and $\left\| \binom{\Phi}{F} \right\|_\infty = 1$. We shall call each such Φ a minimal symbol of $\mathcal{A}(G, F)$. From $\Phi = G - Z_o$ we get $Z_o = G - \Phi$.

It is easy to see that the adjoint operator of $\mathcal{A}$ is

$$\mathcal{A}^* : \mathbf{H}^2_- \oplus \mathbf{H}^2 \to \mathbf{H}^2, \quad \mathcal{A}^* \begin{pmatrix} h_- \\ h \end{pmatrix} = (\Pi_+ G^* \quad \Pi_+ F^*) \begin{pmatrix} h_- \\ h \end{pmatrix}$$
$$= \Pi_+ G^* h_- + \Pi_+ F^* h.$$

The graph of $\mathcal{A}^*$ is

$$\mathcal{G}(\mathcal{A}^*) = \left\{ \begin{pmatrix} h_- \\ h \\ \Pi_+ G^* h_- + \Pi_+ F^* h \end{pmatrix} : h_- \in \mathbf{H}^2_-, h \in \mathbf{H}^2 \right\} .$$

Let

$$\mathcal{M}_2 = \mathcal{G}(\mathcal{A}^*)^{[\perp]}.$$

Corresponding to Lemma 3 and Lemma 2 in [10] for the pure sensitivity case, we have the following lemma.

LEMMA 3.1. *A subspace $\mathcal{N}_2$ of $\mathbf{K}_2$ is maximal negative if and only if it is the graph of a contraction $\mathcal{C} : \mathbf{H}^2 \to \mathbf{L}^2 \oplus \mathbf{H}^2$, i.e.,*

$$\mathcal{N}_2 = \left\{ \begin{pmatrix} \mathcal{C}(x) \\ x \end{pmatrix} : x \in \mathbf{H}^2 \right\},$$

where $\|\mathcal{C}\| \leq 1$.

LEMMA 3.2. *If $\mathcal{N}_2$ is the graph of a bounded linear operator $\mathbf{H}^2 \to \mathbf{L}^2 \oplus \mathbf{H}^2$, then $\mathcal{N}_2$ is S_2-invariant if and only if*

$$\mathcal{N}_2 = \left\{ \begin{pmatrix} \Phi_1 x \\ \Phi_2 x \\ x \end{pmatrix} : x \in \mathbf{H}^2 \right\}$$

for some $\Phi_1 \in \mathbf{L}^\infty$ and some $\Phi_2 \in \mathbf{H}^\infty$.

From Lemmas 3.1 and 3.2 we see that for any S_2-invariant maximal negative subspace $\mathcal{N}_2$, there are two matrices $\Phi_1 \in \mathbf{L}^\infty$ and $\Phi_2 \in \mathbf{H}^\infty$ associated with it. Further, we have the following lemma.

LEMMA 3.3. *For an S_2-invariant maximal negative subspace $\mathcal{N}_2$, $\mathcal{N}_2 \subset \mathcal{M}_2$ if and only if*

$$\mathcal{H}_{\Phi_1} = \mathcal{H}_G \quad \textit{and} \quad \Phi_2 = F.$$

Proof. Since $\mathcal{N}_2$ is maximal negative and S_2-invariant, by Lemmas 3.1 and 3.2, there are two matrices $\Phi_1 \in \mathbf{L}^\infty$ and $\Phi_2 \in \mathbf{H}^\infty$ such that

$$\mathcal{N}_2 = \left\{ \begin{pmatrix} \Phi_1 x \\ \Phi_2 x \\ x \end{pmatrix} : x \in \mathbf{H}^2 \right\}.$$

Further, $\mathcal{N}_2 \subset \mathcal{M}_2$ if and only if for any $x \in \mathbf{H}^2, y_- \in \mathbf{H}^2_-$, and $y \in \mathbf{H}^2$,

$$\begin{aligned}
0 &= \left[\begin{pmatrix} \Phi_1 x \\ \Phi_2 x \\ x \end{pmatrix}, \begin{pmatrix} y_- \\ y \\ \mathcal{H}_G^* y_- + \mathcal{T}_F^* y \end{pmatrix} \right] \\
&= \langle \Phi_1 x, y_- \rangle + \langle \Phi_2 x, y \rangle - \langle x, \mathcal{H}_G^* y_- + \mathcal{T}_F^* y \rangle \\
&= \langle \Pi_- \Phi_1 x, y_- \rangle + \langle \Phi_2 x, y \rangle - \langle x, \mathcal{H}_G^* y_- \rangle - \langle x, \mathcal{T}_F^* y \rangle \\
&= \langle \Pi_- \Phi_1 x, y_- \rangle + \langle \Phi_2 x, y \rangle - \langle x, \Pi_+ G^* y_- \rangle - \langle x, \Pi_+ F^* y \rangle \\
&= \langle \Pi_- \Phi_1 x, y_- \rangle + \langle \Phi_2 x, y \rangle - \langle x, G^* y_- \rangle - \langle x, F^* y \rangle \\
&= \langle \Pi_- \Phi_1 x, y_- \rangle + \langle \Phi_2 x, y \rangle - \langle G x, y_- \rangle - \langle F x, y \rangle \\
&= \langle \Pi_- \Phi_1 x, y_- \rangle + \langle \Phi_2 x, y \rangle - \langle \Pi_- G x, y_- \rangle - \langle F x, y \rangle \\
&= \langle \Pi_- \Phi_1 x - \Pi_- G x, y_- \rangle + \langle \Phi_2 x - F x, y \rangle.
\end{aligned}$$

This means $\mathcal{H}_{\Phi_1} = \mathcal{H}_G$ and $\Phi_2 = F$. □

Remark 1. From $\mathcal{H}_{\Phi_1} = \mathcal{H}_G$ and $\Phi_2 = F$ we can infer that $\mathcal{A}(\Phi_1, F) = \mathcal{A}(G, F)$, which in turn means $G - \Phi_1 \in \mathbf{H}^\infty$.

For clarity we summarize the above three lemmas in the following theorem.

THEOREM 3.4. *$\mathcal{N}_2$ is a maximal negative S_2-invariant subspace and $\mathcal{N}_2 \subset \mathcal{M}_2$ if and only if $\mathcal{N}_2$ is the graph of a multiplication operator $(\Phi_1 \quad F)^{\mathsf{T}} : \mathbf{H}^2 \to \mathbf{L}^2 \oplus \mathbf{H}^2$, where $\Phi_1 \in \mathbf{L}^\infty$ and*

$$\mathcal{A}(\Phi_1, F) = \mathcal{A}(G, F) \quad \textit{and} \quad \left\| \begin{pmatrix} \Phi_1 \\ F \end{pmatrix} \right\|_\infty \leq 1.$$

The above theorem characterizes the optimal mixed sensitivity in terms of S_2-invariant maximal-negative subspaces of $\mathcal{M}_2$. By Ball–Helton's Theorem [1], the S_2-invariant maximal-negative subspaces of $\mathcal{M}_2$ can be determined by the following wandering subspace:

$$\mathbf{L}_2 := \mathcal{M}_2 \cap (S_2\mathcal{M}_2)^{[\perp]}.$$

It is shown in [4] (see also [17]) that for $\|\mathcal{A}\| < 1$, $\dim \mathbf{L}_2 = 2$ and there exist

$$x_1 = \begin{pmatrix} p_1 \\ Fr_1 \\ r_1 \end{pmatrix} \in \mathbf{L}_2, \qquad x_2 = \begin{pmatrix} p_2 \\ Fr_2 \\ r_2 \end{pmatrix} \in \mathbf{L}_2$$

such that $[x_1, x_1] = 1, [x_2, x_2] = -1$, and $[x_1, x_2] = 0$ and that the suboptimal symbols Φ such that $\mathcal{A}(G, F) = \mathcal{A}(\Phi, F)$ and $\|(\begin{smallmatrix} \Phi \\ F \end{smallmatrix})\|_\infty < 1$ is parameterized by the formula

$$\Phi = \frac{p_1\phi + p_2}{r_1\phi + r_2}, \quad \phi \in B_{\mathbf{H}^\infty}.$$

In the optimal case, for $\|\mathcal{A}\| = 1$, consider $\tilde{\mathcal{A}} := \frac{1}{\mu}\mathcal{A}$ for $\mu > 1$; then $\|\tilde{\mathcal{A}}\| < 1$. The corresponding $\tilde{r}_2$ is such that

$$\sup_{\mu > 1} |\tilde{r}_2(z)| < \infty \quad \text{for some } z \in \mathbf{D}$$

if and only if the symbol for $\mathcal{A}$ in $B_{\mathbf{L}^\infty}$ is nonunique. In this case there is a sequence $\{\mu_n\}$ such that $1 < \mu_n \to 1$ and the corresponding $\tilde{p}_1, \tilde{r}_1, \tilde{p}_2, \tilde{r}_2$ converge to p_1, r_1, p_2, r_2 respectively and the optimal symbols (minimal symbols) Φ such that $\mathcal{A}(G, F) = \mathcal{A}(\Phi, F)$ and $\|(\begin{smallmatrix} \Phi \\ F \end{smallmatrix})\|_\infty = 1$ is parameterized by

$$\Phi = \frac{p_1\phi + p_2}{r_1\phi + r_2}, \quad \phi \in B_{\mathbf{H}^\infty}.$$

Now we see that in order to compute explicitly the optimal mixed sensitivity we need to find a base for the wandering subspace $\mathbf{L}_2$.

Remark 2. In [4], abstract formulas similar to those of [17] were derived for a base of $\mathbf{L}_2$. As the formulas in [17], for infinite-dimensional systems, those formulas contain an inverse image of an infinite-dimensional operator, an image of an infinite-dimensional operator, and the norm of an inverse image of the square root of an infinite-dimensional operator.

In order to find a base for the subspace $\mathbf{L}_2$ explicitly, we study $\mathbf{L}_2$ and derive some new characterizations for $\mathbf{L}_2$ in the following. Since $\mathbf{L}_2 := \mathcal{M}_2 \cap (S_2\mathcal{M}_2)^{[\perp]}$, we begin with a characterization of the subspace $\mathcal{M}_2$.

PROPOSITION 3.5. *$(p \quad q \quad r)^{\mathsf{T}} \in \mathcal{M}_2$ if and only if $p - Gr \in \mathbf{H}^2$ and $q - Fr \in \mathbf{H}^2_-$. If $F \in \mathbf{H}^\infty$, $(p \quad q \quad r)^{\mathsf{T}} \in \mathcal{M}_2$ if and only if $p - Gr \in \mathbf{H}^2$ and $q = Fr$.*

Proof. By the definition of $\mathcal{M}_2$, we have

$$\begin{aligned}\left[\begin{pmatrix} p \\ q \\ r \end{pmatrix}, \begin{pmatrix} h_- \\ h \\ \Pi_+ G^* h_- + \Pi_+ F^* h \end{pmatrix}\right] &= \langle p, h_-\rangle_{\mathbf{L}^2} + \langle q, h\rangle_{\mathbf{H}^2} - \langle r, \Pi_+ G^* h_- + \Pi_+ F^* h\rangle_{\mathbf{H}^2} \\ &= \langle p, h_-\rangle_{\mathbf{L}^2} + \langle q, h\rangle_{\mathbf{H}^2} - \langle r, G^* h_- + F^* h\rangle_{\mathbf{L}^2} \\ &= \langle p, h_-\rangle_{\mathbf{L}^2} + \langle q, h\rangle_{\mathbf{H}^2} - \langle Gr, h_-\rangle_{\mathbf{L}^2} - \langle Fr, h\rangle_{\mathbf{L}^2} \\ &= \langle p - Gr, h_-\rangle_{\mathbf{L}^2} + \langle q - Fr, h\rangle_{\mathbf{L}^2} = 0\end{aligned}$$

for any $h_- \in \mathbf{H}^2_-$ and $h \in \mathbf{H}^2$. This is true if and only if $p - Gr \in \mathbf{H}^2$ and $q - Fr \in \mathbf{H}^2_-$.

If $F \in \mathbf{H}^\infty$, then $q - Fr \in \mathbf{H}^2$, and this is true if and only if $q - Fr = 0$, i.e., $q = Fr$. □

Remark 3. From Proposition 3.5 we see that for any $h \in \mathbf{H}^2$, $(h \quad 0 \quad 0)^\top \in \mathcal{M}_2$. Also $(\Pi_- Gr \quad Fr \quad r)^\top \in \mathcal{M}_2$.

Remark 4. For $F \in \mathbf{H}^\infty$, $\mathcal{M}_2$ is S_2-invariant, i.e., $S_2\mathcal{M}_2 \subset \mathcal{M}_2$. In fact, for $(u \quad v \quad w)^\top \in S_2\mathcal{M}_2$, by Lemma 3.6 we have

$$\begin{pmatrix} u \\ v \\ w \end{pmatrix} = S_2 \begin{pmatrix} Gx + y \\ Fx \\ x \end{pmatrix} = \begin{pmatrix} s_1 Gx + s_1 y \\ s_2 Fx \\ s_2 x \end{pmatrix} = \begin{pmatrix} Gs_2 x + s_1 y \\ Fs_2 x \\ s_2 x \end{pmatrix}. \tag{2}$$

By Proposition 3.5, we see that the right-hand side of (2) is in $\mathcal{M}_2$. Therefore $S_2\mathcal{M}_2 \subset \mathcal{M}_2$.

We shall use the following lemma that deals with the commuting properties of the shifts s_1 and s_2 with multiplication operators.

LEMMA 3.6. *For any* $A \in \mathbf{L}^\infty$,

$$As_2 = s_1 A \quad on \quad \mathbf{H}^2.$$

In particular, if $A \in \mathbf{H}^\infty$,

$$As_2 = s_2 A \quad on \quad \mathbf{H}^2.$$

Proof. Immediate from the definitions of the shifts s_1 and s_2. □

Using Proposition 3.5, we can prove the following characterization of $\mathbf{L}_2$. We adopt the proof provided by one referee of the paper. The original proof is much more complicated.

THEOREM 3.7. *For* $F \in \mathbf{H}^\infty$, $(p \quad Fr \quad r)^\top \in \mathbf{L}_2 = \mathcal{M}_2 \cap (S_2\mathcal{M}_2)^{[\perp]}$ *if and only if*

$$p^*, \quad p - Gr, \quad (1 - F^{(*)})r^* - Gp^* \in \mathbf{H}^2.$$

Proof. Noting that $F \in \mathbf{H}^\infty$, by Proposition 3.5, we have

$$\mathcal{M}_2 = \left\{ \begin{pmatrix} Gx + y \\ Fx \\ x \end{pmatrix} : x, y \in \mathbf{H}^2 \right\}.$$

Then $(p \quad Fr \quad r)^\top \in (S_2\mathcal{M}_2)^{[\perp]}$ if and only if for any $x, y \in \mathbf{H}^2$,

$$\left[\begin{pmatrix} p \\ Fr \\ r \end{pmatrix}, S_2 \begin{pmatrix} Gx + y \\ Fx \\ x \end{pmatrix}\right] = 0.$$

By Lemma 3.6, for any $x, y \in \mathbf{H}^2$, we have

$$\begin{aligned} 0 &= \left[\begin{pmatrix} p \\ Fr \\ r \end{pmatrix}, S_2 \begin{pmatrix} Gx + y \\ Fx \\ x \end{pmatrix} \right] \\ &= \langle p, s_1 Gx + s_1 y \rangle + \langle Fr, s_2 Fx \rangle - \langle r, s_2 x \rangle \\ &= \langle p, G s_2 x \rangle + \langle Fr, F s_2 x \rangle - \langle r, s_2 x \rangle + \langle p, s_1 y \rangle \\ &= \langle G^* p - F^{(*)} r - r, s_2 x \rangle + \langle p, s_1 y \rangle. \end{aligned}$$

This means that $Gp - F^{(*)}r - r^* = (G^*p - F^{(*)}r - r)^* \in \mathbf{H}^2$ and $p^* \in \mathbf{H}^2$. The theorem follows. □

Two immediate corollaries of Theorem 3.7 are in order.

COROLLARY 3.8.

$$\begin{pmatrix} (1 - F^{(*)})r^* \\ Fp^* \\ p^* \end{pmatrix} \in \mathbf{L}_2 = (S_2 \mathcal{M}_2)^{[\perp]} \cap \mathcal{M}_2$$

if and only if $(1 - F^{(*)})r^* - Gp^*$, $(1 - F^{(*)})(p - Gr) \in \mathbf{H}^2$.

Remark 5. We note that

$$\begin{aligned} &\left[\begin{pmatrix} p \\ Fr \\ r \end{pmatrix}, \begin{pmatrix} (1 - F^{(*)})r^* \\ Fp^* \\ p^* \end{pmatrix} \right] \\ &\quad = \langle p, (1 - F^{(*)})r^* \rangle + \langle Fr, Fp^* \rangle - \langle r, p^* \rangle \\ &\quad = \langle p, r^* \rangle - \langle p, F^{(*)} r^* \rangle + \langle F^{(*)} r, p^* \rangle - \langle r, p^* \rangle \\ &\quad = 0. \end{aligned}$$

If $1 - F^{(*)}$ has spectral factorization $F_s^{(*)}$, i.e., $1 - F^{(*)} = F_s^{(*)}$ with $F_s \in \mathbf{H}^\infty$, then we have the following corollary.

COROLLARY 3.9.

$$\begin{pmatrix} F_s^* r^* \\ Fr \\ r \end{pmatrix} \in \mathbf{L}_2 = (S_2 \mathcal{M}_2)^{[\perp]} \cap \mathcal{M}_2$$

if and only if $F_s^* r^* - Gr \in \mathbf{H}^2$.

Proof. From Theorem 3.7, $(F_s^* r^* \quad Fr \quad r)^\top \in \mathbf{L}_2$ if and only if

$$\begin{aligned} F_s(F_s^* r^* - Gr) = (1 - F^{(*)})r^* - G F_s r &\in \mathbf{H}^2, \\ F_s^* r^* - Gr &\in \mathbf{H}^2. \end{aligned}$$

But $F_s \in \mathbf{H}^\infty$, so we only need $F_s^* r^* - Gr \in \mathbf{H}^2$. □

Remark 6. We note that

$$\begin{aligned} \left[\begin{pmatrix} F_s^* r^* \\ Fr \\ r \end{pmatrix}, \begin{pmatrix} F_s^* r^* \\ Fr \\ r \end{pmatrix} \right] &= \langle F_s^* r^*, F_s^* r^* \rangle + \langle Fr, Fr \rangle - \langle r, r \rangle \\ &= \langle (1 - F^{(*)}) r^*, r^* \rangle - \langle (1 - F^{(*)}) r, r \rangle = 0. \end{aligned}$$

In view of Corollaries 3.8 and 3.9 and Remarks 5 and 6, when we find a base $\{x_1^0, x_2^0\}$ for $\mathbf{L}_2$ such that $[x_1^0, x_1^0] = [x_2^0, x_2^0] = 0$, since $\mathbf{L}_2$ is not neutral, we have $[x_1^0, x_2^0] \neq 0$. We can define

$$x_1 = \frac{1}{2[x_1^0, x_2^0]} x_1^0 + x_2^0, \qquad x_2 = -\frac{1}{2[x_1^0, x_2^0]} x_1^0 + x_2^0;$$

then $\{x_1, x_2\}$ form a base for $\mathbf{L}_2$ and $[x_1, x_1] = 1, [x_2, x_2] = -1$, and $[x_1, x_2] = 0$.

3.1. Special case: optimal pure sensitivity. As a special case, let us look at the problem of minimizing the pure weighted sensitivity in the $\mathbf{H}^\infty$ norm, which is cast as the following Nehari problem [5], [19]:

Given a matrix $G \in \mathbf{L}^\infty$, find all $Z_o \in \mathbf{H}^\infty$ such that

$$\|G - Z_o\|_\infty = \mu_o = \min_{Z \in \mathbf{H}^\infty} \|G - Z\|_\infty .$$

This problem can be treated as a special case of the mixed sensitivity case with $F = 0$. As in the mixed sensitivity case, we may assume without loss of generality that $\mu_o = 1$.

The operator $\mathcal{A}$ becomes a Hankel operator

$$\mathcal{H}_G : \mathbf{H}^2 \to \mathbf{H}^2_- \qquad (\mathcal{H}_G(h) = \Pi_- Gh) .$$

We want to find all $\Phi \in \mathbf{L}^\infty$ such that $\mathcal{H}_G = \mathcal{H}_\Phi$ and $\|\Phi\|_\infty = 1$. Then $Z_o = G - \Phi$. The graph of $\mathcal{H}_G^*$ is

$$\mathcal{G}(\mathcal{H}_G^*) = \left\{ \begin{pmatrix} y \\ \mathcal{H}_G^* y \end{pmatrix} : y \in \mathbf{H}^2_- \right\} .$$

To simplify the notation, we define

$$\mathcal{M}_1 := \mathcal{G}(\mathcal{H}_G^*)^{[\perp]} .$$

From Theorem 3.4 we know that to find the optimal sensitivity is the same as to characterize the set of all S_1-invariant maximal-negative subspaces of $\mathcal{M}_1$. In order to do so, we need to find a base for the subspace [17], [10]

$$\mathbf{L}_1 := \mathcal{M}_1 \cap (S_1 \mathcal{M}_1)^{[\perp]} .$$

Remark 7. Abstract formulas were derived in [17] for a base of $\mathbf{L}_1$. For infinite-dimensional systems, those formulas contain an inverse image of an infinite-dimensional operator, an image of an infinite-dimensional operator, and the norm of an inverse image of the square root of an infinite-dimensional operator. In the finite-dimensional case, an algorithm was given for computing a base of $\mathbf{L}_1$ [10].

By letting $F = 0$ in Theorem 3.7, we have the following corresponding theorem for the pure sensitivity case.

THEOREM 3.10. $\binom{p}{r} \in \mathbf{L}_1$ *if and only if* $p^*, p - Gr, r^* - Gp^* \in \mathbf{H}^2$.

Theorem 3.10 has the following immediate corollaries:

COROLLARY 3.11. $\binom{p}{r} \in \mathbf{L}_1$ *if and only if* $(r^* \quad p^*)^\top \in \mathbf{L}_1$.

COROLLARY 3.12. $(\pm r^* \quad r)^\top \in \mathbf{L}_1$ *if and only if* $r^* \mp Gr \in \mathbf{H}^2$.

Finding a base for the wandering subspaces $\mathbf{L}_1$ or $\mathbf{L}_2$ analytically is in general very difficult. Yang and Orszag [18] proposed a numerical method to find their approximations. Nevertheless, in the next section, we shall show that in some special cases, we can use Theorem 3.10 to construct explicitly a base for the wandering subspaces $\mathbf{L}_1$ and therefore obtain explicit formulas for the optimal sensitivity.

4. Example. In this section, we shall work on the Hardy spaces on the right half plane. Because our results in the previous sections are on the Hardy spaces on the unit disc and $A(z)$ is in $\mathbf{H}^2$ of the unit disc if and only if $A\left(\frac{s-1}{s+1}\right) = (1+s)B(s)$ for some $B(s)$ in $\mathbf{H}^2$ of the right half plane [12, Theorem, p. 130], we find it convenient to still write $A \in \mathbf{H}^2$ when we actually mean $\frac{1}{1+s}A \in \mathbf{H}^2$.

The purpose of this section is to use the results developed in the previous sections to solve the sensitivity minimization problem for $G = m^* \frac{a+s}{b+s}$, where m is an inner function. This corresponds to the problem of pure sensitivity minimization for the plant $P = m$ and weight $W = \frac{a+s}{b+s}$ [5]. We want to find *all* $Z_{\text{o}} \in \mathbf{H}^\infty$ such that

$$\text{(3)} \qquad \|G - Z_{\text{o}}\|_\infty = \mu_{\text{o}} = \inf_{Z \in \mathbf{H}^\infty} \|G - Z\|_\infty = \|\mathcal{H}_G\|,$$

or equivalently, find *all* $\Phi \in \mathbf{L}^\infty$ such that $\mathcal{H}_\Phi = \mathcal{H}_G$ and $\|\Phi\|_\infty = \mu_{\text{o}}$.

Setting $Z = 0$ in (3) we see that

$$\mu_{\text{o}} \le \|G\|_\infty = \sqrt{\sup_{\omega \in \mathbb{R}} \left\{ \frac{a^2 + \omega^2}{b^2 + \omega^2} \right\}} = \begin{cases} \frac{a}{b}, & a > b, \\ 1, & a \le b. \end{cases}$$

It is standard [15], [9] that the essential spectrum of $\mathcal{H}_G^{(*)}$ is

$$\sigma_{\text{e}}(\mathcal{H}_G^{(*)}) = \left\{ \frac{a^2 + \omega^2}{b^2 + \omega^2} : j\omega \in \sigma_{\text{e}}(m) \right\},$$

where $\sigma_{\text{e}}(m)$ denotes the set of imaginary points which are essential singularities of m. Let

$$\text{(4)} \qquad \rho_{\text{e}} = \sup \left\{ \frac{a^2 + \omega^2}{b^2 + \omega^2} : j\omega \in \sigma_{\text{e}}(m) \right\}$$

be the essential spectral radius of $\mathcal{H}_G^{(*)}$. Noting that $\mu_{\text{o}} = \|\mathcal{H}_G\|$, we have $\mu_{\text{o}}^2 = \|\mathcal{H}_G^{(*)}\|$; thus, μ_{o} is given by the norm of a self-adjoint operator, whose spectrum contains only eigenvalues $\sigma_p(\mathcal{H}_G^{(*)})$ and essential spectrum $\sigma_{\text{e}}(\mathcal{H}_G^{(*)})$. We have $\mu_{\text{o}}^2 = \|\mathcal{H}_G^{(*)}\| = \sup\{\sigma_{\text{e}}(\mathcal{H}_G^{(*)}) \cup \sigma_{\text{e}}(\mathcal{H}_G^{(*)})\} \ge \rho_{\text{e}}$.

We shall consider the case when $a \le b$ and m has an essential singularity at infinity. In this case we see that $1 = \rho_{\text{e}} \le \mu_{\text{o}}^2 \le 1$, so $\mu_{\text{o}} = 1$.

Case I: $a = b$. Take $G = \frac{m^*}{\mu}$. For $\mu > 1$, using Theorem 3.10 we see that

$$\begin{cases} p_1 = 1, \\ r_1 = \frac{m}{\mu}; \end{cases} \qquad \begin{cases} p_2 = \frac{m^*}{\mu}, \\ r_2 = 1 \end{cases}$$

is a base for $\mathbf{L}_1$ and

$$\left[\begin{pmatrix} p_1 \\ r_1 \end{pmatrix}, \begin{pmatrix} p_1 \\ r_1 \end{pmatrix} \right] = \frac{\mu^2 - 1}{\mu^2} > 0, \quad \left[\begin{pmatrix} p_2 \\ r_2 \end{pmatrix}, \begin{pmatrix} p_2 \\ r_2 \end{pmatrix} \right] = \frac{1 - \mu^2}{\mu^2} < 0.$$

Let $x_1 = \frac{\mu}{\sqrt{\mu^2 - 1}}(p_1 \quad r_1)^{\mathrm{T}}$, $x_2 = \frac{\mu}{\sqrt{\mu^2-1}}(p_2 \quad r_2)^{\mathrm{T}}$. Then $[x_1, x_1] = 1$, $[x_2, x_2] = -1$, and $[x_1, x_2] = 0$. It is easy to see that

$$\frac{\mu}{\sqrt{\mu^2 - 1}} r_2 \to \infty$$

as $\mu \to 1$. Thus in this case the minimal symbol is unique. It is easy to see that the unique symbol is $\Phi = m^*$.

Case II: $a < b$. Take $G = \frac{m^*}{\mu}\frac{a+s}{b+s}$. For $\binom{p}{r} \in \mathbf{L}_1$, the trick is to assume that

$$p = P_1 m^* + P_2, \quad r = R_1 m + R_2,$$

where P_1, P_2, R_1, R_2 are rational functions to be computed later.

We have

$$p - G_\mu r = m^* \left(P_1 - \frac{1}{\mu}\frac{a+s}{b+s} R_2 \right) + P_2 - \frac{1}{\mu}\frac{a+s}{b+s} R_1,$$

$$r^* - G_\mu p^* = m^* \left(R_1^* - \frac{1}{\mu}\frac{a+s}{b+s} P_2^* \right) + R_2^* - \frac{1}{\mu}\frac{a+s}{b+s} P_1^*.$$

By Theorem 3.10 , in order for $\binom{p}{r} \in \mathbf{L}_1$, we require that $p - G_\mu r \in \mathbf{H}^2$, $r^* - G_\mu p^* \in \mathbf{H}^2$, $p^* \in \mathbf{H}^2$, and $r \in \mathbf{H}^2$. Using these constraints we see that

$$\begin{cases} P_1 - \frac{1}{\mu}\frac{a+s}{b+s} R_2 & = & 0, \\ P_2 - \frac{1}{\mu}\frac{a+s}{b+s} R_1 & \in & \mathbf{H}^2; \end{cases} \quad \begin{cases} R_1^* - \frac{1}{\mu}\frac{a+s}{b+s} P_2^* & = & 0, \\ R_2^* - \frac{1}{\mu}\frac{a+s}{b+s} P_1^* & \in & \mathbf{H}^2. \end{cases}$$

These conditions in turn yield

$$(5) \qquad \begin{cases} P_2 \left(1 - \frac{1}{\mu^2}\frac{a^2 - s^2}{b^2 - s^2} \right) & \in & \mathbf{H}^2, \\ R_2^* \left(1 - \frac{1}{\mu^2}\frac{a^2 - s^2}{b^2 - s^2} \right) & \in & \mathbf{H}^2. \end{cases}$$

For $\mu > 1$, define $B_\mu^2 := \mu^2 b^2 - a^2 > 0$. Then condition (5) becomes

$$\begin{cases} P_2 \frac{B_\mu^2 - (\mu^2 - 1)s^2}{\mu^2 (b^2 - s^2)} & = & P_2 \frac{(B_\mu - \sqrt{\mu^2 - 1s})(B_\mu + \sqrt{\mu^2 - 1s})}{\mu^2 (b^2 - s^2)} & \in \mathbf{H}^2, \\ R_2^* \frac{B_\mu^2 - (\mu^2 - 1)s^2}{\mu^2 (b^2 - s^2)} & = & R_2^* \frac{(B_\mu - \sqrt{\mu^2 - 1s})(B_\mu + \sqrt{\mu^2 - 1s})}{\mu^2 (b^2 - s^2)} & \in \mathbf{H}^2. \end{cases}$$

Since we require that $p^* \in \mathbf{H}^2$ and $r \in \mathbf{H}^2$, we see that we have the following solutions for $\binom{p}{r}$:

$$\begin{cases} p_1 & = & \alpha_1 \frac{1}{\mu} \frac{a+s}{B_\mu - \sqrt{\mu^2 - 1s}} m^* + \beta_1 \frac{b-s}{B_\mu - \sqrt{\mu^2 - 1s}}, \\ r_1 & = & \beta_1 \frac{1}{\mu} \frac{a-s}{B_\mu - \sqrt{\mu^2 - 1s}} m + \alpha_1 \frac{b+s}{B_\mu - \sqrt{\mu^2 - 1s}}; \end{cases} \quad \begin{bmatrix} P_2 & = & \beta_1 \frac{b-s}{B_\mu - \sqrt{\mu^2 - 1s}} \\ R_2^* & = & \bar{\alpha}_1 \frac{b-s}{B_\mu + \sqrt{\mu^2 - 1s}} \end{bmatrix};$$

$$\begin{cases} p_2 & = & \alpha_2 \frac{1}{\mu} \frac{a+s}{B_\mu + \sqrt{\mu^2 - 1s}} m^* + \beta_2 \frac{b-s}{B_\mu + \sqrt{\mu^2 - 1s}}, \\ r_2 & = & \beta_2 \frac{1}{\mu} \frac{a-s}{B_\mu + \sqrt{\mu^2 - 1s}} m + \alpha_2 \frac{b+s}{B_\mu + \sqrt{\mu^2 - 1s}}; \end{cases} \quad \begin{bmatrix} P_2 & = & \beta_2 \frac{b-s}{B_\mu + \sqrt{\mu^2 - 1s}} \\ R_2^* & = & \bar{\alpha}_2 \frac{b-s}{B_\mu - \sqrt{\mu^2 - 1s}} \end{bmatrix}$$

where the coefficients $\alpha_1, \beta_1, \alpha_2, \beta_2$ satisfy the following interpolation conditions:

$$(6) \qquad \beta_1 (a - C_\mu) m (C_\mu) + \alpha_1 \mu (b + C_\mu) = 0,$$

$$(7) \qquad \alpha_2 (a - C_\mu) m^* (-C_\mu) + \beta_2 \mu (b + C_\mu) = 0,$$

where $C_\mu = \frac{B_\mu}{\sqrt{\mu^2 - 1}}$.

By Theorem 3.10,

$$\begin{pmatrix} p_1 \\ r_1 \end{pmatrix} \in \mathbf{L}_1, \qquad \begin{pmatrix} p_2 \\ r_2 \end{pmatrix} \in \mathbf{L}_1.$$

Since $\dim \mathbf{L}_1 = 2$, $(p_1 \quad r_1)^\top$ and $(p_2 \quad r_2)^\top$ form a base for $\mathbf{L}_1$.

Noting that C_μ is real for $\mu > 1$, we have $m^*(-C_\mu) = \bar{m}(C_\mu)$. From (6) and (7) we can let $\beta_1 = \bar{\alpha}_2$ and $\alpha_1 = \bar{\beta}_2$. This gives us

$$\begin{cases} p_1^* = r_2, \\ r_1^* = p_2. \end{cases}$$

Thus $[(p_1 \quad r_1)^\top, (p_2 \quad r_2)^\top] = \langle p_1, p_2\rangle - \langle r_1, r_2\rangle = \langle 1, p_1^* p_2 - r_1^* r_2\rangle = 0$.

Letting $\beta_1 = \alpha_2 = 1$, we have

$$\begin{cases} p_1 = \frac{1}{\mu\sqrt{\mu^2-1}(C_\mu - s)}\left[\frac{F_\mu m(C_\mu)}{\mu}(a+s)m^* + \mu(b-s)\right], \\ r_1 = \frac{1}{\mu\sqrt{\mu^2-1}(C_\mu - s)}\left[(a-s)m + F_\mu m(C_\mu)(b+s)\right]; \end{cases}$$

$$\begin{cases} p_2 = \frac{1}{\mu\sqrt{\mu^2-1}(C_\mu + s)}\left[(a+s)m^* + F_\mu m^*(-C_\mu)(b-s)\right], \\ r_2 = \frac{1}{\mu\sqrt{\mu^2-1}(C_\mu + s)}\left[\frac{F_\mu m^*(-C_\mu)}{\mu}(a-s)m + \mu(b+s)\right], \end{cases}$$

where $F_\mu = \frac{C_\mu - a}{C_\mu + b}$.

One can compute that

$$p_1 p_1^* - r_1 r_1^* = \frac{\mu^2 - F_\mu^2 |m(C_\mu)|^2}{\mu^4}.$$

Thus

$$\begin{aligned}\left[\begin{pmatrix} p_1 \\ r_1 \end{pmatrix}, \begin{pmatrix} p_1 \\ r_1 \end{pmatrix}\right] &= \langle p_1, p_1\rangle - \langle r_1, r_1\rangle \\ &= \langle p_1 p_1^* - r_1 r_1^*, 1\rangle \\ &= \frac{\mu^2 - F_\mu^2 |m(C_\mu)|^2}{\mu^4} > 0.\end{aligned}$$

Define

$$x_1 = \frac{\mu^2}{\sqrt{\mu^2 - F_\mu^2 |m(C_\mu)|^2}}\begin{pmatrix} p_1 \\ r_1 \end{pmatrix} \quad \text{and} \quad x_2 = \frac{\mu^2}{\sqrt{\mu^2 - F_\mu^2 |m(C_\mu)|^2}}\begin{pmatrix} r_1^* \\ p_1^* \end{pmatrix}.$$

Then $[x_1, x_1] = 1$, $[x_2, x_2] = -1$, and $[x_1, x_2] = 0$.

Next let us look at

$$S_\mu(s) := \frac{\mu^2}{\sqrt{\mu^2 - F_\mu^2 |m(C_\mu)|^2}} r_2 = \frac{F_\mu m^*(-C_\mu)(a-s)m + \mu^2(b+s)}{\sqrt{\mu^2 - F_\mu^2 |m(C_\mu)|^2}\,\mu(B_\mu + \sqrt{\mu^2-1}s)}.$$

For any real number sequence $\{y_n\}$ such that $y_n \to \infty$, if $m(y_n) \to M$, then $|M| \le 1$ since $|m(y_n)| \le 1$. We shall further see that $|M| < 1$. In fact, if there exists $y_n \to \infty$ and $m(y_n) \to M$ and $|M| = 1$, then for any $s \in \mathbb{C}^+$,

$$\sup_{\mu > 1} |S_\mu(s)| = +\infty;$$

we should have a unique optimal sensitivity (minimal symbol). We see that

$$\Phi_1 = \frac{(b-s)\bar{M} + (a+s)m^*}{(a-s)m\bar{M} + (b+s)},$$

being a cluster point of $(\Phi_\mu)_{\mu>1}$ relative to the weak-star topology of $\mathbf{L}^\infty$, should be a minimal symbol for $\mathcal{H}_G$; but since $a \neq b$,

$$\Phi_2 = \frac{a+s}{b+s} m^*$$

is clearly another minimal symbol for $\mathcal{H}_G$, which contradicts the assertion that $\mathcal{H}_G$ has a unique minimal symbol.

We now have $|M| < 1$. Then

$$\sup_{\mu>1} |S_\mu(1)| < +\infty$$

and we have nonunique optimal sensitivities (minimal symbols). Since $|m(s)| \leq 1$ for $s \in \mathbb{C}^+$, there exists a sequence $\{y_n\}$ such that $m(y_n) \to M$ and from the above discussion we know that $|M| < 1$. Let $\mu \to 1$; we get the parameterization of the minimal symbols as

$$\Phi = \frac{[M(a+s)m^* + (b-s)]\,\phi_1 + (a+s)m^* + \bar{M}(b-s)}{[(a-s)m + M(b+s)]\,\phi_1 + \bar{M}(a-s)m + (b+s)} \tag{8}$$

for any $\phi_1 \in B_{\mathbf{H}^\infty}$.

Since

$$\phi = \mathcal{F}(\phi_1) := \frac{\phi_1 + \bar{M}}{1 + M\phi_1}$$

is a bijection from $B_{\mathbf{H}^\infty}$ to $B_{\mathbf{H}^\infty}$ for $|M| < 1$, we see from (8) that

$$\Phi = \frac{(b-s)\phi + (a+s)m^*}{(a-s)m\phi + (b+s)} = m^* \frac{(b-s)m\phi + (a+s)}{(a-s)m\phi + (b+s)}$$

for any $\phi \in B_{\mathbf{H}^\infty}$.

We summarize the results in the following theorem.

THEOREM 4.1. *Let $G = m^* \frac{a+s}{b+s}$ with $b \geq a \geq 0$ and m an inner function with an essential singularity at infinity. Then $\mu_o = 1$ and the $\mathbf{H}^\infty$ optimal sensitivity is parameterized by the formula*

$$\Phi = m^* \frac{(b-s)m\phi + (a+s)}{(a-s)m\phi + (b+s)}, \quad \phi \in B_{\mathbf{H}^\infty}.$$

Remark 8. The above formula is a generalization of a result in [3] that shows that for a pure delay system e^{-ds}, i.e., $m = e^{-ds}$, with weight $W = \frac{a+s}{b+s}$, $a < b$, the optimal pure weighted sensitivity can be parameterized as

$$\Phi = e^{ds} \frac{(b-s)e^{-ds}\phi + (a+s)}{(a-s)e^{-ds}\phi + (b+s)}$$

for any $\phi \in B_{\mathbf{H}^\infty}$.

We remark that for other cases such as the inner function m that has finite essential singularities on the imaginary axis or $a > b$, the optimal sensitivity can also be computed by the present method.

5. Conclusions. Based on some well-known results, we have considered the explicit computation of $\mathbf{H}^\infty$ optimal pure and mixed sensitivities for a class of infinite-dimensional systems by studying certain subspaces of some related Kreĭn spaces. We have given some new characterizations for those subspaces. The pure sensitivity case can be treated as a special case of the mixed sensitivity case.

Our new characterizations are particularly useful both from theoretical and practical points of view. They can be used in the theoretical analysis of the optimal compensators as well as convergence analysis for numerical algorithms since from these new characterizations, explicit $\mathbf{H}^\infty$ optimal weighted pure and mixed sensitivities for a class of infinite-dimensional systems can be obtained as we showed in a simple example that generalized a known result. Practically, they can be used to obtain fast and stable algorithms for computing the $\mathbf{H}^\infty$ optimal performances [18]. In addition, we believe that they can also be used to develop numerically efficient algorithms for computing the suboptimal compensators and spectral factorizations for both irrational and rational systems and that the results can be extended to treat the multivariable systems. An added advantage is that the mathematics used in the paper is no more advanced than that used in the analysis of rational systems [10]. One area of further research is to obtain explicit optimal sensitivity formulas for more general systems and weights.

Acknowledgments. The author thanks the anonymous referees for their valuable suggestions and correction of errors. The proof of Theorem 3.7 was greatly simplified by one of the referees.

REFERENCES

[1] J. A. Ball and J. W. Helton, *A Beurling-Lax theorem for the Lie group $u(m,n)$ which contains most classical interpolation theory*, J. Operator Theory, 9 (1983), pp. 107–142.

[2] J. Bognár, *Indefinite Inner Product Spaces*, Springer-Verlag, New York, 1974.

[3] F. Fagnani, *Problemi di minimizzesione in $\mathbf{H}^\infty$ per sistemi in dimenzione infinita.* Bachelor's Thesis, Universitá degli Studi di Pisa, October 1986.

[4] ———, *An operator-theoretic approach to the mixed-sensitivity minimization problem*, Systems Control Lett., 17 (1991), pp. 227–235.

[5] D. S. Flamm, *Control of delay systems for minimax sensitivity*, Tech. Report LIDS-TH-1560, Massachusetts Institute of Technology Laboratory for Information and Decision Systems, 1986.

[6] D. S. Flamm and H. Yang, *$\mathbf{H}^\infty$-optimal mixed sensitivity for general distributed plants*, in Proc. IEEE Conf. on Decision and Control 1, Honolulu, HI, December 1990, pp. 134–139.

[7] C. Foias, *Commutant lifting techniques for computing optimal $\mathbf{H}^\infty$ controllers*, in Lecture Notes in Math. 1496, E. Mosca and L. Pandolfi, eds., Springer-Verlag, Berlin, New York, 1991, pp. 1–36.

[8] C. Foias and A. Tannenbaum, *On the Nehari problem for a certain class of $\mathbf{L}^\infty$ functions appearing in control theory, II*, J. Funct. Anal., 81 (1988), pp. 207–218.

[9] C. Foias, A. Tannenbaum, and G. Zames, *Some explicit formulae for the singular values of certain Hankel operators with factorizable symbol*, SIAM J. Math. Anal., 19 (1988), pp. 1081–1089.

[10] B. Francis, J. W. Helton, and G. Zames, *$\mathbf{H}^\infty$-optimal feedback controllers for linear multivariable systems*, IEEE Trans. Automatic Control, AC-29 (1984), pp. 888–900.

[11] C. Gu, *Eliminating the genericity conditions in the skew Toeplitz operator algorithm for $\mathbf{H}^\infty$ optimization*, SIAM J. Math. Anal., 23 (1992), pp. 1623–1636.

[12] K. Hoffman, *Banach Spaces of Analytic Functions*, Prentice-Hall, Englewood Cliffs, NJ, 1962.

[13] E. Jonckheere and M. Verma, *A spectral characterization of $\mathbf{H}^\infty$-optimal feedback performance and its efficient computation*, Systems Control Lett., 8 (1986), pp. 13–22.

[14] P. P. Khargonekar, H. Özbay, and A. Tannenbaum, *Four-block problem: stable plants and rational weights*, Internat. J. Control, 50 (1989), pp. 1013–1023.

[15] N. K. Nikol'skiĭ, *Treatise on the Shift Operator*, Springer-Verlag, Berlin and New York, 1986.

[16] D. Sarason, *Generalized interpolation in* $\mathbf{H}^\infty$, Trans. Amer. Math. Soc., 127 (1967), pp. 179–203.

[17] ———, *Operator-theoretic aspects of the Nevanlinna-Pick interpolation problem*, in Operator and Function Theory, S. C. Power, ed., D. Reidel, Dordrecht, the Netherlands, 1985, pp. 279–314.

[18] H. Yang and J. M. Orszag, *Numerical computation of* H^∞ *optimal performance*, J. Sci. Comput., 7 (1992), pp. 289–311.

[19] G. Zames, A. Tannenbaum, and C. Foias, *Optimal* $\mathbf{H}^\infty$ *interpolation: a new approach*, in Modelling, Robustness and Sensitivity Reduction, R. F. Curtain, ed., Groningen, Proc. NATO Adv. Research Workshop, December 1986, pp. 381–398.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 33, No. 2, pp. 365–401, March 1995

© 1995 Society for Industrial and Applied Mathematics

003

ZEROS OF SPECTRAL FACTORS, THE GEOMETRY OF SPLITTING SUBSPACES, AND THE ALGEBRAIC RICCATI INEQUALITY*

ANDERS LINDQUIST[†], GYÖRGY MICHALETZKY[‡], AND GIORGIO PICCI[§]

Abstract. In this paper it is shown how the zero dynamics of (not necessarily square) spectral factors relate to the splitting subspace geometry of stationary stochastic models and to the corresponding algebraic Riccati inequality. The notion of *output-induced subspace* of a minimal Markovian splitting subspace, which is the stochastic analogue of the *supremal output-nulling subspace* in geometric control theory, is introduced. Through this concept, the analysis can be made coordinate-free and straightforward geometric methods can be applied. It is shown how the zero structure of the family of spectral factors relates to the geometric structure of the family of minimal Markovian splitting subspaces in the sense that the relationship between the zeros of different spectral factors is reflected in the partial ordering of minimal splitting subspaces. Finally, the well-known characterization of the solutions of the algebraic Riccati equation is generalized in terms of Lagrangian subspaces invariant under the corresponding Hamiltonian to the larger solution set of the algebraic Riccati inequality.

Key words. zero dynamics, Markovian splitting subspaces, minimal spectral factors, matrix Riccati inequality, algebraic Riccati equation, geometric control theory

AMS subject classifications. 93E03, 93B27, 60G10

1. Introduction.

By now it should be fairly well known that there is a one-to-one correspondence between each pair of the following three fundamental areas of systems theory.

(i) *Minimal spectral factorization* of a rational (full-rank) $m \times m$ spectral density matrix Φ. The problem is to find *all* (square *and* rectangular) rational functions

$$W(s) = C\,(sI - A)^{-1}\,B + D \tag{1.1}$$

(where prime denotes transposition) with poles in the open left half plane, satisfying the factorization equation

$$W(s)W(-s)' = \Phi(s) \tag{1.2}$$

and being *minimal* in the sense that the McMillan degree of W is exactly half of that of Φ. The class of all such minimal spectral factors, each defined modulo right multiplication by a constant orthogonal matrix, will be denoted by $\mathcal{W}$. The subclass of *square* spectral factors will be denoted $\mathcal{W}_0$. Throughout this paper we shall always consider representations for which $(A,\ B,\ C)$ is a minimal triplet and $[\begin{smallmatrix}B\\D\end{smallmatrix}]$ has independent columns. This results in no loss of generality [16].

(ii) Finding all symmetric solutions of the *algebraic Riccati inequality*

$$\Lambda(P) \leq 0, \tag{1.3}$$

* Received by the editors October 15, 1992; accepted for publication (in revised form) October 19, 1993. This work was partially supported by the Swedish Board for Technical Development, the Göran Gustafsson Foundation, LADSEB-CNR, and OTKA grant 2042.

[†] Division of Optimization and Systems Theory, Royal Institute of Technology, 100 44 Stockholm, Sweden (alq@math.kth.se).

[‡] Department of Probability Theory and Statistics, Eötvös Loránd University, Budapest, Hungary (micha@cs.elte.hu).

[§] Department of Electronics and Informatics, University of Padova and LADSEB-CNR, Padova, Italy (picci@paola.dei.unipd.it).

where $\Lambda : \mathbb{R}^{n\times n} \longrightarrow \mathbb{R}^{n\times n}$ is given by

$$\Lambda(P) = AP + PA' + (\bar{C} - CP)'R^{-1}(\bar{C} - CP), \tag{1.4}$$

the matrices $A \in \mathbb{R}^{n\times n}, C, \bar{C} \in \mathbb{R}^{m\times n}$, and $R \in \mathbb{R}^{m\times m}$ being defined through a minimal realization

$$\Phi_+(s) = C(sI - A)^{-1}\bar{C}' + \frac{1}{2}R \tag{1.5}$$

of the positive real part Φ_+ of the spectral density Φ, i.e., the rational matrix function Φ_+ satisfying

$$\Phi(s) = \Phi_+(s) + \Phi_+(-s)' \tag{1.6}$$

and having all its poles in the open left half plane. Here we assume that $R := \Phi(\infty) > 0$.

Let us denote by $\mathcal{P}$ the solution set of (1.3). Then each $P \in \mathcal{P}$ corresponds to a spectral factor (1.1) whose B- and D-matrices are determined by a full-rank matrix factorization of the type

$$\begin{bmatrix} B \\ D \end{bmatrix} [B', D'] = \begin{bmatrix} -AP - PA' & \bar{C}' - PC' \\ \bar{C} - CP & R \end{bmatrix}. \tag{1.7}$$

Obviously the correspondence is one-to-one modulo trivial coordinate transformations [1], [9].

(iii) Describing all *minimal stochastic realizations* of an m-dimensional stationary-increments process $\{y(t); t \in \mathbb{R}\}$ having the (incremental) spectral density Φ. Each stochastic realization is obtained by passing a suitable "white noise" through a filter

$$\xrightarrow{dw} \boxed{W} \xrightarrow{dy} \tag{1.8}$$

having an $m \times p$ minimal spectral factor as its transfer function, thus yielding a linear dynamical model

$$(\Sigma) \quad \begin{cases} dx = Axdt + Bdw, \\ dy = Cxdt + Ddw \end{cases} \tag{1.9}$$

for dy, defined on the whole real line. More precisely, w is a vector Wiener process on $\mathbb{R}$ of a dimension p equal to the number of columns of W. The system Σ is in statistical steady state so that the n-dimensional state process x and the increments of the m-dimensional output process y are jointly stationary. The model Σ is a *minimal* stochastic realization in the sense that there is no other representation of dy of type (1.9) with a state process with fewer components.

In regard to topic (iii), it is actually more natural to consider a coordinate-free representation by assigning to each model Σ the n-dimensional space

$$X = \{a'x(0)\,|a \in \mathbb{R}^n\} \tag{1.10}$$

of random variables. This space is the subspace of an *ambient space* H of the model (1.9), defined as the closure of the linear hull of the following random variables $\{w_i(t) - w_i(\tau);\ i = 1, 2, \dots,\ p;\ t, \tau \in \mathbb{R}\}$ in the topology of the inner product

$$\langle \xi, \eta \rangle = E\{\xi\eta\}, \tag{1.11}$$

where $E\{\cdot\}$ stands for mathematical expectation. The ambient space H is naturally equipped with the shift induced by dw, i.e., the strongly continuous group of unitary operators $\{U_t;\ t \in \mathbb{R}\}$ on H such that $U_t\left[w_i(\tau) - w_i(\sigma)\right] = w_i(\tau + t) - w_i(\sigma + t)$ for all $i = 1, 2, \ldots, p$ and $t, \tau, \sigma \in \mathbb{R}$. All random variables of Σ belong to H, and moreover the processes x and dy are stationary with respect to $\{U_t\}$, i.e., $U_t x_i(\tau) = x_i(\tau + t)$ for all $i = 1, 2, \ldots, n$ and $t, \tau \in \mathbb{R}$ and $U_t\left[y_i(\tau) - y_i(\sigma)\right] = y_i(\tau + t) - y_i(\sigma + t)$ for all $i = 1, 2, \ldots, m$ and $t, \tau, \sigma \in \mathbb{R}$. Minimality of Σ corresponds to minimality of the subspace X in the sense of subspace inclusion and, hence, also in the sense of dimension [16].

Defining the *past* and *future output spaces* as

$$H^- = \text{closure}\left\{a'\left[y(t) - y(s)\right] \middle| a \in \mathbb{R}^m, t, s \leq 0\right\}$$

and

$$H^+ = \text{closure}\left\{a'\left[y(t) - y(s)\right] \middle| a \in \mathbb{R}^m, t, s \geq 0\right\}$$

respectively, it is easy to show and well established in the literature [15], [16], [6] that each X, defined as in (1.10), is a *minimal Markovian splitting subspace* for H^- and H^+, i.e., in particular renders H^- and H^+ conditionally orthogonal given X. Moreover, this property captures the concept of stochastic state space model of dy in a coordinate-free way. Given any X together with its ambient space H, equipped with a shift, we can construct the model Σ modulo the choice of coordinates in the state space [16].

Modulo coordinate-transformations, there is a one-to-one correspondence between the family $\mathcal{X}$ of minimal Markovian splitting subspaces and the solution set $\mathcal{P}$ of the algebraic Riccati inequality (1.3) under which

$$P = E\left\{x(0)x(0)'\right\} \tag{1.12}$$

is the state covariance. Under this correspondence the subset $\mathcal{P}_0 \subset \mathcal{P}$ of solutions of the *algebraic Riccati equation*

$$\Lambda(P) = 0 \tag{1.13}$$

corresponds to the subclass $\mathcal{X}_0 \subset \mathcal{X}$ of stochastic realizations such that [1]

$$X \subset H_0 := H^- \vee H^+, \tag{1.14}$$

i.e., *internal* realizations constructed by using only random quantities contained in the subspace

$$H_0 = \text{closure}\{a'[y(t) - y(s)] | a \in \mathbb{R}^n\}$$

spanned by the output. Under the correspondence mentioned above, $\mathcal{X}_0$ and $\mathcal{P}_0$ correspond to $\mathcal{W}_0 \subset \mathcal{W}$, the subclass of square spectral factors.

Although the structure of the solution set of the algebraic Riccati equation (1.13) is by now fairly well established [27], [20], [26], [13], it is fair to say that the structure of the complete solution set $\mathcal{P}$ of the algebraic Riccati inequality (1.3) is far less

[1] In the upcoming sections, given two subspaces A and B, we shall write $A \vee B$ to denote the closure of $\{\alpha + \beta | \alpha \in A, \beta \in B\}$. To stress that the sum is direct we write instead $A + B$ or, if it is an orthogonal direct sum, $A \oplus B$.

understood; and, except for [10], [16], and [25], little seems to have appeared in the literature since the monograph [9]. We stress that the algebraic Riccati inequality and the set $\mathcal{P}$ are important in many areas of systems and control, including dissipative systems and H^∞ control.

In this respect, one purpose of this paper is to provide new results on the structure of $\mathcal{P}$ and new concepts for the study and classification of this set based on the *zero structure* of the family $\mathcal{W}$ of minimal spectral factors W. The work reported here is a continuation and a deepening of the results presented in [16] and [19]. In particular it was shown in [16] that

1. The set $\mathcal{P}$ (which is bounded and convex) has *facets*, each of which is uniquely defined by a pair of solutions of the algebraic Riccati equation (1.13). For each $P \in \mathcal{P}$ there is a minimal facet $[P_{0-}, P_{0+}]$ containing P, called the *tightest local frame* of P, defined as the set of all solutions Q of the algebraic Riccati inequality (1.3) satisfying the relation $P_{0-} \leq Q \leq P_{0+}$, where

$$P_{0-} := \sup \{P_0 \in \mathcal{P}_0 \mid P_0 \leq P\},$$

$$P_{0+} := \inf \{P_0 \in \mathcal{P}_0 \mid P \leq P_0\}.$$

Here, for any $P_1, P_2 \in \mathcal{P}$, $P_1 \leq P_2$ means that $P_2 - P_1$ is nonnegative definite. The tightest bounds of P, i.e., P_{0-} and P_{0+}, can be computed as the limit solutions of the matrix Riccati differential equation $\dot{\Pi} = \Lambda(\Pi)$, with initial condition $\Pi(0) = P$, as t tends to $-\infty$ and ∞ respectively.

2. The *open tightest frame* (P_{0-}, P_{0+}) of $P \in \mathcal{P}$, consisting of all $Q \in [P_{0-}, P_{0+}]$ having P_{0-} and P_{0+} as tight bounds, can be characterized in terms of the *zeros* of the corresponding minimal spectral factor W. If (W_{0-}, W_{0+}) is the pair of square minimal spectral factors corresponding to P_{0-} and P_{0+}, then the zeros of W are precisely the common zeros of W_{0-} and W_{0+}.

In this paper we greatly expand on the above characterization of facets and tight frames providing necessary and sufficient conditions in terms of zeros (or, better, the *zero dynamics*) of spectral factors. To this end, in §2, we first provide a geometric characterization of the zero dynamics in the stochastic framework (Theorem 2.9). In particular, we demonstrate how the zero structure of each W can be recovered directly from the corresponding *output-induced subspace* $X \cap H_0$ and a related compressed shift. We introduce a dual control problem and show that its maximal output-nulling subspace consists of precisely those $a \in \mathbb{R}^n$ for which $a'x(0) \in X \cap H_0$ and that these a are also the zero directions of W. In this way we not only provide the appropriate connection to geometric control theory [3], [28] but also obtain elegant coordinate-free proofs of the main theorems of §§2 and 3.

Next, in §3, we analyze the relation between partial ordering of minimal splitting subspaces and zeros and characterize the ordering in terms of invariant subspaces for the zero dynamics and right half-plane zeros. The results on ordering are very intuitive and are in agreement with some early observations of Anderson [2] and Robinson [23]. The characterizations in terms of invariant subspaces extend those known thus far for square spectral factors and the algebraic Riccati equation, as for example reported in the survey of Kucera [13].

Thus far all results are coordinate-free. Then, in §4, we introduce coordinates and translate the geometric characterizations of §§2 and 3 in terms of covariances and solutions of the algebraic Riccati inequality. Through this analysis we also obtain a natural generalization of the well-known characterization (Potter [22], MacFarlane

[17]; also see [26]) of $\mathcal{P}$ in terms of the n-dimensional Lagrangian subspaces $\mathcal{L} \subset \mathbb{R}^{2n}$, invariant under multiplication by the Hamiltonian $\mathcal{H}$ corresponding to Φ. In fact, in §5, we show that the $\mathcal{H}$-invariant isotropic subspaces $\mathcal{L}$ of dimension $k \leq n$ are in one-to-one correspondence with the facets of $\mathcal{P}$ whose elements P have identical zero structure. Under this correspondence

$$\mathcal{L} = \begin{bmatrix} I \\ P \end{bmatrix} \mathcal{V}^*, \tag{1.15}$$

where $\mathcal{V}^* \subset \mathbb{R}^n$ is the space on which the zero dynamics of W is defined and which corresponds in $\mathcal{X}$ to the output-induced subspace $X \cap H_0$ of X.

We make extensive cross reference between the three frameworks of $\mathcal{P}, \mathcal{X}$, and $\mathcal{W}$; and there are some very good reasons for this. The geometric splitting subspace theory provides a very natural setting also for analyzing the algebraic Riccati inequality. In fact, several geometric results that are linked to such concepts as *splitting* and *internal subspace* have less obvious counterparts in the $\mathcal{P}$-setting and could easily have been overlooked had it not been for the interaction with the geometry of splitting subspaces.

2. Zero dynamics and splitting subspaces. It is well known by now that the poles of a spectral factor W can be expressed in terms of the shift $\{U_t\}$ and the corresponding splitting subspace X [16]. In fact, the compressed forward shift on X,

$$U_t(X) := E^X U_t \,|_X \quad \text{for} \quad t \geq 0 \tag{2.1}$$

(where E^X is the orthogonal projector onto X), is a strongly continuous and uniformly bounded semigroup so that

$$U_{t+\tau}(X) = U_t(X) U_\tau(X); \tag{2.2}$$

and therefore there is an operator $F : X \to X$ such that

$$U_t(X) = e^{Ft}. \tag{2.3}$$

Then it can be shown that

$$\{\text{poles of } W\} = \sigma(F), \tag{2.4}$$

i.e., the poles of W are precisely the eigenvalues of F. To see this, take $a \in \mathbb{R}^n$ and integrate (1.9) to obtain

$$a'x(t) = a'e^{At}x(0) + \int_0^t a'e^{A(t-s)}B\,dw(s), \tag{2.5}$$

the last term of which is orthogonal to X. Consequently, $E^X U_t a'x(0) = a'e^{At}x(0)$, i.e., $e^{Ft}a'x(0) = a'e^{At}x(0)$, showing that A' is in fact a matrix representation of F.

The basic question that we shall address in this section is the following. Is there an analogous geometric characterization of the *zeros* of W in terms of $\{U_t\}$ and X? As we shall see, the answer to this question is yes.

To simplify matters, in this paper we shall make the blanket assumption that the spectral density Φ is *coercive*, i.e., Φ has no zeros on the imaginary axis or at infinity. In particular this implies that

$$R := \Phi(\infty) > 0 \tag{2.6}$$

so that all minimal spectral factors W are of dimension $p \times m$ with $p \geq m$ and of full rank m almost everywhere in the complex plane and, hence, right invertible. Let

$$\left\{ \begin{array}{l} \dot{x} = Ax + Bu, \\ y = Cx + Du \end{array} \right. \tag{2.7}$$

be a minimal realization of W. Recall [7] that a complex number λ is called a *right zero* of W (or, equivalently, of the state-space system (2.7)) if, for some $\left[\begin{smallmatrix} x_0 \\ u_0 \end{smallmatrix}\right] \neq 0$, $u(t) = u_0 e^{\lambda t}$, $x(t) = x_0 e^{\lambda t}$ satisfy (2.7) while at the same time keeping the output $y(t)$ (identically) zero for all $t \in \mathbb{R}$.

It is well known and trivial to check that $\lambda \in \mathbb{C}$ is a right zero of W if and only if there are nonzero solutions of [2]

$$\begin{bmatrix} A - \lambda I & B \\ C & D \end{bmatrix} \begin{bmatrix} x_0 \\ u_0 \end{bmatrix} = 0. \tag{2.8}$$

More generally it can be shown [28] that constraining the dynamic variables x and u in (2.7) to yield an identically zero output $y \equiv 0$ requires confining, for all times $t \in \mathbb{R}$, the state $x(t)$ of the system (2.7) to a particular subspace $\mathcal{V}^* = \mathcal{V}^*(A, B, C, D) \subset \mathbb{R}^n$ called the *maximal output nulling subspace* of the system (2.7). The inputs u that keep $x(t)$ in $\mathcal{V}^*$ for all $t \in \mathbb{R}$ can be generated by suitable linear state feedback laws

$$u = Kx + Lv, \quad x \in \mathcal{V}^*, \tag{2.9}$$

where L is such that $\operatorname{Im} BL \subset \mathcal{V}^*$, $DL = 0$ and v is an unconstrained input function. Any K achieving this is called a *friend* of $\mathcal{V}^*$ [28]. It can be shown that $\mathcal{V}^*$ is actually the largest subspace $\mathcal{V} \subset \mathbb{R}^n$ for which there is a feedback matrix K such that

$$(A + BK)\mathcal{V} \subset \mathcal{V} \subset \ker(C + DK). \tag{2.10}$$

It follows from the discussion above that all x_0 solving (2.8) belong to $\mathcal{V}^*(A, B, C, D)$. Conversely, the subspace $\mathcal{V}^*$ can be associated to the right zeros of (2.7) in the following sense. If K is a friend of $\mathcal{V}^*$ and u is generated by a feedback law (2.9), all solutions of

$$\dot{x} = (A + BK)x + BLv, \quad x(0) \in \mathcal{V}^*, \tag{2.11}$$

belong to $\mathcal{V}^*$ for all times t and all inputs v. Pick λ_0 in the spectrum of $(A + BK)|_{\mathcal{V}^*}$, let x_0 be the corresponding eigenvector, and set $u_0 := Kx_0$. Then it is trivial to check that $\left[\begin{smallmatrix} x_0 \\ u_0 \end{smallmatrix}\right]$ solves (2.8) for $\lambda = \lambda_0$, and hence λ_0 is a right zero of (2.7). Those zeros that are reachable modes for the system (2.11) can actually be moved arbitrarily in the complex plane by a suitable choice of v. Those that are *not* reachable are fixed and are called *invariant zeros* of W. They are in fact even independent of the particular choice of the matrix K [28]. The *maximal reachability subspace* $\mathcal{R}^*(A, B, C, D)$ of W is precisely the maximal subspace of $\mathcal{V}^*$ that is reachable by inputs produced by feedback laws of the form (2.9). If $\mathcal{R}^*(A, B, C, D) = 0$, then all zeros are invariant.

In our setup the spectral factors W are most naturally viewed as operators acting on input functions from the left, and it is more appropriate to consider *left zeros* instead. These are defined simply as the right zeros of the transpose W'. Given a

[2] Note that there are infinitely many λ for which the matrix in (2.8) has a nonzero kernel when $p > m$, and hence there are infinitely many right zeros in this case.

minimal realization of W as in (2.7), a complex number λ is then a left zero of W if and only if there is a nonzero vector $\left[\begin{smallmatrix} z_0 \\ u_0 \end{smallmatrix}\right]$ that solves

$$\begin{bmatrix} A' - \lambda I & C' \\ B' & D' \end{bmatrix} \begin{bmatrix} z_0 \\ u_0 \end{bmatrix} = 0. \tag{2.12}$$

It is easy to show that the vectors z_0 solving (2.12) for some λ form a subspace $\mathcal{V} \subset \mathbb{R}^n$ that is (A', C')-invariant and output-nulling. In fact, $\mathcal{V}$ is a subspace of the *maximal output nulling subspace* $\mathcal{V}^* := \mathcal{V}^*(A', C', B', D')$ of the *dual* system

$$(\Sigma') \quad \begin{cases} \dot{z} = A'z + C'u, \\ v = B'z + D'u \end{cases} \tag{2.13}$$

corresponding to W'. We note that the maximal reachability subspace of Σ', i.e., the subspace $\mathcal{R}^*(A', C', B', D')$ is just the zero space, since W' is left invertible [12] (Theorem 3.36). In other words, the left zeros of W are all invariant.

Now, since $\mathcal{R}^*(A', C', B', D') = 0$, it can be shown that there is a friend K', whose restriction to $\mathcal{V}^*$ is unique, making $\mathcal{V}^*$ $(A' + C'K')$-invariant. The autonomous system

$$\dot{z}(t) = (A' + C'K')z(t), \qquad z(0) \in \mathcal{V}^*, \tag{2.14}$$

with state space $\mathcal{V}^*$, will be called the (left) *zero dynamics* of Σ (or of W) [4], [21]. The eigenvalues of the feedback matrix $(A' + C'K')|_{\mathcal{V}^*}$ are the (*left*) *zeros* of W. As we have pointed out above, all left zeros are invariant. Clearly the invariant zeros of W are the same from the left and from the right. There are, however, noninvariant right zeros of W that are not left zeros (since, in general, $\mathcal{R}^*(A, B, C, D) \neq 0$). From now on we shall only consider left zeros and left zero dynamics, and therefore we shall drop the attribute "left."

Note that the zero dynamics of W is naturally defined only modulo similarity, i.e., modulo coordinate transformations in the state space of minimal realizations (1.1) of W. The vector space $\mathcal{V}^* := \mathcal{V}^*(A', C', B', D')$ will be called the space of *zero directions* of W.

For later reference we shall now explicitly compute the zero dynamics of W for the special case under consideration. To this end, it is convenient to write the system (1.9) in standard form taking

$$\begin{bmatrix} B \\ D \end{bmatrix} = \begin{bmatrix} B_1 & B_2 \\ R^{1/2} & 0 \end{bmatrix}, \tag{2.15}$$

where $R = DD'$ and $R^{1/2}$ is the symmetric square root of R. This can be achieved by an orthogonal coordinate transformation in input space, which of course will not affect the zeros of the spectral factor W. Eliminating the noise dw_1 in

$$(\Sigma) \quad \begin{cases} dx = Axdt + B_1 dw_1 + B_2 dw_2, \\ dy = Cxdt + R^{1/2} dw_1 \end{cases} \tag{2.16}$$

produces a state representation

$$dx = \Gamma x dt + B_1 R^{-1/2} dy + B_2 dw_2 \tag{2.17}$$

in feedback form where Γ is the feedback matrix

$$\Gamma = A - B_1 R^{-1/2} C. \tag{2.18}$$

Let us return to the dual control system (2.13). Then, setting the output v equal to zero yields

$$\left\{\begin{array}{l} \dot{z} = A'z + C'u, \\ 0 = B_1'z + R^{1/2}u, \\ 0 = B_2'z \end{array}\right. \tag{2.19}$$

or, eliminating the control u,

$$\left\{\begin{array}{l} \dot{z} = \Gamma'z, \\ B_2'z = 0. \end{array}\right. \tag{2.20}$$

Consequently, the maximal output-nulling subspace $\mathcal{V}^*$ is precisely

$$\mathcal{V}^* = \langle \Gamma \mid B_2 \rangle^{\perp}, \tag{2.21}$$

i.e., the orthogonal complement of the reachability space

$$\langle \Gamma \mid B_2 \rangle = \operatorname{Im}(B_2, \Gamma B_2, \Gamma^2 B_2, \ldots) \tag{2.22}$$

in $\mathbb{R}^n$. Now, it follows from the discussion above that the invariant zeros of W are precisely the eigenvalues of $\Gamma'|_{\mathcal{V}^*}$, for the maximal reachability space $\mathcal{R}^*$ of the autonomous dynamics (2.20) is zero. Consequently, $\Gamma'|_{\mathcal{V}^*}$ is the generator of the zero dynamics of W. In particular

$$\{\text{zeros of } W\} = \sigma\{\Gamma'|_{\mathcal{V}^*}\}. \tag{2.23}$$

Next we turn to the stochastic version of this theory. For this we need the following definition.

DEFINITION 2.1. *Let X be a Markovian splitting subspace. A subspace $Y \subset X$ is called output induced if*

(i) $Y \subset H_0$;

(ii) $U_t Y \subset Y \vee H^+_{[0,t]}$ *for* $t \geq 0$,

where $H^+_{[0,t]}$ is the subspace spanned by the output dy on the finite interval $[0,t]$,, i.e.,

$$H^+_{[0,t]} = \text{closure}\,\{a'\,[y(\tau) - y(s)]\,|a \in \mathbb{R}^m, \tau, s \in [0,t]\,\};$$

(iii) $U_t Y \subset Y \vee H^-_{[t,0]}$ *for* $t \leq 0$,

where $H^-_{[t,0]}$ is spanned by the output on $[t,0]$.

The following proposition, the proof of which will be postponed to the appendix, establishes the fact that an output-induced subspace is actually a stochastic counterpart of an (A,B)-invariant subspace in geometric control theory.

PROPOSITION 2.1. *Let $Y \subset X \cap H_0$ be output-induced. Then*

$$FY \subset Y \vee \operatorname{Im} N \tag{2.24}$$

where the linear operators $F: X \to X$ and $N: \mathbb{R}^m \to X$ are defined by (2.3) *and*

$$Na = \lim_{h \downarrow 0} \frac{1}{h} E^X a'[y(h) - y(0)], \tag{2.25}$$

respectively.

As we have already noted above, F has the matrix representation A' in the basis in X consisting of the components of $x(0)$. Moreover, it was proven in [14] that

$$Cx(0) = \lim_{h\downarrow 0} \frac{1}{h} E^X[y(h) - y(0)] \tag{2.26}$$

and consequently $Na = a'Cx(0)$, i.e., N has the matrix representation C' in the basis $x(0)$. Therefore, condition (2.24) is equivalent to (A', C')-invariance of the representative of Y in the aforementioned coordinate system. To make this correspondence more precise we shall consider next the problem of finding the *maximal* output-induced subspace of a given minimal Markovian splitting subspace.

THEOREM 2.2. *Let X be a minimal Markovian splitting subspace. Then there is a maximal output-induced subspace of X, namely, $Y^* := X \cap H_0$. The subspace Y^* is maximal in the sense that $Y \subset Y^*$ for any other output-induced subspace Y of X.*

There is a close connection between the concept of maximal output-induced subspace of a minimal Markovian splitting subspace and the zero dynamics of the corresponding minimal spectral factor. This connection is best understood by regarding the realization (1.9).

LEMMA 2.3. *Let $X \in \mathcal{X}$, and let* (1.9) *be a corresponding minimal realization. Then*

$$X \cap H_0 = \{a'x(0) | \, a \in \mathcal{V}^*(A', C', B', D')\}.$$

Proof. First take $\xi \in X \cap H_0$. Then ξ has a representation $\xi = a'x(0)$ where $a \in \mathbb{R}^n$. We shall prove that $a \in \mathcal{V}^* := \mathcal{V}^*(A', C', B', D')$. We immediately see that

$$\xi = a'x(0) = \int_{-\infty}^{0} a'e^{-At}B\,dw(t). \tag{2.27}$$

On the other hand, since $\xi \in H_0$, there is a representation

$$\xi = \int_{-\infty}^{\infty} \hat{u}(i\omega)'\,d\hat{y}(i\omega), \tag{2.28}$$

where $\hat{u}$ is a vector function on the imaginary axis that is L_2 with respect to the matrix measure $\frac{1}{2\pi}\Phi(i\omega)d\omega$ and $d\hat{y}$ is the spectral measure [24] of the process dy, i.e.,

$$y(t) - y(s) = \int_{-\infty}^{\infty} \frac{e^{i\omega t} - e^{i\omega s}}{i\omega} d\hat{y}.$$

This spectral measure may be written

$$d\hat{y} = Wd\hat{w}$$

in terms of the spectral factor (1.1), the transfer function of (1.9), and the spectral measure $d\hat{w}$ of the generating noise dw of (1.9). Consequently,

$$\xi = \int_{-\infty}^{\infty} \hat{u}(i\omega)'W(i\omega)d\hat{w}(i\omega), \tag{2.29}$$

where $\hat{f} := W'\hat{u}$ is an L_2 function on the imaginary axis with inverse Fourier transform

$$f(t) = \int_{-\infty}^{t} B'e^{A'(t-s)}C'u(s)\,ds + D'u(t), \tag{2.30}$$

where u is the inverse Fourier transform of $\hat{u}$ in the L_2 sense. (To see that $\hat{u}$ is L_2 note that $\Phi(\infty)$ is nonsingular by assumption.) Then (2.29) may be written

$$\xi = \int_{-\infty}^{\infty} f(-t)'\, dw(t) \tag{2.31}$$

in the time domain [24], [15], and, in view of (2.27), we must have

$$f(t) = \begin{cases} B'e^{A't}a & \text{for } t \geq 0, \\ 0 & \text{for } t \leq 0. \end{cases} \tag{2.32}$$

Hence, if we set

$$v(t) := B'\left\{e^{A't}(-a) + \int_{-\infty}^{t} e^{A'(t-s)}C'u(s)\, ds\right\} + D'u(t) \tag{2.33}$$

$$= B'\left\{e^{A't}[-a + \bar{z}(0)] + \int_{0}^{t} e^{A'(t-s)}C'u(s)\, ds\right\} + D'u(t), \tag{2.34}$$

where

$$\bar{z}(0) = \int_{-\infty}^{0} e^{-A's}C'u(s)\, ds,$$

it is seen from (2.32) that $v(t) = 0$ for $t \geq 0$, and hence u is an output-nulling input for the dual control system

$$(\Sigma') \quad \begin{cases} \dot{z} = A'z + C'u, \\ v = B'z + D'u \end{cases} \tag{2.35}$$

initiated at $z(0) = -a + \bar{z}(0)$. Therefore $-a + \bar{z}(0) \in \mathcal{V}^*$. On the other hand, (2.30) and (2.32) show that the output of Σ' with control u and initial condition $z(-\infty) = 0$ is identically zero on the negative real axis. Therefore the corresponding state trajectory

$$\bar{z}(t) = \int_{-\infty}^{t} e^{A'(t-s)}C'u(s)\, ds$$

belongs to $\mathcal{V}^*$ for $t \leq 0$. Hence, in particular, $\bar{z}(0) \in \mathcal{V}^*$, and consequently $a \in \mathcal{V}^*$ as claimed.

To prove the converse statement, we first note that the coercivity of Φ ensures that $\Gamma'|_{\mathcal{V}^*}$ has no eigenvalues on the imaginary axis, since the zeros of a minimal spectral factor W are also zeros of the spectral density Φ. Therefore $\mathcal{V}^*$ can be decomposed into a direct sum

$$\mathcal{V}^* = \mathcal{V}^*_- \dotplus \mathcal{V}^*_+, \tag{2.36}$$

where $\mathcal{V}^*_-$ is the sum of the generalized eigenspaces corresponding to eigenvalues of $\Gamma'|_{\mathcal{V}^*}$ with negative real part and $\mathcal{V}^*_+$ is the corresponding subspace for eigenvalues with positive real parts. Both $\mathcal{V}^*_-$ and $\mathcal{V}^*_+$ are of course invariant for Γ'. We want to prove that, if $a \in \mathcal{V}^*$, then $a'x(0) \in X \cap H_0$. To this end, take $a \in \mathcal{V}^*$ and let $a = a_- + a_+$ where $a_- \in \mathcal{V}^*_-$ and $a_+ \in \mathcal{V}^*_+$. Since, in view of (2.21), $\mathcal{V}^* \perp \operatorname{Im} B_2$, (2.17) yields

$$d(a'x) = a'\Gamma x dt + a'B_1R^{-1/2}dy \tag{2.37}$$

for any $a \in \mathcal{V}^*$. Therefore, by choosing a basis in $\mathcal{V}^*$ consistent with the direct sum decomposition (2.36) , (2.37) produces two equations relative to $\mathcal{V}^*_-$ and $\mathcal{V}^*_+$, which by Γ-invariance can be integrated separately on the negative and positive time axis respectively. It then follows that

$$a'_- x(0) = \int_{-\infty}^{0} a'_- e^{-\Gamma t} B_1 R^{-1/2} dy(t) \quad \text{for} \quad a_- \in \mathcal{V}^*_- \tag{2.38}$$

and

$$a'_+ x(0) = \int_{0}^{\infty} a'_+ e^{-\Gamma t} B_1 R^{-1/2} dy(t) \quad \text{for} \quad a_+ \in \mathcal{V}^*_+, \tag{2.39}$$

and hence $a'_- x(0) \in X \cap H^-$ and $a'_+ x(0) \in X \cap H^+$ so that $a'x(0) \in X \cap H_0$, proving the lemma. □

Remark. Note that the basic idea of this construction is that $\mathcal{V}^*$ acts dually in the model (1.9) as a maximal "exogenous-noise-nulling" subspace in the sense that multiplying (1.9) by an $a \in \mathcal{V}^*$ removes the influence of the noninternal components of the input noise dw. An alternative and perhaps more elegant way of seeing this is to consider the adjoint control system

$$(\Sigma^*) \quad \left\{ \begin{array}{l} \dot{z} = -A'z + C'u, \\ v = -B'z + D'u \end{array} \right. \tag{2.40}$$

with transfer function $W^*(s) = W(-s)$, instead of the dual system Σ' defined by (2.13). Clearly Σ^* and Σ' have the same output-nulling subspaces $\mathcal{V}$ and, in particular, the same $\mathcal{V}^*$. (In fact, by a computation similar to the one given above for Σ', we see that the generator of the zero dynamics of Σ^* is $-\Gamma'|_{\mathcal{V}^*}$.) The study of linear functionals $a'x(0)$ of the state at time zero leads naturally to considering the adjoint system Σ^*. Given the stochastic system (1.9), differentiating the bilinear form $z'x$ yields

$$d(z'x) = z'dx + \dot{z}'xdt \tag{2.41}$$
$$= u'dy - v'dw, \tag{2.42}$$

showing that the exogenous noise is blocked out if $z(0) \in \mathcal{V}^*$, i.e., $v = 0$. Then

$$d(z'x) = u'dy$$

can be integrated to establish that $z(0)'x(0) \in X \cap H_0$.

The same idea is used in the following proof.

Proof of Theorem 2.2. Let $\xi \in X \cap H_0$. Then, by Lemma 2.3, $\xi = a'x(0)$ where $a \in \mathcal{V}^*$. Consequently, integrating (2.17) and noting that $\mathcal{V}^* \perp \mathrm{Im} B_2$, we obtain

$$a'x(t) = a'e^{\Gamma t}x(0) + \int_0^t a'e^{\Gamma(t-s)} B_1 R^{-1/2}\, dy(s). \tag{2.43}$$

Since $\mathcal{V}^*$ is Γ'-invariant, $e^{\Gamma' t} a \in \mathcal{V}^*$ and hence the first term in the sum (2.43) belongs to $X \cap H_0$ (Lemma 2.3). Consequently, $X \cap H_0$ satisfies the conditions of Definition 2.1 and is thus output-induced. Since all output-induced subspaces are contained in $X \cap H_0$, it must be maximal. □

The fact that the zero dynamics of W is autonomous is reflected in the following lemma, which will be proved in the appendix.

LEMMA 2.4. *Under the coercivity assumption above, $X \cap H^+_{[0,t]} = 0$ for $t \geq 0$ and $X \cap H^-_{[t,0]} = 0$ for $t \leq 0$ so that the vector sums are direct in* (ii) *and* (iii) *of Definition* 2.1.

In view of Lemma 2.4, an equivalent way of stating Theorem 2.2 is to say that

$$U_t \{X \cap H_0\} \subset X \cap H_0 + H^+_{[0,t]} \quad \text{for} \quad t \geq 0 \tag{2.44}$$

and

$$U_t \{X \cap H_0\} \subset X \cap H_0 + H^-_{[t,0]} \quad \text{for} \quad t \leq 0. \tag{2.45}$$

Note that the direct sum property in Lemma 2.4 is lost as $t \to \infty$, since H^- and H^+ in general have nontrivial intersections with $X \cap H_0$, namely, $X \cap H^-$ and $X \cap H^+$ respectively.

Now, in view of (2.44) and (2.45), there are *oblique* time-varying projectors

$$\pi_t : (X \cap H_0) + H^+_{[0,t]} \to X \cap H_0$$

and

$$\bar{\pi}_t : (X \cap H_0) + H^-_{[-t,0]} \to X \cap H_0,$$

the first being the projection onto $X \cap H_0$ parallel to $H^+_{[0,t]}$ and the second projection onto $X \cap H_0$ parallel to $H^-_{[-t,0]}$. The projectors play the role of feedback in geometric control theory in confining the motion of the state to the subspace $X \cap H_0$. Accordingly, we form the compressed shift operators $V_t(X)$ and $\bar{V}_t(X)$ on $X \cap H_0$ by the relations

$$V_t(X)\xi = \pi_t U_t \xi \tag{2.46}$$

and

$$\bar{V}_t(X)\xi = \bar{\pi}_t U_t^* \xi. \tag{2.47}$$

LEMMA 2.5. *The families $\{V_t(X);\ t \geq 0\}$ and $\{\bar{V}_t(X); t \geq 0\}$ of linear operators are strongly continuous semigroups on $X \cap H_0$.*

Proof. Let $\xi \in X \cap H_0$, and form

$$\begin{aligned} V_t(X)V_s(X) &= \pi_t U_t \pi_s U_s \xi && (2.48)\\ &= \pi_{t+s} U_t \pi_s U_s \xi && (2.49)\\ &= V_{t+s}(X)\xi - \pi_{t+s} U_t (1 - \pi_s) U_s \xi, && (2.50) \end{aligned}$$

where we have used the fact that $\pi_{t+s}|_{X \cap H_0 + H^+_{[0,t]}} = \pi_t$ for $s \geq 0$. But $(1 - \pi_s)U_s\xi \in H^+_{[0,t]}$, and hence

$$U_t(1 - \pi_s)U_s\xi \in H^+_{[0,t+s]};$$

therefore the last term in (2.50) equals zero, establishing the semigroup property for $\{V_t(X);\ t \geq 0\}$. To prove strong continuity, note that, if $t \leq T$, $V_t(X)\xi = \pi_t U_t \xi = \pi_T U_t \xi$, which tends to ξ as $t \to 0$. The rest follows from a symmetric argument. □

Consequently there are infinitesimal generators, i.e., operators $G, \bar{G} : X \cap H_0 \to X \cap H_0$ such that

$$V_t(X) = e^{Gt} \tag{2.51}$$

and

$$\bar{V}_t(X) = e^{\bar{G}t}. \tag{2.52}$$

LEMMA 2.6. *For each $t \geq 0$,*

$$\bar{V}_t(X) = V_t(X)^{-1}; \tag{2.53}$$

i.e., in particular,

$$\bar{G} = -G. \tag{2.54}$$

Proof. Let $\xi \in X \cap H_0$. Then

$$\begin{aligned} \bar{V}_t(X)V_t(X)\xi &= \bar{\pi}_t U_t^* \pi_t U_t \xi & (2.55)\\ &= \xi - \bar{\pi}_t U_t^*(1 - \pi_t)U_t\xi. & (2.56) \end{aligned}$$

Since $(1 - \pi_t)U_t\xi \in H^+_{[0,t]}$, we have

$$U_t^*(1 - \pi_t)U_t\xi \in H^-_{[-t,0]},$$

and therefore the last term of (2.56) is zero. □

Consequently, we may define $V_t(X)$ also for negative t. In fact, setting

$$V_t(X) = \bar{V}_{-t}(X)$$

is equivalent to defining $V_t(X)$ for all $t \in \mathbb{R}$ by means of (2.46) with $\pi_{-t} = \bar{\pi}_t$ for $t \leq 0$. Hence the family of operators $\{V_t(X);\ t \in \mathbb{R}\}$ is actually a *group.*

The following proposition characterizes the output-induced subspaces of X as the invariant subspaces for the group $\{V_t(X);\ t \in \mathbb{R}\}$.

PROPOSITION 2.7. *The output-induced subspaces of X are precisely the G-invariant subspaces of $X \cap H_0$.*

Proof. First suppose that $Y \subset X$ is output-induced. Then

$$U_tY \subset Y + H^+_{[0,t]} \quad \text{for} \quad t \geq 0, \tag{2.57}$$

so applying the projection π_t to both sides we see that $e^{Gt}Y \subset Y$. Conversely, suppose that $Y \subset X \cap H_0$ is e^{Gt}-invariant. From (2.44) we have that

$$U_tY \subset X \cap H_0 + H^+_{[0,t]} \quad \text{for} \quad t \geq 0. \tag{2.58}$$

We want to show that $X \cap H_0$ in (2.58) can be exchanged for Y so that (2.57) is obtained. However, this is obvious by applying the projector π_t to (2.58) and noting that, by assumption, $e^{Gt}Y \subset Y$. Trivially, the corresponding statement for $t \leq 0$ follows from (2.45) by an analogous argument. □

We shall identify two particularly important G-invariant subspaces of $X \cap H_0$, namely, the *past-output-induced* subspace $X \cap H^-$ and the the *future-output-induced* subspace $X \cap H^+$. In fact, suppose that $\xi \in X \cap H^-$ and $t \geq 0$. Then $U_t^* \xi \in H^-$ and

$$(1 - \bar{\pi}_t) U_t^* \xi \in H^-_{[-t,0]} \subset H^-,$$

and hence

$$e^{-Gt}\xi = \bar{V}_t(X \cap H_0)\xi = \bar{\pi}_t U_t^* \xi \in X \cap H^-,$$

because the range of $\bar{\pi}_t$ is contained in X. Therefore $X \cap H^-$ is G-invariant. A symmetric argument shows that $X \cap H^+$ is also G-invariant. Consequently, by Proposition 2.7, $X \cap H^-$ and $X \cap H^+$ are output-induced subspaces of X.

Coercivity also implies that $H^- \cap H^+ = 0$ [14] so that the sum

$$H_0 = H^- + H^+ \tag{2.59}$$

is direct. The following lemma states in particular that the maximal output-induced subspace can be represented as a direct sum of $X \cap H^-$ and $X \cap H^+$.

LEMMA 2.8. *Let H^-, H^+, H_0 be defined as in* §1, *and let X be a splitting subspace. Then*

$$X \cap H_0 = \left(X \cap H^-\right) + \left(X \cap H^+\right) \tag{2.60}$$

where the sum is direct.

For the proof let us first recall that a Markovian splitting subspace can be uniquely represented as the intersection

$$X = S \cap \bar{S} \tag{2.61}$$

of a pair $(S, \bar{S})$ of subspaces of the ambient subspace H that satisfy

$$S \supset H^- \quad \text{and} \quad \bar{S} \supset H^+, \tag{2.62}$$

the invariance properties

$$U_t^* S \subset S \quad \text{and} \quad U_t \bar{S} \subset \bar{S} \quad \text{for all} \quad t \geq 0, \tag{2.63}$$

and intersect perpendicularly in the sense that

$$H = S^\perp \oplus X \oplus \bar{S}^\perp, \tag{2.64}$$

where $S^\perp$ and $\bar{S}^\perp$ are the orthogonal complements in H of S and $\bar{S}$ respectively (see, e.g., [16]). We shall write $X \sim (S, \bar{S})$ to refer to this representation. The class $\mathcal{X}$ of minimal Markovian splitting subspaces consists precisely of the $X \sim \left(S, \bar{S}\right)$ that are *observable*, i.e.,

$$\bar{S} = H^+ \vee S^\perp, \tag{2.65}$$

and *constructible*, i.e.,

$$S = H^- \vee \bar{S}^\perp \tag{2.66}$$

(see [16]).

Proof of Lemma 2.8. Since $X \cap H^- \subset X \cap H_0$ and $X \cap H^+ \subset X \cap H_0$, it trivially holds that

$$X \cap H_0 \supset (X \cap H^-) \vee (X \cap H^+)\,. \tag{2.67}$$

Since $H_0 = H^- + H^+$ is a direct sum, so is that of (2.67). Hence it just remains to show that the converse inclusion holds. To this end suppose that $\lambda \in X \cap H_0$. Since $\lambda \in H_0 = H^- + H^+$, there are unique $\alpha \in H^-$ and $\beta \in H^+$ such that

$$\lambda = \alpha + \beta.$$

Then, since $\lambda \in X \subset \bar{S}$ and $\beta \in H^+ \subset \bar{S}$, we have $\alpha = \lambda - \beta \in \bar{S}$, and hence

$$\alpha \in \bar{S} \cap H^- = \bar{S} \cap S \cap H^- = X \cap H^-$$

Then $\beta = \lambda - \alpha \in X$, i.e., $\beta \in X \cap H^+$. This completes the proof of the lemma.
□

Remark. The fact that $X \cap H^-$ and $X \cap H^+$ are output-induced can be seen from first principles using Lemma 2.8. In fact, from (2.61), we see that $X \cap H^- = \bar{S} \cap H^-$, and hence

$$U_t\left\{X \cap H^-\right\} \subset \bar{S} \cap \left(H^- + H^+_{[0,t]}\right) \tag{2.68}$$

$$= \bar{S} \cap H^- + H^+_{[0,t]}. \tag{2.69}$$

Here the first inclusion follows from the U_t-invariance (2.63) of $\bar{S}$ and the second equality from Lemma 2.8, noting that $\bar{S} \sim (H, \bar{S})$ is a splitting subspace and $\bar{S} \supset H^+_{[0,t]}$. This shows that $X \cap H^-$ satisfies condition (ii) of Definition 2.1. A symmetric argument proves condition (iii), while condition (i) is trivially satisfied. Hence $X \cap H^-$ is output-induced. In the same way we show that $X \cap H^+$ is output-induced.

The following theorem is one of the main results of this paper, tying together the geometry of minimal Markovian splitting subspaces to the zero dynamics of minimal spectral factors.

THEOREM 2.9. *Let X be a minimal Markovian splitting subspace, and let W be the corresponding spectral factor. Then the group $\{V_t(X);\ t \in \mathbb{R}\}$ acting on the maximal output-induced subspace $X \cap H_0$, of X, is isomorphic to the zero dynamics* (2.14) *of W in the sense that the linear bijective map $T : \mathcal{V}^*(A', C', B', D') \to X \cap H_0$, defined by $Ta = a'x(0)$, makes the following diagram commutative:*

$$\begin{tikzcd} X \cap H_O \arrow[r, "V_t(X)"] & X \cap H_0 \\ \mathcal{V}^* \arrow[u, "T"] \arrow[r, "e^{(A'+C'K')t}"] & \mathcal{V}^* \arrow[u, "T"'] \end{tikzcd}$$

In particular,

$$\{\text{zeros of } W\} = \sigma(G), \tag{2.70}$$

where $\sigma(G)$ is the spectrum of the infinitesimal generator of the group $\{V_t(X);\ t \in \mathbb{R}\}$. The restricted groups $V_t^-(X) := V_t(X)|_{X \cap H^-}$ and $V_t^+(X) := V_t(X)|_{X \cap H^+}$, $t \in \mathbb{R}$, describe the asymptotically stable and antistable zero dynamics of W, the respective generators

$$G_s = G|_{X \cap H^-} \quad \text{and} \quad G_u = G|_{X \cap H^+}$$

having spectra $\sigma(G_s)$ and $\sigma(G_u)$ coinciding with the zeros of W with respectively negative and positive real parts.

Hence, in particular, $\dim(X \cap H_0)$, $\dim(X \cap H^-)$, and $\dim(X \cap H^+)$ are respectively the total number of zeros of W, the number zeros in the open left half plane (*stable zeros*), and the number of zeros in the open right half plane (*antistable zeros*). (The last statement is actually a splitting subspace version of Theorem 4.1 in [11] (see also [1]) as we shall see in §4 upon introducing state covariances.) If X is internal and $\dim X = n$, then there are exactly n zeros. If $X \cap H_0 = 0$, there are no zeros.

We shall call G the *generator of the zero dynamics* of X. Since in this paper we consider the special case when R is nonsingular, we may, as we have already pointed out, write the zero dynamics (2.14) as

$$\dot{z} = \Gamma' z, \quad z \in \mathcal{V}^*, \tag{2.71}$$

where Γ is defined by (2.18). By Lemma 2.3, the map T in the commutative diagram of Theorem 2.9 assigns the value $a'x(0) \in X \cap H_0$ to each $a \in \mathcal{V}^*$, i.e.,

$$T : a \to a'x(0). \tag{2.72}$$

Proof. Take $a \in \mathcal{V}^*$ so that $a'x(0) \in X \cap H_0$. Then (2.43) holds. From this sum with the first term in $X \cap H_0$ and the second in $H^+_{[0,t]}$ for $t \geq 0$, we obtain

$$\pi_t U_t a'x(0) = \pi_t a'x(t) = a'e^{\Gamma t}x(0)$$

for $t \geq 0$, i.e.,

$$e^{Gt}a'x(0) = a'e^{\Gamma t}x(0).$$

Hence $G\,[a'x(0)] = a'\Gamma x(0)$, i.e., $GTa = T\Gamma' a$, proving the similarity

$$G = T\Gamma'|_{\mathcal{V}^*}T^{-1}. \tag{2.73}$$

Moreover, note that (2.38) and (2.39) imply that

$$T\mathcal{V}^*_- \subset X \cap H^- \quad \text{and} \quad T\mathcal{V}^*_+ \subset X \cap H^+.$$

However, since, by Lemma 2.8 and (2.36) the two vector sums

$$T\mathcal{V}^* = T\mathcal{V}^*_- + T\mathcal{V}^*_+$$

and

$$X \cap H_0 = X \cap H^- + X \cap H^+$$

are direct and $T\mathcal{V}^* = X \cap H_0$ (Lemma 2.3), it must hold that

$$T\mathcal{V}^*_- = X \cap H^- \quad \text{and} \quad T\mathcal{V}^*_+ = X \cap H^+. \tag{2.74}$$

Then, by retracing the first part of the proof with $\mathcal{V}^*$ replaced by $\mathcal{V}^*_-$ and $\mathcal{V}^*_+$, we establish the similarity relations

$$G_s = T\Gamma'|_{\mathcal{V}^*_-}T^{-1} \quad \text{and} \quad G_u = T\Gamma'|_{\mathcal{V}^*_+}T^{-1},$$

which clearly shows that G_s is stable and G_u is antistable. This completes the proof of the theorem. □

Next we shall derive some representation formulas for the restrictions of the group $\{V_t(X);\ t \in \mathbb{R}\}$ to the complementary invariant subspaces $X \cap H^-$ and $X \cap H^+$. These relations are connected to the generalization to the Riccati inequality of certain projection results concerning the algebraic Riccati equation due to Willems [27]. This will be discussed in §5.

Because of the direct sum decomposition (2.59), any $\eta \in H_0$ has a unique decomposition

$$\eta = \pi_- \eta + \pi_+ \eta \tag{2.75}$$

where $\pi_- : H_0 \to H^-$ is the projection on H^- along H^+ and $\pi_+ : H_0 \to H^+$ is the projection on H^+ along H^-.

LEMMA 2.10. *Let $t \geq 0$. Then if $\xi \in X \cap H^-$, we have $\pi_- U_t \xi \in X \cap H^-$; and, dually, if $\xi \in X \cap H^+ m$ it follows that $\pi_+ U_t^* \xi \in X \cap H^+$. Moreover, the restrictions of $V_t(X)$ to the complementary invariant subspaces $X \cap H^-$ and $X \cap H^+$ coincide with the above compressed shifts $\pi_- U_t : X \cap H^- \to X \cap H^-$ and $\pi_+ U_t^* : X \cap H^+ \to X \cap H^+$ respectively, i.e.,*

$$V_t^-(X) := V_t(X)|_{X \cap H^-} = \pi_- U_t\, |_{X \cap H^-}$$

and

$$V_{-t}^+(X) := V_{-t}(X)|_{X \cap H^+} = \pi_+ U_t^*\, |_{X \cap H^+}.$$

Proof. Let $t \geq 0$, and take $\xi \in X \cap H^-$. Since $X \cap H^-$ is output-induced (see, e.g., the remark before Theorem 2.9),

$$U_t \xi \in X \cap H^- + H^+_{[0,t]}.$$

Therefore, since $X \cap H^- \subset H^-$ and $H^+_{[0,t]} \subset H^+$, we have

$$\pi_- U_t \xi = \pi_t U_t \xi = V_t(X) \xi.$$

The $(\pi_- U_t)$-invariance of $X \cap H^-$ now follows from the $V_t(X)$-invariance. A symmetric result yields the corresponding result for $X \cap H^+$. □

3. Zeros and ordering. In this section we shall study the zero structure of the family of all minimal (analytic) spectral factors by using a partial ordering of the family $\mathcal{X}$ of all minimal Markovian splitting subspaces that are defined in some common probabilistic setting. Such a setting can be described by a sufficiently large common Hilbert space $\hat{H}$ containing H_0. It can be shown [16, §§5.2 and 5.3] that it suffices to take $\hat{H}$ to be of the form

$$\hat{H} = H_0 \oplus H\,(d\eta)\,,$$

where $d\eta$ is some n-dimensional Wiener process independent of dy and $H(d\eta)$ is the space generated by the increments of the components of η. The Hilbert space $\hat{H}$ is endowed with a shift $\{\hat{U}_t; t \in \mathbb{R}\}$, namely, the one induced by $(dy, d\eta)$; and the ambient space of each minimal X in this setting is a doubly invariant subspace of $\hat{H}$ containing H_0. The shift $\{U_t\}$ corresponding to $X \in \mathcal{X}$ is just the restriction of $\{\hat{U}_t\}$ to its ambient space H. Recall that the ambient space H has a representation $H(dw)$, where the Wiener process dw may be identified with the driving noise of a minimal stochastic realization (1.9) corresponding to X.

In [16] we introduced a partial order of $\mathcal{X}$ defined as follows. Given two minimal Markovian splitting subspaces, X_1 and X_2, we say that $X_1 \leq X_2$ if

$$||E^{X_1}\lambda|| \leq ||E^{X_2}\lambda|| \quad \text{for all } \lambda \in H^+$$

or, equivalently,

$$||E^{X_2}\lambda|| \leq ||E^{X_1}\lambda|| \quad \text{for all } \lambda \in H^-.$$

With the above choice of Hilbert space $\hat{H}$, it can be shown that $\leq$ is a bona fide partial ordering relation of $\mathcal{X}$; i.e., in particular, $X_1 \leq X_2$ and $X_2 \leq X_1$ imply that $X_1 = X_2$. Moreover, $\mathcal{X}$ has a maximal and a minimal element, X_+ and X_-, in this ordering, i.e.,

$$X_- \leq X \leq X_+ \tag{3.1}$$

for each $X \in \mathcal{X}$, where $X_- := E^{H^-}H^+$ and $X_+ := E^{H^+}H^-$ are respectively the forward and the backward predictor spaces. Clearly both X_- and X_+ belong to $\mathcal{X}_0$.

It can be seen from (3.1) that any $X \in \mathcal{X}$ is bounded from below and from above by elements in $\mathcal{X}_0$, namely, by X_- and X_+ respectively. In this context, a relevant question is whether these internal bounds could be tightened. In [16] it was shown that, for each $X \in \mathcal{X}$, there are unique $X_{0-}, X_{0+} \in \mathcal{X}_0$ so that

$$X_1 \leq X_{0-} \leq X \leq X_{0+} \leq X_2$$

for all $X_1, X_2 \in \mathcal{X}_0$ such that $X_1 \leq X \leq X_2$. In other words

$$X_{0-} = \max\{X_0 \in \mathcal{X}_0 \mid X_0 \leq X\},$$

$$X_{0+} = \min\{X_0 \in \mathcal{X}_0 \mid X \leq X_0\}$$

are unique, and we call them the *tightest internal bounds* of X.

At several instances that follow we shall consider a restriction of some linear operator to an invariant subspace. Whenever such a restriction occurs, the invariance is automatically implied and will not be stated explicitly.

LEMMA 3.1. *Let* $X_1, X_2 \in \mathcal{X}$, *and suppose that* $X_1 \leq X_2$. *Then*

(i) $X_1 \cap H^+ \subset X_2 \cap H^+$ and $X_2 \cap H^- \subset X_1 \cap H^-$;

(ii) $V_t^-(X_1)|_{X_2 \cap H^-} = V_t^-(X_2), \quad t \in \mathbb{R}$;

(iii) $V_t^+(X_2)|_{X_1 \cap H^+} = V_t^+(X_1), \quad t \in \mathbb{R}$.

Proof. (i) Recall that if $X \sim (S, \bar{S})$ is a minimal Markovian splitting subspace then the corresponding tightest lower internal bound $X_{0-} \sim (S_{0-}, \bar{S}_{0-})$ has the property that $S_{0-} = S \cap H_0$ (Theorem 6.11 in [16]). Now, if $X_1 \leq X_2$, then, with self-explanatory notation, $(X_1)_{0-} \leq X_1 \leq X_2$, and consequently $(X_1)_{0-} \leq (X_2)_{0-}$ or, equivalently, $S_1 \cap H_0 \subset S_2 \cap H_0$, which implies that $S_1 \cap H^+ \subset S_2 \cap H^+$. But, in view of (2.61) and (2.62), this is equivalent to $X_1 \cap H^+ \subset X_2 \cap H^+$. A symmetric argument yields $X_2 \cap H^- \subset X_1 \cap H^-$.

(ii) First take $t \geq 0$. Then, by Lemma 2.10,

$$V_t^-(X) = \pi_- U_t|_{X \cap H^-}$$

for any $X \in \mathcal{X}$, where $\pi_- : H_0 \to H^-$ is the oblique projection parallel to H^+. Therefore, since $X_2 \cap H^- \subset X_1 \cap H^-$ and these spaces are both invariant for the compressed shift $\pi_- U_t$ (Lemma 2.10),

$$V_t^-(X_1)|_{X_2 \cap H^-} = V_t^-(X_2) \tag{3.2}$$

for $t \geq 0$. However, for any $X \in \mathcal{X}$,

$$V_t^-(X) = V_t(X)|_{X \cap H^-}$$

for all $t \in \mathbb{R}$, and hence (3.2) may be written

$$V_t(X_1)|_{X_2 \cap H^-} = V_t(X_2)|_{X_2 \cap H^-} \quad \text{for} \quad t \geq 0,$$

which is a statement about groups and consequently holds for all $t \in \mathbb{R}$.

(iii) The proof follows from a symmetric argument to that used to prove (ii), first proving the the statement for $t \leq 0$ and then invoking the group property. □

COROLLARY 3.2. *Let $X \in \mathcal{X}$. Then*

$$V_t^-(X) = V_t^-(X_-)|_{X \cap H^-} = V_t(X_-)|_{X \cap H^-} \tag{3.3}$$

and

$$V_t^+(X) = V_t^+(X_+)|_{X \cap H^+} = V_t(X_+)|_{X \cap H^+}. \tag{3.4}$$

Proof. To prove (3.3) just take $X_1 = X_-$ and $X_2 = X$ in Lemma 3.1 and then observe that $V_t^-(X_-) = V_t(X_-)$. A symmetric argument yields (3.4). □

We see from this lemma that if W, W_-, and W_+ are the spectral factors of X, X_-, and X_+ respectively, then the stable zeros W are also zeros of W_- and the antistable zeros of W are zeros of W_+. We also see that W_- is the minimum phase spectral factor, all its zeros being stable, and that W_+ is the maximum phase spectral factor with only antistable zeros.

Lemma 3.1 with Corollary 3.2 has a number of other important consequences that will be discussed later. Before turning to this, however, we shall complete the analysis of the relation between subspace inclusion of the type exhibited in statement (i) of Lemma 3.1.

LEMMA 3.3. *Let $X_1, X_2 \in \mathcal{X}_0$. Then, for each $X \in \mathcal{X}$,*

(i) $X_1 \leq X \Longleftrightarrow X_1 \cap H^+ \subset X \cap H^+$,

(ii) $X \leq X_2 \Longleftrightarrow X_2 \cap H^- \subset X \cap H^-$.

Moreover, $X_1 = X_{0-}$ if and only if $X_1 \cap H^+ = X \cap H^+$ and $X_2 = X_{0+}$ if and only if $X_2 \cap H^- = X \cap H^-$.

Proof. We begin by proving (i). In view of Lemma 3.1, it remains to prove that $X_1 \cap H^+ \subset X \cap H^+$ implies that $X_1 \leq X$, which, by Theorem 6.8(ii) in [16], is equivalent to $S_1 \subset S$. This in turn is certainly implied by $S_1 \subset S \cap H_0$.

Now, for any splitting subspace $X \sim (S, \bar{S})$, S is itself a splitting subspace, namely, $S \sim (S, H)$; and consequently Lemma 2.8 implies that

$$S \cap H_0 = H^- + X \cap H^+, \tag{3.5}$$

because, by (2.61) and (2.62), $S \cap H^- = H^-$ and $S \cap H^+ = S \cap \bar{S} \cap H^+ = X \cap H^+$.

Then, by (3.5), $X_1 \cap H^+ \subset X \cap H^+$ implies that $S_1 = S_1 \cap H_0 \subset S \cap H_0$, proving (i). A completely symmetric argument yields (ii). By Theorem 6.11 in [16], $X_1 = X_{0-}$ is equivalent to $S_1 = S \cap H_0$. This implies that $S_1 \cap H^+ = S \cap H^+$, i.e.,

$$X_1 \cap H^+ = X \cap H^+. \tag{3.6}$$

On the other hand, there is only one $X_1 \in \mathcal{X}_0$ satisfying (3.6), because (3.6) and

$$S_1 = H^- + X_1 \cap H^+$$

determine S_1 uniquely and for minimal Markovian splitting subspaces there is a one-to-one correspondence between X and S as can be seen from (2.65). Hence we have shown that (3.6) is equivalent to $X_1 = X_{0-}$. In the same way we show that

$$X_2 \cap H^- = X \cap H^-$$

is equivalent to $X_2 = X_{0+}$. □

THEOREM 3.4. *Let $X_1, X_2 \in \mathcal{X}_0$, and suppose $X_1 \le X_2$. Then:*

(i) *For each $X \in \mathcal{X}$,*

$$X_1 \le X \le X_2 \Longleftrightarrow X_1 \cap X_2 \subset X.$$

Moreover, $X_1 = X_{0-}$ if and only if $X_1 \cap X_2 = X \cap X_2$ and $X_2 = X_{0+}$ if and only if $X_1 \cap X_2 = X \cap X_1$.

(ii) *If $X_1 \cap X_2 \subset X$, then $X_1 \cap X_2$ is a $V_t(X)$-invariant subspace for each $t \in \mathbb{R}$, i.e.,*

$$G[X_1 \cap X_2] \subset X_1 \cap X_2. \tag{3.7}$$

Conversely, any G-invariant subspace $Z \subset X \cap H_0$ takes the form $Z = X_1 \cap X_2$ for some unique $X_1, X_2 \in \mathcal{X}_0$ such that $X_1 \le X \le X_2$.

The proof of this theorem is rather long and technical. For this reason we shall first give some interpretations of the results stated so far and postpone the proof of Theorem 3.4 to the end of the section.

COROLLARY 3.5. *Let at least one of $X_1, X_2 \in \mathcal{X}$ be internal, and suppose that $X_1 \le X_2$. Then*

$$V_t(X_1)|_{X_1 \cap X_2} = V_t(X_2)|_{X_1 \cap X_2} \tag{3.8}$$

for all $t \in \mathbb{R}$.

Proof. We want to prove that, for any $\lambda \in X_1 \cap X_2$,

$$V_t(X_1)\xi = V_t(X_2)\xi$$

for all $t \in \mathbb{R}$. To this end, first suppose that $t \ge 0$ and set $\xi_i := V_t(X_i)$, $i = 1, 2$. Then $\xi_i = \pi_t^{(i)} U_t \lambda$, where, for each $i = 1, 2$,

$$\pi_t^{(i)} : X_i \cap H_0 + H^+_{[0,t]} \to X_i \cap H_0$$

is the oblique projector onto $X_i \cap H_0$ parallel to $H^+_{[0,t]}$. Hence there are $\eta_1, \eta_2 \in H^+_{[0,t]}$ such that

$$U_t \lambda = \xi_1 + \eta_1 = \xi_2 + \eta_2.$$

Now, applying the invariance result of Theorem 3.4 twice, first taking $X = X_1$ and then $X = X_2$, we see that both ξ_1 and ξ_2 must belong to $X_1 \cap X_2$. But

$$X_1 \cap X_2 + H^+_{[0,t]}$$

is a direct sum (Lemma 2.3); and hence we must have $\xi_1 = \xi_2$ (and $\eta_1 = \eta_2$), establishing (3.8) for $t \geq 0$. Because of the group property, (3.8) actually holds for all $t \in \mathbb{R}$. □

Recalling the characterization of Proposition 2.7 of output-induced subspaces of $X \in \mathcal{X}$, we immediately have the following important corollary of Theorem 3.4.

COROLLARY 3.6. *The output-induced subspaces $Y \subset X \in \mathcal{X}$ are precisely the subspaces of the form $Y = X_1 \cap X_2$ where $X_1, X_2 \in \mathcal{X}_0$ are internal bounds of X, i.e., $X_1 \leq X \leq X_2$.*

As an illustration of Corollary 3.6 we shall give representations of the output-induced subspaces $X \cap H_0$, $X \cap H^-$, and $X \cap H^+$ as intersections of internal minimal Markovian splitting subspaces. As we have already seen, these output-induced subspaces are of special importance in the classification of the zero structure of minimal spectral factors.

PROPOSITION 3.7. *Let $X \in \mathcal{X}$ have tightest internal bounds X_{0-} and X_{0+}. Then*

(i) $X \cap H^- = X \cap X_- = X_{0+} \cap X_-$,

(ii) $X \cap H^+ = X \cap X_+ = X_{0-} \cap X_+$,

(iii) $X \cap H_0 = X \cap X_{0-} = X \cap X_{0+} = X_{0-} \cap X_{0+}$.

Proof. In view of the last statement of Theorem 3.4(i), it only remains to prove that

$$X \cap H^- = X \cap X_-, \tag{3.9}$$

$$X \cap H^+ = X \cap X_+, \tag{3.10}$$

and

$$X \cap H_0 = X_{0-} \cap X_{0+}. \tag{3.11}$$

Taking $X_1 = X_-$ and $X_2 = X$ in Lemma 3.1(i) and recalling that $X_- \subset H^-$, we see that $X \cap H^- \subset X_- \cap H^- \subset X_-$ and, hence, $X \cap H^- \subset X \cap X_-$. Trivially, $X_- \subset H^-$ also implies that $X \cap X_- \subset X \cap H^-$, and hence (3.9) follows. Relation (3.10) follows by symmetry. To prove (3.11), let $X \sim (S, \bar{S})$. Then, by Theorem 6.11 in [16], $S_{0-} = S \cap H_0$ and $\bar{S}_{0+} = \bar{S} \cap H_0$. Hence

$$X_{0-} \cap X_{0+} = S_{0-} \cap \bar{S}_{0+} = S \cap \bar{S} \cap H_0 = X \cap H_0$$

because $X_{0-} \leq X_{0+}$ and hence $S_{0-} \subset S_{0+}$ and $\bar{S}_{0+} \subset \bar{S}_{0-}$. □

Recall that the group $\{V_t(X)\}$ acting on the maximal output-induced subspace $X \cap H_0$ can be identified with the *zero dynamics* of the minimal spectral factor W corresponding to X because of the isomorphism of Theorem 2.9. Similarly the groups $\{V_t^-(X)\}$ and $\{V_t^+(X)\}$ on $X \cap H^-$ and $X \cap H^+$ respectively can be identified with the *stable*, respectively, the *antistable*, *zero dynamics* of W. The partial ordering of minimal Markovian splitting subspaces induces a partial ordering of the stable and antistable zero dynamics of the corresponding spectral factors. We shall say that $\{V_t^-(X_1)\}$ acting on $X_1 \cap H^-$ is a *restriction* of $\{V_t^-(X_2)\}$ acting on $X_2 \cap H^-$ if

$$X_1 \cap H^- \subset X_2 \cap H^-$$

and

$$V_t^-(X_1) = V_t^-(X_2)|_{X_1 \cap H^-}.$$

In the same way we can define restrictions of antistable zero dynamics. Clearly restriction is a partial-order relation.

THEOREM 3.8. *Let $X_1, X_2 \in \mathcal{X}$ with at least one of them be internal, and let W_1 and W_2 be the corresponding minimal spectral factors. Then if $X_1 \leq X_2$:*

(i) *The stable zero dynamics of W_2 is a restriction of the stable zero dynamics W_1. In particular, all stable zeros of W_2 are zeros of W_1.*

(ii) *The antistable zero dynamics of W_1 is a restriction of the antistable zero dynamics W_2. In particular, all antistable zeros of W_1 are zeros of W_2.*

(iii) *The zero dynamics of W_1 and W_2 coincide on the intersection $X_1 \cap X_2$ (i.e., a relation such as* (3.8) *holds).*

Proof. Statements (i) and (ii) are just restatements of (ii) and (iii) of Lemma 3.1, while statement (iii) is a reformulation of Corollary 3.5. □

COROLLARY 3.9. *Let X_{0-} and X_{0+} be the tightest internal bounds of $X \in \mathcal{X}$, and let W_{0-}, W_{0+}, and W be the corresponding minimal spectral factors. Then the zeros of W are precisely the common zeros of W_{0-} and W_{0+}.*

Proof. This follows immediately from Proposition 3.7(iii) and Theorem 3.8(iii). □

From Corollary 3.9 we see that if X_- and X_+ are the tightest internal bounds of X, which in fact is the "generic" situation, then the corresponding spectral factor has no zeros. In fact, W_- and W_+ have no common zero. The other extreme is the situation when X is internal so that $X_{0-} = X = X_{0+}$. Then W has n zeros.

The following corollary of Theorem 3.4 is a splitting-subspace version of an invariance result, due to Willems [27], formulated in the context of the algebraic Riccati equation. It will be used in §5.

COROLLARY 3.10. *Let G_+ be the zero generator of X_+. Then there is a one-to-one correspondence between G_+-invariant subspaces $Z \subset X_+$ and $X \in \mathcal{X}_0$ under which $Z = X \cap X_+$ and $X \sim (S, \bar{S})$ where*

$$S = H^- + Z \quad \text{and} \quad \bar{S} = H^+ \vee S^\perp.$$

Similarly, if G_- is the zero generator of X_-, there is a one-to-one correspondence between G_--invariant subspaces $Z \subset X_-$ and $X \in \mathcal{X}_0$ under which $Z = X \cap X_-$ and $X \sim (S, \bar{S})$ where

$$\bar{S} = H^+ + Z \quad \text{and} \quad S = H^+ \vee \bar{S}^\perp.$$

Proof of Theorem 3.4(i). ($\Rightarrow$) We first prove that if $X_1 \leq X_2$ and X_1, X_2 are internal, then

$$X_1 \cap X_2 = (X_1 \cap X_2 \cap H^-) + (X_1 \cap X_2 \cap H^+). \tag{3.12}$$

The inclusion $\supset$ is trivial, and we use the procedure of the proof of Lemma 2.8 to prove the converse. To this end, take $\lambda \in X_1 \cap X_2$. Then, by Lemma 2.8,

$$\lambda = X_2 \cap \left[(X_1 \cap H^-) + (X_1 \cap H^+)\right].$$

Set $\lambda = \alpha + \beta$ where $\alpha \in X_1 \cap H^-$ and $\beta \in X_1 \cap H^+$. But since $S_1 \subset S_2$ (see proof of Lemma 3.3),

$$X_1 \cap H^+ = S_1 \cap H^+ \subset S_2 \cap H^+ = X_2 \cap H^+ \subset X_2,$$

and therefore $\beta \in X_2$. Hence $\alpha = \lambda - \beta \in X_2$ so that $\alpha \in X_1 \cap X_2 \cap H^-$ and $\beta \in X_1 \cap X_2 \cap H^+$, as required. This proves (3.12). Now if $X_1 \leq X \leq X_2$, then by Lemma 3.3, $X_2 \cap H^- \subset X \cap H^-$, and therefore

$$X_1 \cap X_2 \cap H^- \subset X_1 \cap X \cap H^- = \bar{S}_1 \cap \bar{S} \cap H^-,$$

where we also have used (2.61) and (2.62). But

$$\bar{S} \cap H^- \subset \bar{S} \cap H_0 = \bar{S}_{0+}$$

by Theorem 6.11 in [16], and since $X_1 \leq X_{0+}$, it follows that $\bar{S}_{0+} \subset \bar{S}_1$. Hence

$$X_1 \cap X_2 \cap H^- \subset \bar{S} \cap H^- = X \cap H^-.$$

In the same way we show that

$$X_1 \cap X_2 \cap H^+ \subset X \cap H^+,$$

and therefore (3.12) and Lemma 2.8 imply that

$$X_1 \cap X_2 \subset X \cap H_0 \subset X.$$

($\Leftarrow$) Next suppose that $X_1 \cap X_2 \subset X$. Then

$$X_1 \cap X_2 \cap H^+ \subset X \cap H^+.$$

But $X_1 \cap X_2 \cap H^+ = S_1 \cap S_2 \cap H^+$, which in view of the fact that $X_1 \leq X_2$ and hence $S_1 \subset S_2$ (see, e.g., (3.5) and Lemma 3.3) is the same as $S_1 \cap H^+$. Since $S_1 \cap H^+ = X_1 \cap H^+$, we have

$$X_1 \cap H^+ \subset X \cap H^+,$$

which, by Lemma 3.3, is equivalent to $X_1 \leq X$. In the same way we show that $X \leq X_2$.

We turn next to the second statement of the theorem, concerning tight internal bounds. Since $X_1 \leq X_2$ and X_1 and X_2 are internal, $S_1 \subset S_2$ and $\bar{S}_2 \subset \bar{S}_1$ (Theorem 6.8 in [16]). Hence, in view of (2.61),

$$X_1 \cap X_2 = S_1 \cap \bar{S}_2.$$

Now $S_1 = S \cap H_0$ if and only if $X_1 = X_{0-}$ (Theorem 6.11 in [16]), in which case

$$X_1 \cap X_2 = S \cap \bar{S}_2 = X \cap \bar{S}_2.$$

But since $X_1 \cap X_2 \subset S_2$, this is the same as

$$X_1 \cap X_2 = X \cap \bar{S}_2 \cap S_2 = X \cap X_2.$$

The rest follows analogously. □

Proof of Theorem 3.4(ii). First suppose that $\xi \in X_1 \cap X_2 \subset X$, and let $t \geq 0$. Then, for $i = 1, 2$, $U_t\xi \in \bar{S}_i$; and therefore, since

$$(1 - \pi_t)U_t\xi \in H^+_{[0,t]} \subset \bar{S}_i$$

we have $\pi_t U_t \xi \in \bar{S}_i$, i.e.,

$$V_t(X \cap H_0)\xi \in \bar{S}_i \quad \text{for} \quad i = 1, 2. \tag{3.13}$$

A symmetric argument yields

$$\bar{V}_t(X \cap H_0)\xi \in S_i \quad \text{for} \quad i = 1, 2. \tag{3.14}$$

Now from (3.13) and (3.14) we have $G\xi \in \bar{S}_1 \cap \bar{S}_2$ and $\bar{G}\xi \in S_1 \cap S_2$. But the group property of Theorem 2.9 implies that $\bar{G} = -G$, so therefore

$$G\xi \in S_1 \cap S_2 \cap \bar{S}_1 \cap \bar{S}_2 = X_1 \cap X_2,$$

proving the invariance property (3.7).

Finally, we prove the converse statement on G-invariance. Thus, suppose that $Z \subset X \cap H_0$ is G-invariant. Then in view of the decomposition (2.60) of Lemma 2.8 and the fact that both $X \cap H^-$ and $X \cap H^+$ are G-invariant, there is a decomposition

$$Z = Z_s + Z_u \tag{3.15}$$

such that $Z_s \subset X \cap H^-$ is G_s-invariant and $Z_u \subset X \cap H^+$ is G_u-invariant (Theorem 2.9).

We show first that there is a one-to-one correspondence between G_u-invariant subspaces $Z_u \subset X \cap H^+$ and splitting subspaces $X_u \in \mathcal{X}_0$ such that $X_u \leq X$, under which $Z_u = X_u \cap H^+$ and $S_u = H^- + Z_u$. To this end, take $t \geq 0$ and recall that $e^{G_u} Z_u = \pi_+ U_t^* Z_u$; and therefore, since $(1 - \pi_+)U_t^* Z_u \subset H^-$, it follows that $G_u Z_u \subset Z_u$ is equivalent to

$$U_t^* \left(H^- + Z_u\right) \subset \left(H^- + Z_u\right),$$

because $U_t^* H^- \subset H^-$. Set $S_u := H^- + Z_u$ and $\bar{S}_u := H^+ \vee Z_u{}^\perp$. Then $X_u \sim (S_u, \bar{S}_u)$ belongs to $\mathcal{X}_0$. (See the discussion in §2 and [15] or [16].) Since $S_u \sim (S_u, H_0)$ is itself a splitting subspace, Lemma 2.8 yields

$$S_u = H^- + \left(X_u \cap H^+\right) \tag{3.16}$$

for $S_u \cap H^- = H^-$ and $S_u \cap H^+ = X_u \cap H^+$. Hence we must have

$$Z_u = X_u \cap H^+;$$

and since $Z_u \subset X$, we have $X_u \cap H^+ \subset X \cap H^+$, from which we see that $X_u \leq X$ (Lemma 3.3). Consequently we have established the required one-to-one correspondence between G_u-invariant $Z_u \subset X \cap H^+$ and $X_u \in \mathcal{X}_0$ such that $X_u \leq X$.

In the same way we prove the symmetric statement that there is a one-to-one correspondence between G_s-invariant subspaces $Z_s \subset X \cap H^-$ and $X_s \in \mathcal{X}_0$ such that $X_s \geq X$, under which $Z_s = X_s \cap H^-$ and $S_s = H^+ + Z_s$.

Now, returning to the decomposition (3.15), we have shown that there are splitting subspaces $X_1, X_2 \in \mathcal{X}_0$ such that $X_1 \leq X \leq X_2$ and

$$Z = \left(X_1 \cap H^-\right) + \left(X_2 \cap H^+\right).$$

Let $\tilde{X}$ be an abitrary element in $\mathcal{X}$ having X_1 and X_2 as tightest internal bounds. Then by Lemma 3.3,

$$Z = \left(\tilde{X} \cap H^-\right) + \left(\tilde{X} \cap H^+\right),$$

i.e., $Z = \tilde{X} \cap H_0$ (Lemma 2.8). Proposition 3.7(iii) then yields $Z = X_1 \cap X_2$, proving the last statement of the theorem. □

COROLLARY 3.11. *Let $X \in \mathcal{X}$ and $X_0 \in \mathcal{X}_0$ be arbitrary, and let G be the zero generator of X. Then*

$$G\left[X \cap X_0\right] \subset X \cap X_0.$$

Conversely, any G-invariant subspace Z can be written $Z = \tilde{X} \cap X_0$ where $\tilde{X} \in \mathcal{X}$, $X_0 \in \mathcal{X}_0$, and X_0 is either the tightest upper or tightest lower internal bound of $\tilde{X}$.

Proof. Take $\xi \in X \cap X_0$ and $t \geq 0$. Then by the same procedure as in the proof of Theorem 3.4, $V_t(X \cap H_0)\xi \in \bar{S}_0$ and $\bar{V}_t(X \cap H_0)\xi \in S_0$, i.e., $G\xi \in \bar{S}_0$ and $-G\xi \in S_0$, and consequently $G\xi \in S_0 \cap \bar{S}_0 = X_0$. But, by definition, $G\xi \in X \cap H_0 \subset X$, and therefore $G\xi \in X \cap X_0$. This proves the required invariance. The inverse statement follows from the proof of Theorem 3.4. In fact, Z can be written as $Z = X_1 \cap X_2$ where X_1 and X_2 are tight internal bounds of $\tilde{X} \in \mathcal{X}$. Then from the last statement of Theorem 3.4(i), $Z = \tilde{X} \cap X_1 = \tilde{X} \cap X_2$. □

Proof of Corollary 3.10. Just noting that $G_s = G_-$ for $X = X_-$, $G_u = G_+$ for $X = X_+$, and $X_- \leq X \leq X_+$, the statements of the corollary are seen to be special cases of the corresponding results in the proof of Theorem 3.4. □

4. Introducing coordinates.

In this section we shall, among other things, reformulate the geometric results of §3 in the dual deterministic setting of linear functionals of the state at time zero. This will lead to characterizations in terms of state covariances and will facilitate the application of some of these results to the algebraic Riccati inequality in §5.

To this end, we shall now equip each $X \in \mathcal{X}$ with a basis chosen uniformly over the family $\mathcal{X}$, in a way first suggested in [5]. Let $\{\xi_1, \xi_2, \dots, \xi_n\}$ be an arbitrary basis in X_+. Such a basis corresponds to a model (1.1) with a state process $\{x_+(t); t \in \mathbb{R}\}$ such that

$$x_+(0) = \begin{bmatrix} \xi_1 \\ \xi_2 \\ \vdots \\ \xi_n \end{bmatrix}.$$

(See, e.g., [16] for the construction.) Now, for an arbitrary $X \in \mathcal{X}$, we define

$$x_k(0) = E^X \xi_k, \qquad k = 1, 2, \dots. \tag{4.1}$$

This can be seen to be a basis in X, and $x(0)$ is the state vector at zero of a model (1.1) having the same A and C matrices as that of $x_+(0)$.

There are several reasons why this construction is the right one. First, if for each $X \in \mathcal{X}$ we define the state covariance

$$P = E\{x(0)x(0)'\}, \tag{4.2}$$

then it was shown in [16] that

$$X_1 \leq X_2 \iff P_1 \leq P_2 \tag{4.3}$$

(where, as before, $P_1 \leq P_2$ means that $P_2 - P_1$ it positive semidefinite). In particular, (3.1) corresponds to

$$P_- \leq P \leq P_+, \tag{4.4}$$

$\mathcal{X}$ to the solution set $\mathcal{P}$ of the algebraic Riccati inequality $\Lambda(P) \leq 0$, and $\mathcal{X}_0$ to the subfamily $\mathcal{P}_0$ of solutions of the algebraic Riccati equation $\Lambda(P) = 0$, thus connecting the geometric theory of stochastic realization with that of Anderson [1] and Faurre et al. [9].

Secondly, the above family of bases is consistent in the sense that representations coincide on intersecting splitting subspaces as explained in the following lemma.

LEMMA 4.1. *Let $X_1, X_2 \in \mathcal{X}$. Then if $\lambda \in X_1 \cap X_2$, there is a unique $a \in \mathbb{R}^n$ such that*

$$\lambda = a'x_1(0) = a'x_2(0)$$

where $x_1(0)$ and $x_2(0)$ are bases of X_1 and X_2 respectively constructed as in (4.1).

Proof. Suppose that $\lambda = a_1'x_1(0) = a_2'x_2(0)$. Then by Theorem 6.12 in [16],

$$E^{X-}\lambda = a_1'x_-(0) = a_2'x_-(0);$$

and hence we must have $a_1 = a_2$, as claimed. □

Next we shall give a result that will be instrumental in establishing the correspondence between families of output-induced subspaces and covariance matrices P. To this end, given $X \in \mathcal{X}$ and the corresponding basis (4.1), define the linear map $T : \mathbb{R}^n \to X$ as

$$Ta = a'x(0). \tag{4.5}$$

This is a natural extension to $\mathbb{R}^n$ of the map T defined in §2. Clearly T is a bijection, and in view of Lemma 4.1,

$$T_1^{-1}|_{X_1 \cap X_2} = T_2^{-1}|_{X_1 \cap X_2} \tag{4.6}$$

if T_1 corresponds to X_1 and T_2 corresponds to X_2; hence, with some care, we may simply write T^{-1} whenever there is no risk for misunderstanding.

LEMMA 4.2. *Let $X_1, X_2 \in \mathcal{X}$ and $X_1 \leq X_2$, and let at least one of X_1 and X_2 be internal. Then*

$$T^{-1}(X_1 \cap X_2) = \ker(P_2 - P_1)$$

where P_1 and P_2 are the covariances corresponding to X_1 and X_2 respectively.

Proof. Let $\lambda \in X_1 \cap X_2$ and $T^{-1}(\lambda) = a$. Then $a'x_1(0) = a'x_2(0)$, and therefore

$$a'(P_2 - P_1)a = 0, \tag{4.7}$$

and therefore $a \in \ker(P_2 - P_1)$. Conversely, suppose that

$$a \in \ker(P_2 - P_1). \tag{4.8}$$

Since $X_1 \leq X_2$ and at least one of X_1 and X_2 is internal,

$$a'x_1(0) = E^{X_1}a'x_2(0)$$

(Proposition 6.12 in [16]), i.e., $[a'x_2(0) - a'x_1(0)] \perp a'x_1(0)$. Therefore, since

$$a'x_2(0) = [a'x_2(0) - a'x_1(0)] + a'x_1(0),$$

we have

$$E\left|a'x_2(0) - a'x_1(0)\right|^2 = a'(P_2 - P_1)a.$$

Consequently, by (4.8), $a'x_2(0) = a'x_1(0) \in X_1 \cap X_2$, i.e., $a \in T^{-1}(X_1 \cap X_2)$. □

We are now in a position to reformulate the first part of Theorem 3.4 in terms of covariances, thus obtaining an amplification of Theorem 9.1 and Lemma 9.3 in [16]. In the parameterization $\mathcal{P}$ of $\mathcal{X}$ the tightest internal bounds X_{0-} and X_{0+} of $X \in \mathcal{X}$ will be denoted P_{0-} and P_{0+} respectively. Recall that (P_{0-}, P_{0+}) denotes the *open tightest frame* of P, i.e., the set of all $P \in \mathcal{P}$ having P_{0-} and P_{0+} as their tightest upper and lower bounds in $\mathcal{P}_0$.

THEOREM 4.3. *Let $P_1, P_2 \in \mathcal{P}_0$ and $P \in \mathcal{P}$. Then*

(i) $P_1 \leq P \leq P_2 \Longleftrightarrow \ker(P_2 - P_1) \subset \ker(P_2 - P)$
with $\ker(P_2 - P_1) = \ker(P_2 - P)$ *if and only if* $P_1 = P_{0-}$; *and*

(ii) $P_1 \leq P \leq P_2 \Longleftrightarrow \ker(P_2 - P_1) \subset \ker(P - P_1)$
with $\ker(P_2 - P_1) = \ker(P - P_1)$ *if and only if* $P_2 = P_{0+}$.

Proof. Let $T : \mathbb{R}^n \to X$ be the bijection defined above, i.e., $T(a) = a'x(0)$. If $X_1 \leq X \leq X_2$, then $X_1 \cap X_2 \subset X$ by Theorem 3.4. Hence Lemma 4.2 can be applied with the same T^{-1} so that $X_1 \cap X_2$, $X \cap X_2$, and $X \cap X_1$ correspond to $\ker(P_2 - P_1)$, $\ker(P_2 - P)$, and $\ker(P - P_1)$ respectively under the bijection. Therefore,

$$\ker(P_2 - P_1) \subset \ker(P_2 - P) \cap \ker(P - P_1). \tag{4.9}$$

Also $X_2 = X_{0+}$ if and only if $X_1 \cap X_2 = X \cap X_2$, i.e., $\ker(P_2 - P_1) = \ker(P_2 - P)$. To prove the converse statement observe that any element $\xi \in X_1 \cap X_2$ can be written in the form $\xi = a'x_1(0) = a'x_2(0)$, where $a \in \ker(P_2 - P_1)$. So if $\ker(P_2 - P_1) \subset \ker(P_2 - P)$, then $a \in \ker(P_2 - P)$, i.e., $a'x_2(0) = a'x(0)$; therefore $\xi \in X$, which implies that $X_1 \cap X_2 \subset X$, which is equivalent to $X_1 \leq X \leq X_2$ by Theorem 3.4. This proves (i). Statement (ii) is proved in the same way. □

We shall now provide an explicit representation of $\mathcal{V}^*$ and its Γ'-invariant subspaces $\mathcal{V}$ in terms of covariance matrices.

As pointed out in §1, the set $\mathcal{P}$ is a parametrization of the family $\mathcal{X}$ of minimal Markovian splitting subspaces. In fact, a uniform choice of bases produces a unique state process x for each $X \in \mathcal{X}$ and hence a unique $P := E\{x(0)x(0)'\}$. Modulo orthogonal transformations in the input space, there is a unique minimal stochastic realization (1.9) corresponding to x that may be written in standard form

$$(\Sigma) \quad \begin{cases} dx = Axdt + B_1dw_1 + B_2dw_2, \\ dy = Cxdt + R^{1/2}dw_1. \end{cases} \tag{4.10}$$

A uniform choice of bases also fixes the matrices A and C to be the same for all $X \in \mathcal{X}$. Conversely, for each $P \in \mathcal{P}$, we have a minimal spectral factor

$$W(s) = C(sI - A)^{-1}(B_1, B_2) + (R^{1/2}, 0)$$

where

$$B_1 = (\bar{C} - CP)'R^{-1/2} \tag{4.11}$$

and B_2 is a full-rank factor of $-\Lambda(P)$, i.e.,

$$\Lambda(P) = -B_2 B_2', \tag{4.12}$$

and (in a suitable Hilbert space $\hat{H}$ as discussed in the beginning of §3) a unique stochastic realization (4.10), in turn defining a unique X.

Moreover, the uniform choice of bases associates to each $X \in \mathcal{X}$ a maximal output-nulling subspace $\mathcal{V}^* = \mathcal{V}^*(A', C', B', D')$ of the dual system (2.13) and a feedback matrix

$$\Gamma = A - B_1 R^{-1/2} C'. \tag{4.13}$$

We recall that $\mathcal{V}^* = \langle \Gamma \mid B_2 \rangle^{\perp}$. As explained in the proof of Lemma 2.3, (2.36), $\mathcal{V}^*$ can be decomposed into a direct sum

$$\mathcal{V}^* = \mathcal{V}^*_- + \mathcal{V}^*_+$$

of Γ'-invariant subspaces, $\mathcal{V}^*_-$ and $\mathcal{V}^*_+$, coresponding to the stable and the antistable modes of $\Gamma' \,|_{\mathcal{V}^*}$ respectively.

LEMMA 4.4. *Let $P \in \mathcal{P}$, and let $\mathcal{V}^*$ be the corresponding output-nulling subspace. Then*

(i) $\mathcal{V}^* = \ker(P - P_{0-}) = \ker(P_{0+} - P) = \ker(P_{0+} - P_{0-})$,
(ii) $\mathcal{V}^*_- = \ker(P - P_-) = \ker(P_{0+} - P_-)$,
(iii) $\mathcal{V}^*_+ = \ker(P_+ - P) = \ker(P_+ - P_{0-})$.

Proof. In view of Lemma 2.3 and (2.74), $\mathcal{V}^* = T^{-1}(X \cap H_0)$, $\mathcal{V}^*_- = T^{-1}(X \cap H^-)$, and $\mathcal{V}^*_+ = T^{-1}(X \cap H^+)$. Then applying Lemma 4.2 to Proposition 3.7 yields the desired result. □

Consider two covariance matrices P_1 and P_2 in $\mathcal{P}$ such that $P_1 \leq P_2$. We shall next establish the relation between the corresponding pairs of output-nulling subspaces $(\mathcal{V}^*_-)_1$, $(\mathcal{V}^*_+)_1$ and $(\mathcal{V}^*_-)_2$, $(\mathcal{V}^*_+)_2$ and the corresponding feedback matrices (4.13), Γ_1 and Γ_2. The following chain of results provides dual versions of Lemma 3.1, Corollary 3.2, and Lemma 3.3 in §3.

LEMMA 4.5. *Let at least one of $P_1, P_2 \in \mathcal{P}$ belong to $\mathcal{P}_0$, and suppose that $P_1 \leq P_2$. Then*

(i) $(\mathcal{V}^*_+)_1 \subset (\mathcal{V}^*_+)_2$ and $(\mathcal{V}^*_-)_2 \subset (\mathcal{V}^*_-)_1$,
(ii) $\Gamma'_1 \,|_{(\mathcal{V}^*_-)_2} = \Gamma'_2 \,|_{(\mathcal{V}^*_-)_2}$,
(iii) $\Gamma'_1 \,|_{(\mathcal{V}^*_+)_1} = \Gamma'_2 \,|_{(\mathcal{V}^*_+)_1}$.

Proof. The proof follows directly by applying Proposition 3.7 and Lemma 4.2 to Lemma 3.1. □

The following corollary illustrates the role of $\mathcal{V}^*_-$ and $\mathcal{V}^*_+$ as the stable and unstable Γ'-invariant subspaces of $\mathcal{V}^*$.

COROLLARY 4.6. *Let $P \in \mathcal{P}$, and let Γ be the corresponding feedback matrix* (4.13). *Then*

$$\Gamma' \,|_{\mathcal{V}^*_-} = \Gamma'_- \,|_{\mathcal{V}^*_-} \quad \text{and} \quad \Gamma' \,|_{\mathcal{V}^*_+} = \Gamma'_+ \,|_{\mathcal{V}^*_+},$$

where Γ_- and Γ_+ are the feedback matrices corresponding to P_- and P_+ respectively.

Proof. Take $P_1 = P_-$ and $P_2 = P$ in Lemma 4.5(ii) to prove (i). The second statement follows by setting $P_1 = P_+$ and $P_2 = P$ in Lemma 4.5(iii). □

LEMMA 4.7. *Let $P_1, P_2 \in \mathcal{P}_0$. Then for each $P \in \mathcal{P}$,*

(i) $P_1 \leq P \iff \ker(P_+ - P_1) \subset \ker(P_+ - P)$
with $\ker(P_+ - P_1) = \ker(P_+ - P)$ *if and only if* $P_1 = P_{0-}$; *and*
(ii) $P \leq P_2 \iff \ker(P_2 - P_-) \subset \ker(P - P_-)$
with $\ker(P_2 - P_-) = \ker(P - P_-)$ *if and only if* $P_2 = P_{0+}$.

In other words,
(i) $P_1 \leq P \iff (\mathcal{V}_+^*)_1 \subset \mathcal{V}_+^*$
with $(\mathcal{V}_+^*)_1 = \mathcal{V}_+^*$ *if and only if* $P_1 = P_{0-}$; *and*
(ii) $P \leq P_2 \iff (\mathcal{V}_-^*)_2 \subset \mathcal{V}_-^*$
with $(\mathcal{V}_-^*)_2 = \mathcal{V}_-^*$ *if and only if* $X_2 = X_{0+}$.

Proof. Follows immediately from Lemma 3.3. It is also a simple corollary of Theorem 4.3. □

The following theorem gives, for an arbitrary $P \in \mathcal{V}$, a complete characterization of all Γ'-invariant subspaces in $\mathcal{V}$, i.e., the output-nulling subspaces of the dual control system (2.13).

THEOREM 4.8. *Let* Γ *be the feedback matrix* (4.13) *corresponding to* $P \in \mathcal{P}$. *Then if* $P_1, P_2 \in \mathcal{P}_0$ *and* $P_1 \leq P \leq P_2$, *the subspace*

$$\ker(P_2 - P_1)$$

is Γ'*-invariant. Conversely, any* Γ'*-invariant subspace* $\mathcal{V} \subset \mathcal{V}^*$ *has a representation*

$$\mathcal{V} = \ker(P_2 - P_1)$$

for some $P_1, P_2 \in \mathcal{P}_0$ *such that* $P_1 \leq P \leq P_2$.

Proof. Follows by applying Lemma 4.2 to Theorem 3.4. □

Concerning Theorem 3.4, of which Theorem 4.8 is an isomorphic version, we may add that, thanks to Lemma 4.2, a simpler and more transparent proof of the invariances can be given. For example, to prove the G-invariance of $X_1 \cap X_2$ in Theorem 3.4, take $\xi \in X_1 \cap X_2$. Then, by Lemma 4.1, there is an $a \in \mathbb{R}^n$ such that

$$\xi = a'x_1(0) = a'x_2(0).$$

Since X_1 and X_2 are internal, the corresponding B_2-matrices are zero; i.e., for $t \geq 0$ and $i = 1, 2$,

$$U_t\xi = a'e^{\Gamma_i t}x_i(0) + \int_0^t a'e^{\Gamma_i(t-s)}(B_1)_i R^{-1/2}\,dy(s),$$

and therefore

$$e^{Gt}\xi = \pi_t U_t \xi = a'e^{\Gamma_1 t}x_1(0) = a'e^{\Gamma_2 t}x_2(0) \in X_1 \cap X_2.$$

Consequently $X_1 \cap X_2$ is G-invariant.

In the same way as above we obtain from Corollary 3.5 the following result characterizing intersecting zero dynamics.

LEMMA 4.9. *Let at least one of* $P_1, P_2 \in \mathcal{P}$ *belong to* $\mathcal{P}_0$, *and suppose that* $P_1 \leq P_2$. *Then*

$$\Gamma_1' \,|_{\ker(P_2 - P_1)} = \Gamma_2' \,|_{\ker(P_2 - P_1)}.$$

An important consequence of this lemma and the fact that $\mathcal{V}^*$ is constant over the open tightest frame (P_{0-}, P_{0+}) (Lemma 4.4) is that the zero dynamics is the same for all $P \in (P_{0-}, P_{0+})$. In fact, by Lemma 4.9,

$$\Gamma_{0-}' \,|_{\ker(P - P_{0-})} = \Gamma' \,|_{\ker(P - P_{0-})}$$

and, by Lemma 4.4, $\ker(P - P_{0-}) = \ker(P_{0+} - P_{0-})$.

The next proposition, which is due to Molinari [20] (also see [16, Lemma 10.2]), also belongs to the general area of invariance results described in this section and corresponds to Corollary 3.11.

PROPOSITION 4.10. *Let $P \in \mathcal{P}$ and $P_0 \in \mathcal{P}_0$ be arbitrary. Then all subspaces of the form*

$$\mathcal{V} = \ker(P - P_0)$$

are Γ'-invariant subspaces of $\mathcal{V}^*$.

5. Invariant subspaces and the algebraic Riccati inequality. In this section we shall generalize the well-known Potter-MacFarlane characterization of the (symmetric) solutions of the algebraic Riccati equation

$$\Lambda(P) = 0, \tag{5.1}$$

in terms of subspaces invariant under the Hamiltonian matrix, to the algebraic Riccati inequality

$$\Lambda(P) \leq 0. \tag{5.2}$$

Setting

$$F := A - \bar{C}' R^{-1} C, \tag{5.3}$$

we may write

$$\Lambda(P) = FP + PF' + PC'R^{-1}CP + \bar{C}'R^{-1}\bar{C}, \tag{5.4}$$

which corresponds to the *Hamiltonian* matrix

$$\mathcal{H} = \begin{bmatrix} F' & C'R^{-1}C \\ -\bar{C}'R^{-1}\bar{C} & -F \end{bmatrix}. \tag{5.5}$$

It is well known [17], [22], [18] that the solution set $\mathcal{P}_0$ of the algebraic Riccati equation is in a one-to-one correspondence with the class of Lagrangian $\mathcal{H}$-invariant subspaces $\mathcal{L}$ of $\mathbb{R}^{2n}$. Recall that a subspace $\mathcal{L}$ is Lagrangian if it is *isotropic* in the sense that if $x, y \in \mathcal{L}$, then

$$x' \begin{bmatrix} 0 & I \\ -I & 0 \end{bmatrix} y = 0 \tag{5.6}$$

and it is of maximal dimension n. Under this correspondence $\mathcal{L} = \operatorname{Im}\left[\begin{smallmatrix} I \\ P \end{smallmatrix}\right]$. The purpose of this section is to show that a similar correspondence holds for the solution set $\mathcal{P}$ of the algebraic Riccati inequality (5.2) and that this correspondence is related to the zero structure described above. In this respect a crucial observation is the following.

PROPOSITION 5.1. *Let $P \in \mathcal{P}$, and let $\mathcal{V}^*$ be the maximal output-nulling subspace of the corresponding dual system* (2.13). *Then $\mathcal{V}^*$ is the largest Γ'-invariant subspace of $\mathbb{R}^n$ such that*

$$\Lambda(P)\,|_{\mathcal{V}^*} = 0 \tag{5.7}$$

where Γ *is defined by* (2.18) *or, equivalently,*

$$\Gamma = F + PC^{'}R^{-1}C. \tag{5.8}$$

Proof. In view of (2.21), $\mathcal{V}^*$ is the largest $\Gamma^{'}$-invariant subspace orthogonal to the columns of B_2; and consequently, since $\Lambda(P) = -B_2B_2^{'}$, $\mathcal{V}^*$ is the largest $\Gamma^{'}$-invariant subspace for which (5.7) holds. □

Now, recall from §4 that to each $P \in \mathcal{P}$ there is a direct-sum decomposition

$$\ker(P - P_-) + \ker(P_+ - P) = \ker(P_{0+} - P_{0-}) \tag{5.9}$$

where P_{0-}, $P_{0+} \in \mathcal{P}_0$ are the tightest lower and upper internal bounds of P. In view of Lemma 4.4, this is equivalent to

$$\mathcal{V}^* = \mathcal{V}_-^* + \mathcal{V}_+^*. \tag{5.10}$$

As we have seen in §4 $\mathcal{V}_-^*$ is $\Gamma^{'}_-$-invariant and $\mathcal{V}_+^*$ is $\Gamma^{'}_+$-invariant. Moreover, if $a \in \mathcal{V}_-^*$ and $b \in \mathcal{V}_+^*$, then $a^{'}(P_+ - P)b = a^{'}(P - P_-)b = 0$ and, consequently, $\mathcal{V}_-^*$ and $\mathcal{V}_+^*$ are $(P_+ - P_-)$-orthogonal, i.e.,

$$a'(P_+ - P_-)b = 0 \quad \text{for all } a \in \mathcal{V}_-^*, b \in \mathcal{V}_+^*. \tag{5.11}$$

In §4 (Lemma 4.4) we saw that $\mathcal{V}_-^* = \ker(P_{0+} - P_-)$ and $\mathcal{V}_+^* = \ker(P_+ - P_{0-})$, so decomposition (5.9) may also be written

$$\ker(P_{0+} - P_-) + \ker(P_+ - P_{0-}) = \ker(P_{0+} - P_{0-}), \tag{5.12}$$

only involving covariance matrices belonging to $\mathcal{P}_0$.

If P is a solution of the algebraic Riccati equation (5.1), i.e., $P \in \mathcal{P}_0$, then $P = P_{0-} = P_{0+}$ and both (5.9) and (5.12) reduce to the $(P_+ - P_-)$-orthogonal decomposition

$$\ker(P - P_-) + \ker(P_+ - P) = \mathbb{R}^n \tag{5.13}$$

of the whole $\mathbb{R}^n$. This corresponds to the situation studied by J. C. Willems [27]. To set up notation and make contact with the gemetric theory of splitting subspaces we shall here restate Willems's result.

To this end, let $X \in \mathcal{X}_0$ and consider the stochastic version of (5.13), namely,

$$X = X \cap X_- + X \cap X_+, \tag{5.14}$$

obtained via Lemma 4.2 or directly from Lemma 2.8 and Proposition 3.7. Applying the projectors π_- and π_+ of (2.75) to (5.14) shows that

$$\pi_- X = X \cap X_- \quad \text{and} \quad \pi_+ X = X \cap X_+,$$

which can be translated into $\mathbb{R}^n$ via the bijective map $T : \mathbb{R}^n \to X$ of (4.5) to yield

$$\operatorname{Im}\Pi_- = \ker(P - P_-) \quad \text{and} \quad \operatorname{Im}\Pi_+ = \ker(P_+ - P).$$

Here $\Pi_- : \mathbb{R}^n \to \mathbb{R}^n$ and $\Pi_+ : \mathbb{R}^n \to \mathbb{R}^n$ are complementary projection operators defined as $\Pi_- = T^{-1}\pi_-|_X T$ and $\Pi_+ = T^{-1}\pi_+|_X T$ respectively. Now take $a \in \mathbb{R}^n$

and form the projections $a_- := \Pi_- a$ and $a_+ := \Pi_+ a$. From (5.13) we see that $a = a_- + a_+$, $Pa_- = P_- a_-$, and $Pa_+ = P_+ a_+$ so that $Pa = P_- \Pi_- a + P_+ \Pi_+ a$ for all $a \in \mathbb{R}^n$. Consequently,

$$P = P_- \Pi_- + P_+ \Pi_+.$$

LEMMA 5.2. (J. C. Willems). *Let Γ_- and Γ_+ be the feedback matrices, given by* (4.13) *and* (4.11)*, corresponding to P_- and P_+ respectively. Then:*

(i) *There is a one to one correspondence between Γ_-'-invariant subspaces $\mathcal{V}_- \subset \mathbb{R}^n$ and $P \in \mathcal{P}_0$ under which*

$$\mathcal{V}_- = \ker(P - P_-) \tag{5.15}$$

and

$$P = P_- \Pi_- + P_+(I - \Pi_-), \tag{5.16}$$

where Π_- is the $(P_+ - P_-)$-orthogonal projector of $\mathbb{R}^n$ onto $\mathcal{V}_-$.

(ii) *Dually, there is a one-to-one correspondence between Γ_+'-invariant subspaces $\mathcal{V}_+ \subset \mathbb{R}^n$ and $P \in \mathcal{P}_0$ under which*

$$\mathcal{V}_+ = \ker(P_+ - P) \tag{5.17}$$

and

$$P = P_-(I - \Pi_+) + P_+ \Pi_+, \tag{5.18}$$

where Π_+ is the $(P_+ - P_-)$-orthogonal projector of $\mathbb{R}^n$ onto $\mathcal{V}_+$.

Proof. By Lemma 4.2, $\mathcal{V}_-$ corresponds to $Z = X \cap X_-$ and $\mathcal{V}_+$ to $Z = X \cap X_+$ in Corollary 3.10. Moreover, Γ_- and Γ_+ correspond to G_- and G_+ respectively, and therefore the lemma follows. □

In summary, by Lemma 5.2, any $P \in \mathcal{P}_0$ corresponds to two subspaces, $\mathcal{V}_-^* = \ker(P - P_-)$, invariant for Γ_-', and $\mathcal{V}_+^* = \ker(P_+ - P)$, invariant for Γ_+', which by (5.13) are complementary, i.e., sum to all of $\mathbb{R}^n$. If $P \in \mathcal{P}$ does not belong to $\mathcal{P}_0$, however, (5.13) is replaced by (5.9). Therefore, if we insist on representing the invariant subspaces $\mathcal{V}_-^*$ and $\mathcal{V}_+^*$ in terms of solutions of the algebraic Riccati equation, as stated in Lemma 5.2, then there will still be representations of the type $\mathcal{V}_-^* = \ker(P_0 - P_-)$ and $\mathcal{V}_+^* = \ker(P_+ - P_0)$, but now we can no longer use the same P_0. Formula (5.12) is precisely a manifestation of this fact.

The following notation will be used in the sequel. If $\mathcal{L}$ is a k-dimensional subspace of $\mathbb{R}^{2n}$ with basis matrix $L \in \mathbb{R}^{2n \times k}$, define $\tau(\mathcal{L})$ to be the subspace in $\mathbb{R}^n$ spanned by the truncated matrix obtained by removing the bottom n rows of L.

We are now in a position to state the main result of this section.

THEOREM 5.3. *Let $\mathcal{P}$ be the solution set of the matrix Riccati inequality* (5.2)*, and let $\mathcal{H}$ be the Hamiltonian matrix* (5.5)*. Then there is a one-to-one correspondence between the isotropic $\mathcal{H}$-invariant subspaces $\mathcal{L} \subset \mathbb{R}^{2n}$ of dimension $k \leq n$ and the family of open tightest frames (P_{0-}, P_{0+}) of $\mathcal{P}$. Under this correspondence*

$$\mathcal{L} = \begin{bmatrix} I \\ P \end{bmatrix} \mathcal{V}^* \tag{5.19}$$

for any $P \in (P_{0-}, P_{0+})$, where $\mathcal{V}^* \subset \mathbb{R}^n$ is the subspace of zero directions

$$\mathcal{V}^* = \ker(P_{0+} - P_{0-}) \tag{5.20}$$

and $k = \dim \mathcal{L}$ is the number of zeros of the spectral factor W corresponding to P. Conversely, given any isotropic $\mathcal{H}$-invariant subspace $\mathcal{L} \subset \mathbb{R}^{2n}$ of dimension $k \leq n$, the matrices P_{0-} and P_{0+} are obtained from Lemma 5.2, formulas (5.16) and (5.18), as the elements in $\mathcal{P}_0$ corresponding to the invariant subspaces $\mathcal{V}_- = \tau(\mathcal{L}_-)$ and $\mathcal{V}_+ = \tau(\mathcal{L}_+)$, where $\mathcal{L}_-$ and $\mathcal{L}_+$ are the subspaces of $\mathcal{L}$ consisting of sums of stable and antistable eigenspaces of $\mathcal{H}$.

Proof. First suppose that $P \in \mathcal{P}$ has the tightest local frame (P_{0-}, P_{0+}), and define $\mathcal{L}$ by (5.19) and (5.20). Clearly, (5.19) is independent of the choice of $P \in (P_{0-}, P_{0+})$. In fact, if $P_1, P_2 \in (P_{0-}, P_{0+})$, then by Lemma 4.4, $\mathcal{V}^* = \ker(P_1 - P_{0-}) = \ker(P_2 - P_{0-})$, and hence it follows that $(P_2 - P_1)a = 0$ for all $a \in \mathcal{V}^*$. Now a straightforward calculation, using (5.4) and the fact that $\Lambda(P)\mathcal{V}^* = 0$ (Proposition 5.1), shows that

$$\mathcal{H}\mathcal{L} = \begin{bmatrix} I \\ P \end{bmatrix} \Gamma' \mathcal{V}^*.$$

Since $\Gamma' \mathcal{V}^* \subset \mathcal{V}^*$, this yields $\mathcal{H}\mathcal{L} \subset \mathcal{L}$ as claimed. The fact that $P' = P$ ensures that $\mathcal{L}$ is isotropic.

Conversely, suppose that $\mathcal{L} \subset \mathbb{R}^{2n}$ is any $\mathcal{H}$-invariant isotropic subspace of dimension $k \leq n$. Then $\mathcal{L}$ is a direct sum of generalized eigenspaces of $\mathcal{H}$; and since these eigenspaces are contained in either $\operatorname{Im}\left[\begin{smallmatrix} I \\ P_- \end{smallmatrix}\right]$ or $\operatorname{Im}\left[\begin{smallmatrix} I \\ P_+ \end{smallmatrix}\right]$ (for $\mathbb{R}^{2n}$ is a direct sum of these subspaces), we have the direct sum decomposition

$$\mathcal{L} = \mathcal{L}_- + \mathcal{L}_+ \tag{5.21}$$

where $\mathcal{L}_- := \mathcal{L} \cap \operatorname{Im}\left[\begin{smallmatrix} I \\ P_- \end{smallmatrix}\right]$ and $\mathcal{L}_+ := \mathcal{L} \cap \operatorname{Im}\left[\begin{smallmatrix} I \\ P_+ \end{smallmatrix}\right]$ are both $\mathcal{H}$-invariant, because $\operatorname{Im}\left[\begin{smallmatrix} I \\ P_- \end{smallmatrix}\right]$ and $\operatorname{Im}\left[\begin{smallmatrix} I \\ P_+ \end{smallmatrix}\right]$ are. Therefore, there are full-rank matrices M_- and M_+ such that

$$\mathcal{L}_- = \operatorname{Im} \begin{bmatrix} I \\ P_- \end{bmatrix} M_- \quad \text{and} \quad \mathcal{L}_+ = \operatorname{Im} \begin{bmatrix} I \\ P_+ \end{bmatrix} M_+. \tag{5.22}$$

But $\operatorname{Im}\left[\begin{smallmatrix} I \\ P_- \end{smallmatrix}\right]$ is $\mathcal{H}$-invariant and

$$\mathcal{H} \begin{bmatrix} I \\ P_- \end{bmatrix} = \begin{bmatrix} I \\ P_- \end{bmatrix} \Gamma'_-,$$

and consequently

$$\mathcal{H} \begin{bmatrix} I \\ P_- \end{bmatrix} M_- = \begin{bmatrix} I \\ P_- \end{bmatrix} \Gamma'_- M_-.$$

Therefore, since $\mathcal{L}_-$, represented by (5.22), is $\mathcal{H}$-invariant, $\operatorname{Im} M_-$ must be Γ'_--invariant. In the same way we show that $\operatorname{Im} M_+$ is Γ'_+- invariant. Consequently, it follows from Lemma 5.2 that there are unique $P_{0-}, P_{0+} \in \mathcal{P}_0$ so that

$$\mathcal{V}_- := \operatorname{Im} M_- = \ker(P_{0+} - P_-) \tag{5.23}$$

and

$$\mathcal{V}_+ := \operatorname{Im} M_+ = \ker(P_+ - P_{0-}). \tag{5.24}$$

It remains to show that $P_{0-} \leq P_{0+}$ so that (P_{0-}, P_{0+}) may form a tightest local frame and we may identify $\mathcal{V}_-$ and $\mathcal{V}_+$ with $\mathcal{V}_-^*$ and $\mathcal{V}_+^*$ respectively. To this end, note that since

$$\mathcal{L} = \operatorname{Im} \begin{bmatrix} M_- & M_+ \\ P_-M_- & P_+M_+ \end{bmatrix}$$

is isotropic,

$$\begin{bmatrix} M_- & M_+ \\ P_-M_- & P_+M_+ \end{bmatrix}' \begin{bmatrix} 0 & I \\ -I & 0 \end{bmatrix} \begin{bmatrix} M_- & M_+ \\ P_-M_- & P_+M_+ \end{bmatrix} = 0,$$

i.e.,

$$M_-'(P_+ - P_-)M_+ = 0.$$

Consequently, $\mathcal{V}_-$ and $\mathcal{V}_+$ are $(P_+ - P_-)$-orthogonal. In other words,

$$\mathcal{V}_+ \subset (\mathcal{V}_-)^\circ \tag{5.25}$$

where $^\circ$ denotes the $(P_+ - P_-)$-orthogonal complement in $\mathbb{R}^n$. Now, in view of (5.23) and decomposition (5.13),

$$(\mathcal{V}_-)^\circ = \ker(P_+ - P_{0+}). \tag{5.26}$$

Therefore,

$$\ker(P_+ - P_{0-}) = \mathcal{V}_+ \subset (\mathcal{V}_-)^\circ = \ker(P_+ - P_{0+}),$$

so it follows from Lemma 4.7 that $P_{0-} \leq P_{0+}$, as claimed.

Now, let $P \in \mathcal{P}$ be an arbitrary element in the open tightest frame (P_{0-}, P_{0+}). Then by (5.12), (5.23), and (5.24),

$$\mathcal{V} := \mathcal{V}_- + \mathcal{V}_+ = \ker(P_{0+} - P_{0-}),$$

and hence, by Lemma 4.4, $\mathcal{V} = \mathcal{V}^*$, the space of zero directions corresponding to P. Moreover, $\mathcal{V}_-$ and $\mathcal{V}_+$ are actually $\mathcal{V}_-^*$ respectively $\mathcal{V}_+^*$. □

Theorem 5.3 is a generalization of the well-known result linking solutions in $\mathcal{P}_0$ to $\mathcal{H}$-invariant Lagrangian subspaces [17], [22], [18], in which special situation the equivalence classes of Theorem 5.3 are singletons and the invariant subspaces are n-dimensional. The fewer zeros that the spectral factor corresponding to P has, the larger is the equivalence class (the tightest local frame) and the smaller is the dimension of the invariant subspace $\mathcal{L}$.

Appendix. In this appendix we shall give the proofs deferred from §2.

Proof of Proposition 2.1. Suppose that

$$U_tY \subset Y \vee H^+_{[0,t]}. \tag{A.1}$$

Let $\xi \in Y$. Then

$$U_t\xi = \lambda_t + \eta_t \tag{A.2}$$

where $\lambda_t \in Y$ and $\eta_t \in H^+_{[0,t]}$. Since $Y \subset X \cap H_0$, we must have $\lambda_t = e^{Gt}\xi$; and therefore, applying the orthogonal projector E^X to (A.2), we obtain

$$e^{Ft}\xi = e^{Gt}\xi + E^X\eta_t. \tag{A.3}$$

Hence, for $t > 0$,

$$\frac{1}{t}(e^{Ft} - I)\xi = \frac{1}{t}(e^{Gt} - I)\xi + \frac{1}{t}E^X\eta_t, \tag{A.4}$$

and consequently

$$\lim_{t\downarrow 0} \frac{1}{t}E^X\eta_t \tag{A.5}$$

exists and, by the definition (2.25) of the operator N, must belong to $\operatorname{Im} N$. Therefore, since $e^{Gt}\xi \in Y$ (Proposition 2.7), we have $\xi \in Y \vee \operatorname{Im} N$, i.e.,

$$FY \subset Y \vee \operatorname{Im} N \tag{A.6}$$

as claimed. □

Remark. Let $\nu \in \operatorname{Im} N$ be the limit (A.5). Then, from (A.5), we see that

$$\nu = (F - G)\xi$$

is a linear function of ξ, and therefore there is a map $L : Y \to \mathbb{R}^m$ such that $\nu = NL\xi$; consequently,

$$G = F - NL.$$

Proof of Lemma 2.4. If $\xi \in X \cap H_0$, then $\xi = a'x(0)$ with $a \in \mathcal{V}^*$. Therefore it follows immediately from (2.43) and the fact that $\mathcal{V}^*$ is Γ'-invariant that for $t \geq 0$,

$$U_t\xi \in X \cap H_0 \vee H^+_{[0,t]}$$

and

$$U_t^*\xi \in X \cap H_0 \vee H^-_{[-t,0]},$$

so it only remains to show that these vector sums are direct, i.e., that $X \cap H^+_{[0,t]} = 0$ and $X \cap H^-_{[-t,0]} = 0$.

By stationarity $X \cap H^-_{[-t,0]} = 0$ if and only if $(U_tX) \cap H^+_{[0,t]} = 0$. To prove the latter, suppose $\eta \in (U_tX) \cap H^+_{[0,t]}$. We want to prove that η must be zero. To this end, note that

$$\hat{\eta} := E^{H^+_{[0,t]}}\eta = \eta. \tag{A.7}$$

Now there is an $a \in \mathbb{R}^n$ such that $\eta = a'x(t)$ and hence

$$\hat{\eta} = a'\hat{x}(t),$$

where $\hat{x}(t)$ is the Kalman-filter estimate. It is well known that $\Pi(t) := E\{\hat{x}(t)\hat{x}(t)'\}$ satisfies the Riccati differential equation

$$\dot{\Pi} = \Lambda(\Pi), \quad \Pi(0) = 0,$$

which has the limit P_- as $t \to \infty$; see, e.g., [9] or [16]. It is now easy to see that $Q := P_- - \Pi$ satisfies the homogeneous Riccati equation

$$\dot{Q} = \Gamma_- Q + Q\Gamma_-' - QC'R^{-1}CQ, \quad Q(0) = P_- > 0.$$

Since $Q(0) > 0$, $M(t) = Q(t)^{-1}$ exists on some finite interval $[0, t_1]$ and it is readily seen that it satisfies the Lyapunov differential equation

$$\dot{M} = -M\Gamma_- - \Gamma_-' M + C'R^{-1}C, \quad M(0) = P_-^{-1} > 0$$

there. Integrating we obtain

$$M(t) = e^{-\Gamma_-' t} M(0) e^{-\Gamma_- t} + \int_0^t e^{-\Gamma_-'(t-s)} C'R^{-1}C e^{-\Gamma_-(t-s)} \, ds$$

where the first term is positive definite and the second nonnegative definite. Consequently, $M(t) > 0$ for all finite t and hence $Q(t) > 0$ for all finite t.

Now, from (A.7) we have that

$$a' \left[P - \Pi(t) \right] a = 0.$$

But $P - \Pi(t) \geq P_- - \Pi(t) = Q > 0$. Hence $a = 0$, and therefore $\eta = 0$. The proof that $X \cap H_{[0,t]}^+ = 0$ follows from a symmetric argument. □

REFERENCES

[1] B. D. O. Anderson, *The inverse problem of stationary covariance generation*, J. Statist. Phys., 1 (1969), pp. 133–147.

[2] B. D. O. Anderson, *Algebraic properties of minimal degree spectral factors*, Automatica, 9 (1973), pp. 491–500.

[3] G. Basile and G. Marro, *Controlled and conditioned invariant subspaces in linear system theory*, J. Optim. Theory Appl., 3 (1973), pp. 306–316.

[4] C. I. Byrnes and A. Isidori, *A frequency domain philosophy for nonlinear systems, with applications to stabilization and to adaptive control*, in Proc. 23rd IEEE Conf. Decision and Control, Las Vegas, Nevada, 1984, pp. 1569–1573.

[5] P. E. Caines and D. Delchamps, *Splitting subspaces, spectral factorization and the positive real equation: structural features of the stochastic realization problem*, in Proc. 1980 IEEE Conf. Decision and Control, Albuquerque, New Mexico, 1980, pp. 358–362.

[6] P. E. Caines, *Linear Stochastic Systems*, Wiley, New York, 1988.

[7] C. A. Desoer and J. D. Schulman, *Zeros and poles of matrix transfer functions and their dynamical interpretations*, IEEE Trans. Circuits and Systems, CAS-21 (1974), pp. 3–8.

[8] H. Dym and H. P. McKean, *Gaussian Processes, Function Theory, and the Inverse Spectral Problem*, Academic Press, New York, 1976.

[9] P. Faurre, M. Clerget, and F. Germain, *Opérateurs Rationnels Positifs*, Dunod, Paris, 1979.

[10] L. E. Faibusovich, *Matrix Riccati inequality: Existence of solutions*, Systems Control Lett., 9 (1987), pp. 59–64.

[11] M. Green, *Balanced stochastic realizations*, Linear Algebra Appl., 98 (1983), pp. 369–402.

[12] M. L. J. Hautus and L. M. Silverman, *System structure and singular control*, Linear Algebra Appl., 50 (1974), pp. 3–8.

[13] V. Kučera, *Algebraic Riccati equation: Hermitian and definite solutions*, in The Riccati Equation, S. Bittanti, A. J. Laub, and J. C. Willems, eds., Springer-Verlag, New York, 1991, pp. 53–88.

[14] A. Lindquist and G. Picci, *Forward and backward semimartingale representations for stationary increment processes*, Stochastics, 15 (1985), pp. 1–50.

[15] ———, *Realization theory for multivariate stationary Gaussian processes*, SIAM J. Control Optim., 23 (1985), pp. 809–857.

[16] A. Lindquist and G. Picci, *A geometric approach to modelling and estimation of linear stochastic systems*, J. Math. Systems, Estimation, Control, 1 (1990), pp. 241–333.

[17] A. G. J. MacFarlane, *An eigenvector solution of the optimal linear regulator problem*, J. Electron. Control, 14 (1963), pp. 496–501.

[18] K. Mårtensson, *On the matrix Riccati equation*, Inform. Sci., 3 (1971), pp. 17–49.

[19] Gy. Michaletzky, *Zeros of (non-square) spectral factors and canonical correlations*, in Proc. 11th IFAC World Congress, Tallinn, Estonia, 1991, pp. 167–172; also in preprint collection: 1990, pp. 221–226.

[20] B. P. Molinari, *The time-invariant linear-quadratic optimal-control problem*, Automatica, 13 (1977), pp. 347–357.

[21] A. Isidori and C. Moog, *On the nonlinear equivalent of the notion transmission zeros*, in Modelling and Adaptive Control, C. I. Byrnes and A. Kurzhansky, eds., Springer-Verlag, New York, 1986.

[22] J. E. Potter, *Matrix quadratic solutions*, SIAM J. Appl. Math., 14 (1966), pp. 496–501.

[23] E. A. Robinson, *Properties of the Wold decomposition of stationary Gaussian stochastic processes*, Theory Probab. Appl., 8 (1963), pp. 187–195.

[24] Yu. A. Rosanov, *Stationary Random Processes*, Holden-Day, New York, 1963.

[25] C. Scherer, *The solution set of the algebraic Riccati equation and the algebraic Riccati inequality*, Linear Algebra Appl., 153 (1991), pp. 99–122.

[26] M. Shayman, *Geometry of the algebraic Riccati equation*, SIAM J. Control Optim., 21 (1983), pp. 375–394.

[27] J. C. Willems, *Least squares stationary optimal control and the algebraic Riccati equation*, IEEE Trans. Automat. Control, AC-16 (1971), pp. 621–634.

[28] W. M. Wonham, *Linear Multivariable Control: A Geometric Approach*, Springer-Verlag, New York, 1977.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 33, No. 2, pp. 402–418, March 1995

© 1995 Society for Industrial and Applied Mathematics
004

A GLOBALLY CONVERGENT SUCCESSIVE APPROXIMATION METHOD FOR SEVERELY NONSMOOTH EQUATIONS*

LIQUN QI† AND XIAOJUN CHEN†

Abstract. This paper presents a globally convergent successive approximation method for solving $F(x) = 0$ where F is a continuous function. At each step of the method, F is approximated by a smooth function f_k, with $\|f_k - F\| \to 0$ as $k \to \infty$. The direction $-f_k'(x_k)^{-1}F(x_k)$ is then used in a line search on a sum of squares objective. The approximate function f_k can be constructed for nonsmooth equations arising from variational inequalities, maximal monotone operator problems, nonlinear complementarity problems, and nonsmooth partial differential equations. Numerical examples are given to illustrate the method.

Key words. global convergence, successive approximation, integration convolution

AMS subject classifications. 90C30, 90C33

1. Introduction.

Let $F : R^n \to R^n$ be a continuous but not necessarily differentiable function. We consider the system of nonlinear equations

$$F(x) = 0, \qquad x \in R^n. \tag{1}$$

The recent literature of such nonsmooth equations includes [1]–[3], [6]–[8], [10]–[13], [15], [17], [19], [21].

If F is smooth, a popular method for solving (1) is the damped Newton method [4], [9]

$$\left\{ \begin{array}{ll} \text{Solve} & F(x_k) + F'(x_k)d = 0 \text{ to get } d_k \\ \text{Set} & x_{k+1} = x_k + \alpha_k d_k, \end{array} \right. \tag{2}$$

where the step size α_k in $(0, 1]$ is chosen by a line search.

Han, Pang, and Rangaraj [6] generalized the damped Newton method to solve the nonsmooth equation (1) using the idea of an "iteration function." Let $\theta : R^n \to R_+$ be defined by

$$\theta(x) = \frac{1}{2}F(x)^T F(x).$$

Damped Newton method with iteration function (IF). Let $\rho, \sigma \in (0, 1)$ be given. Let $G : R^{n\times n} \to R^n$ be a given iteration function. Let $x_0 \in R^n$ be arbitrary. Set $k = 0$.

$$\left\{ \begin{array}{ll} \text{Solve} & F(x_k) + G'(x_k, d) = 0 \text{ to get } d_k \\ \text{Set} & x_{k+1} = x_k + \rho^{m_k} d_k, \end{array} \right.$$

where m_k is the smallest nonnegative integer m such that

$$\theta(x_k + \rho^m d_k) - \theta(x_k) \le -2\sigma\rho^m\theta(x_k).$$

Global convergence was established in [6] under four assumptions on G and $F^T G$. In general, $G(x, \cdot)$ is nonlinear. This implies that a system of nonlinear equations (generally easier than (1)) is solved at each step in the above method.

* Received by the editors August 24, 1992; accepted for publication (in revised form) October 20, 1993. This research was supported by the Australian Research Council.

† School of Mathematics, University of New South Wales, Sydney, New South Wales 2052, Australia.

Recently, Gabriel and Pang [7] proposed a trust region algorithm using iteration functions. They also required certain assumptions on the iteration functions to establish convergence of their algorithm. Poliquin and Qi [14] proved that, in the case of nonsmooth optimization, the assumptions on the iteration functions actually implied restrictions on the original function. There are other globally convergent methods for nonsmooth equations [10]–[13], [20]. These methods either assume conditions much stronger than continuity or only work for some special problems.

In this paper, we introduce a successive approximation method. Let $||\cdot||$ denote the Euclidean norm. At the kth step, we approximate F by a smooth function f_k such that $F = f_k + g_k$, where

$$||g_k|| \equiv \sup\{||g_k(x)|| : x \in R^n\} \le \alpha ||F(x_k)||$$

and $\alpha \in (0,1)$ is a fixed constant. The algorithm uses $f'_k(x_k)$, wherever a derivative of F at x_k is needed.

There are two outstanding advantages of the new algorithm over existing methods. The first advantage is that a linear approximation is made at each step, so the subproblem is a system of linear equations. Known globally convergent methods for solving nonsmooth equations do not have this feature. The second advantage is that the conditions required to establish convergence and implement the new algorithm are very general. We establish global convergence of this algorithm under the following assumptions on F: continuity of F, boundedness of a level set, and nonsingularity of f'_k at x_k for all k and at x^*, an accumulation point of $\{x_k\}$. To implement our algorithm we require F to be locally Lipschitzian. Under these assumptions, we may construct f_k with the desired accuracy. The basic tool is the integration convolution. In some special cases, we have other ways to construct f_k.

Although we discuss the linear convergence of this algorithm in §3, we do not intend to pursue a higher rate of convergence for this method. There are already several superlinearly convergent methods [10], [11], [19–21] and a superlinear convergence theory [13], [15] for solving nonsmooth equations. One may construct a hybrid algorithm that is globally and superlinearly convergent using the new algorithm and a known superlinearly convergent algorithm with the methodology proposed in [15]. We do not go into the details of such a construction. The merit of our algorithm is that it may solve some severely nonsmooth equations, such as nonsmooth equations arising from the variational inequality problem for a general convex set and from the maximal monotone operator problem (see §5).

In §2 we describe the successive approximation method and prove its global convergence.

In §3 we consider the rate of convergence.

In §4 we discuss how to construct a successive approximation function for a nonsmooth function F using integration convolution.

In §5 we investigate some applications of our algorithm.

In §6 we give numerical results with the successive approximation method.

2. Method and global convergence.

DEFINITION 1. *Let $\alpha \in (0,1)$ be a constant. At the* kth *step of the iteration methods described in this and the next sections, we call*

$$F = f_k + g_k$$

a normal decomposition of F, *if* f_k *is smooth and* $\|g_k\| \leq \alpha\|F(x_k)\|$, *whenever* $F(x_k) \neq 0$.

We shall give some examples of normal decompositions in §5.

Let

$$\theta(x) = \frac{1}{2}F(x)^T F(x)$$

and

$$\theta_k(x) = \frac{1}{2}f_k(x)^T f_k(x).$$

Our method can be described as follows.

The successive approximation method (SAM). Given $\rho, \alpha \in (0,1)$, an initial vector $x_0 \in R^n$ and a normal decomposition $F = f_0 + g_0$ with $\|g_0\| \leq \frac{\alpha}{2}\|F(x_0)\|$, let $0 < \sigma < 1 - \alpha$. For $k \geq 0$:

1. Solve $F(x_k) + f'_k(x_k)d = 0$ to get d_k.

2. Set $x_{k+1} = x_k + \rho^{m_k}d_k$, where m_k is the smallest nonnegative integer m such that

$$\theta_k(x_k + \rho^m d_k) - \theta_k(x_k) \leq -2\sigma\rho^m\theta(x^k).$$

3. If $F(x_{k+1}) = 0$, stop. If $\|g_k\| < \alpha\|F(x_{k+1})\|$, we let $f_{k+1} = f_k$ and $g_{k+1} = g_k$. Otherwise, we construct a new normal decomposition

$$F = f_{k+1} + g_{k+1},$$

with $\|g_{k+1}\| \leq \min\{\frac{\alpha}{2}\|F(x_{k+1})\|, \frac{1}{2}\|g_k\|\}$.

Assumption 1. The level set

$$D_0 = \{x \in R^n : \theta(x) \leq (1+\alpha)^2\theta(x_0)\}$$

is bounded.

Assumption 2. $f'_k(x_k)$ are nonsingular for all k.

LEMMA 1. *Suppose that* $F(x_k) \neq 0$ *and* $F = f_k + g_k$ *is a normal decomposition of* F. *Then there exists a scalar* $t_k \in (0,1]$ *such that for all* $t \in (0, t_k]$

$$\theta_k(x_k + td_k) - \theta_k(x_k) \leq -2\sigma t\theta(x_k). \tag{3}$$

Proof. Notice that $\theta'_k(x_k) = f'_k(x_k)^T f_k(x_k)$ and $f'_k(x_k)d_k = -F(x_k)$. We have

$$\begin{aligned}\theta_k(x_k + td_k) - \theta_k(x_k) &= \frac{1}{2}(f_k(x_k + td_k)^T f_k(x_k + td_k) - f_k(x_k)^T f_k(x_k)) \\ &= td_k^T f'_k(x_k)^T f_k(x_k) + o(t) \\ &= -tF(x_k)^T F(x_k) + tF(x_k)^T g_k(x_k) + o(t).\end{aligned}$$

Since $\|g_k\| < \alpha\|F(x_k)\|$,

$$\begin{aligned}\theta_k(x_k + td_k) - \theta_k(x_k) &\leq -2t\theta(x_k) + t\|F(x_k)\|\|g_k(x_k)\| + o(t) \\ &< -2t\theta(x_k) + t\alpha\|F(x_k)\|\|F(x_k)\| + o(t) \\ &= -2t(1-\alpha)\theta(x_k) + o(t).\end{aligned}$$

Since $\sigma < 1 - \alpha$, there exists $t_k \in (0,1]$ such that for all $t \in (0, t_k]$, (3) holds. □

Lemma 1 indicates that the SAM is well defined under Assumption 2.

THEOREM 1. *Suppose that Assumptions* 1 *and* 2 *hold. Then the* SAM *is well defined and for all k*

$$x_k \in D_0. \tag{4}$$

Let $\{x_k\}$ be a sequence produced by the SAM. *If furthermore for an accumulation point x^* of $\{x_k\}, f_k'(x^*)$ is nonsingular for all large k, then*

$$\lim_{k\to\infty} F(x_k) = 0 \tag{5}$$

and

$$F(\tilde{x}) = 0$$

for all accumulation points $\tilde{x}$ of $\{x_k\}$.

Proof. Without loss of generality, we may assume that F is not smooth. Hence $\|g_k\| > 0$ for any k.

By Lemma 1, the SAM is well defined. We now prove (4). Without loss of generality, we assume that $F(x_k) \neq 0$ for all k. Let $K = \{0\} \cup \{k : \|g_{k-1}\| \geq \alpha\|F(x_k)\|\}$. Assume that K consists of $k_0 = 0 < k_1 < k_2 < \cdots$. Let k be an arbitrary nonnegative integer. Let k_j be the largest number in K such that $k_j \leq k$. Then

$$f_k = f_{k_j}, \qquad g_k = g_{k_j},$$

and

$$\begin{aligned}
\|F(x_k)\| &= \|f_k(x_k) + g_k(x_k)\| \\
&= \|f_{k_j}(x_k) + g_{k_j}(x_k)\| \\
&\leq \|f_{k_j}(x_k)\| + \|g_{k_j}(x_k)\| \\
&\leq \|f_{k_j}(x_{k_j})\| + \|g_{k_j}\| \\
&= \|F(x_{k_j}) - g_{k_j}(x_{k_j})\| + \|g_{k_j}\| \\
&\leq \|F(x_{k_j})\| + \|g_{k_j}(x_{k_j})\| + \|g_{k_j}\| \\
&\leq \|F(x_{k_j})\| + 2\|g_{k_j}\|.
\end{aligned}$$

If $j = 0$, then $\|F(x_k)\| \leq \|F(x_0)\| + \alpha\|F(x_0)\|$, since $\|g_0\| \leq \frac{\alpha}{2}\|F(x_0)\|$. If $j \geq 1$, then

$$\begin{aligned}
\|F(x_k)\| &\leq \|F(x_{k_j})\| + 2\|g_{k_j}\| \leq \frac{1}{\alpha}\|g_{k_j-1}\| + \|g_{k_{j-1}}\| \\
&\leq \left(\frac{1}{\alpha} + 1\right)\frac{1}{2^{j-1}}\|g_0\| \leq (1+\alpha)\frac{1}{2^{j-1}}\|F(x_0)\|.
\end{aligned} \tag{6}$$

In both cases it follows that $\theta(x_k) \leq (1+\alpha)^2\theta(x_0)$. This implies that (4) holds.

We now prove the second part of the theorem. If K is infinite, then for any $k \geq 0$ there exists $k_j \in K$, the largest number in K such that $k_j \leq k$ and (6) holds. The limit in the right-hand side of (6) is zero. This proves (5).

Hence, to prove (5), it suffices to prove that K is infinite. Suppose K is finite and assume $\hat{k} > k$ for all $k \in K$. Then $\|g_{\hat{k}}\| < \alpha\|F(x_k)\|$ for $k \geq \hat{k}$. Hence for all $k \geq \hat{k}$,

$$f_k \equiv f_{\hat{k}}, \quad g_k \equiv g_{\hat{k}}, \tag{7}$$

and

$$\theta(x_k) = \frac{1}{2}\|F(x_k)\|^2 > \frac{1}{2\alpha^2}\|g_{\hat{k}}\|^2 \equiv \hat{\epsilon} > 0. \tag{8}$$

Suppose that K_0 is a subsequence of $\{0, 1, \ldots\}$ such that $\{x_k : k \in K_0\}$ converges to x^*. By (7) and the condition of this theorem, $f'_{\hat{k}}(x^*)$ is nonsingular. Since $\lim_{k\to\infty, k\in K_0} x_k = x^*$ and $f'_{\hat{k}}(\cdot)$ is a continuous function, $\{\|f'_{\hat{k}}(x_k)^{-1}\| : k \in K_0\}$ is uniformly bounded. Therefore, there exists $L > 0$ such that $\|d_k\| = \|f'_{\hat{k}}(x_k)^{-1}F(x_k)\| \leq L$ for all $k \geq \hat{k}$, $k \in K_0$. Since $\theta'_{\hat{k}}(\cdot)$ is continuous, we have $\delta > 0$ such that for all x satisfying $\|x - x^*\| \leq \delta$

$$\|\theta'_{\hat{k}}(x) - \theta'_{\hat{k}}(x^*)\| \leq \frac{1 - \sigma - \alpha}{L}\hat{\epsilon}. \tag{9}$$

Since $\lim_{k\to\infty, k\in K_0} x_k = x^*$, we have $\bar{k} > \hat{k}$ such that for all $k > \bar{k}$, $k \in K_0$

$$\|x_k - x^*\| \leq \frac{\delta}{2}. \tag{10}$$

Let $t^* \in (0, 1)$ be such that

$$t^* L < \frac{\delta}{2}. \tag{11}$$

By (10) and (11), for all $k > \bar{k}$, $k \in K_0$, $t \in (0, t^*]$, and $\eta \in (0, 1)$, we have

$$\|x_k + \eta t d_k - x^*\| \leq \delta. \tag{12}$$

Now by (9) and (12), for all $k > \bar{k}, k \in K_0$, and $t \in (0, t^*]$, we have

$$\begin{aligned}
&|\ \theta_{\hat{k}}(x_k + td_k) - \theta_{\hat{k}}(x_k) - td_k^T\theta'_{\hat{k}}(x^*)\ | \\
&\leq t\|d_k\| \int_0^1 \|\theta'_{\hat{k}}(x_k + \eta t d_k) - \theta'_{\hat{k}}(x^*)\| d\eta \\
&\leq t(1 - \sigma - \alpha)\hat{\epsilon}.
\end{aligned} \tag{13}$$

Therefore, for all $k \geq \bar{k}, k \in K_0$, and $t \in (0, t^*]$,

$$\begin{aligned}
&\theta_{\hat{k}}(x_k + td_k) - \theta_{\hat{k}}(x_k) \\
&\quad \leq td_k^T\theta'_{\hat{k}}(x^*) + t(1 - \sigma - \alpha)\hat{\epsilon} \\
&\quad \leq td_k^T\theta'_{\hat{k}}(x_k) + t\|d_k\|\|\theta'_{\hat{k}}(x^*) - \theta'_{\hat{k}}(x_k)\| + t(1 - \sigma - \alpha)\hat{\epsilon} \\
&\quad \leq td_k^T\theta'_{\hat{k}}(x_k) + tL\frac{1 - \sigma - \alpha}{L}\hat{\epsilon} + t(1 - \sigma - \alpha)\hat{\epsilon} \\
&\quad \leq td_k^T f'_{\hat{k}}(x_k)^T f_{\hat{k}}(x_k) + 2t(1 - \sigma - \alpha)\hat{\epsilon} \\
&\quad = -tF(x_k)^T f_{\hat{k}}(x_k) + 2t(1 - \sigma - \alpha)\hat{\epsilon} \\
&\quad = -2t\theta(x_k) + tF(x_k)^T g_{\hat{k}}(x_k) + 2t(1 - \sigma - \alpha)\hat{\epsilon} \\
&\quad \leq -2t\theta(x_k) + t\|F(x_k)\|\|g_{\hat{k}}\| + 2t(1 - \sigma - \alpha)\theta(x_k) \\
&\quad \leq -2t\theta(x_k) + 2t\alpha\theta(x_k) + 2t(1 - \sigma - \alpha)\theta(x_k) \\
&\quad = -2t\sigma\theta(x_k).
\end{aligned}$$

This implies that for all $k \geq \bar{k}$, $k \in K_0$, we have $\rho^{m_k - 1} \geq t^*$, i.e.,

$$\rho^{m_k} \geq \rho t^*. \tag{14}$$

By (8), (14), and the construction of our algorithm, for all $k \geq \bar{k}, k \in K_0$

$$\theta_{\hat{k}}(x_{k+1}) - \theta_{\hat{k}}(x_k) \leq -2\sigma\rho^{m_k}\theta(x_k) \leq -2\rho t^*\sigma\hat{\epsilon} < 0.$$

However, by (7) and the construction of our algorithm, $\theta_{\hat{k}}(x_k)$ is nonincreasing for $k \geq \bar{k}$. This implies $\theta_{\hat{k}}(x_k) \to -\infty$ as $k \to \infty$. This contradicts the fact that $\theta_{\hat{k}}(x_k) \geq 0$ for all k. Hence, K cannot be finite. This proves (5). The final conclusion of this theorem simply follows (5) and the continuity of F. □

Remark 1. We may inductively apply the proof of Lemma 1 and the first part of Theorem 1 to prove (3) and (4). In this way, we may reduce Assumption 2 to "$f'_k(x_k)$ are nonsingular for all k satisfying $x_k \in D_0$."

Remark 2. In [17] trust region methods using the decomposition of F for nonsmooth equations were presented. In the second one, successive approximation was used and $F(x_k)$ and $F'(x_k)$ used in classical trust region methods were replaced by $f(x_k)$ and $f'(x_k)$, respectively. If we use successive approximation and replace all $F(x_k)$ in the SAM by $f_k(x_k)$, then we can also prove the global convergence with the technique of [17].

3. Convergence rate. In order to give a convergence rate, we consider a modification of the SAM.

Modified SAM (MSAM). Given $\rho, \alpha \in (0,1)$, $c \in (0, \frac{1}{1+\alpha})$, an initial vector $x_0 \in R^n$, and a normal decomposition $F = f_0 + g_0$ with $\|g_0\| \leq \frac{\alpha}{2}\|F(x_0)\|$, let $0 < \sigma < 1 - \alpha$. For $k \geq 0$:

1. Solve $F(x_k) + f'_k(x_k)d = 0$ to get d_k.
If

$$\frac{\|F(x_k + d_k)\|}{\|F(x_k)\|} \leq c, \tag{15}$$

we let $x_{k+1} = x_k + d_k$, $f_{k+1} = f_k$, and $g_{k+1} = g_k$. Otherwise, we do Steps 2 and 3 of the SAM.

THEOREM 2. *Theorem* 1 *holds for the* MSAM.

Proof. Let K be the set of k such that (15) holds. If K is finite, then the MSAM is essentially the same as the SAM. Hence Theorem 1 holds in this case. Suppose now K is infinite. Let k_i and k_{i+1} be two consecutive numbers in K. If $k_{i+1} = k_i + 1$, then

$$||F(x_{k_{i+1}})|| \leq c||F(x_{k_i})||.$$

Otherwise, with an argument similar to the first part of the proof of Theorem 1, for any k satisfying $k_i + 1 \leq k \leq k_{i+1}$

$$||F(x_k)|| \leq (1+\alpha)||F(x_{k_i+1})||.$$

Hence, for any k satisfying $k_i + 1 \leq k \leq k_{i+1}$,

$$\begin{aligned} ||F(x_k)|| &\leq c(1+\alpha)||F(x_{k_i})|| \\ &= c_1||F(x_{k_i})|| \leq c_1^i||F(x_{k_1})||, \end{aligned}$$

where $c_1 = c(1+\alpha) \in (0,1)$. This shows that (5) holds. Then the conclusion follows. □

The proof of Theorem 2 shows that the MSAM is globally convergent and the norm $||F(x_{k_i})||$ reduces linearly in i if (15) holds infinitely many times. In the following theorem, we show that under some conditions the linear convergence rate can be realized and that $F = f_k + g_k$ need not be a normal decomposition any more for all large k.

THEOREM 3. *Suppose that the conditions of Theorem* 1 *hold and that x^* is an accumulation point of $\{x_k\}$ generated by the* MSAM. *Suppose that there exist a positive integer $\hat{k}$ and positive numbers r, l_1, l_2, and β such that $x_{\hat{k}} \in S(x^*, r)$ and for all $x, y \in S(x^*, r)$*

$$\|f'_{\hat{k}}(x) - f'_{\hat{k}}(x^*)\| \le l_1\|x - x^*\|,$$

$$\beta \ge \|f'_{\hat{k}}(x^*)^{-1}\|,$$

$$\|g_{\hat{k}}(x) - g_{\hat{k}}(y)\| \le l_2\|x - y\|,$$

and

$$\beta l_1 r < 1, \qquad \frac{\beta}{1 - \beta l_1 r}(2l_1 r + l_2) \le c.$$

Then for all $k \ge \hat{k}$, $f_k = f_{\hat{k}}, g_k = g_{\hat{k}}$, and

$$\|x_{k+1} - x^*\| \le c\|x_k - x^*\|. \tag{16}$$

Furthermore, for all $k \ge \hat{k}$, (15) *also holds.*

Proof. By Theorem 2, $F(x^*) = 0$. For any $x \in S(x^*, r)$, we have

$$\|f'_{\hat{k}}(x) - f'_{\hat{k}}(x^*)\| \le l_1 r.$$

By the Perturbation Lemma, for all $x \in S(x^*, r)$,

$$\|f'_{\hat{k}}(x)^{-1}\| \le \frac{\beta}{1 - l_1 r\beta} =: \beta_1.$$

Let $\bar{x}_{\hat{k}+1} = x_{\hat{k}} + d_{\hat{k}}$. Then from $x_{\hat{k}} \in S(x^*, r)$, we have

$$\begin{aligned}
&\|\bar{x}_{\hat{k}+1} - x^*\| \\
&\quad = \|x_{\hat{k}} - x^* - f'_{\hat{k}}(x_{\hat{k}})^{-1}(F(x_{\hat{k}}) - F(x^*))\| \\
&\quad = \left\| f'_{\hat{k}}(x_{\hat{k}})^{-1}\left(\int_0^1 (f'_{\hat{k}}(x^*) - f'_{\hat{k}}(x_{\hat{k}} + t(x^* - x_{\hat{k}})))(x_{\hat{k}} - x^*)dt \right.\right. \\
&\qquad\qquad \left.\left. + (f'_{\hat{k}}(x_{\hat{k}}) - f'_{\hat{k}}(x^*))(x_{\hat{k}} - x^*) - (g_{\hat{k}}(x_{\hat{k}}) - g_{\hat{k}}(x^*))\right)\right\| \\
&\quad \le \beta_1 \left(\int_0^1 l_1(1-t)\|x_{\hat{k}} - x^*\|^2\,dt + l_1\|x_{\hat{k}} - x^*\|^2 + l_2\|x_{\hat{k}} - x^*\|\right) \qquad (17) \\
&\quad \le \beta_1 \left(l_1\left(\frac{1}{2} + 1\right)\|x_{\hat{k}} - x^*\| + l_2\right)\|x_{\hat{k}} - x^*\| \\
&\quad \le \beta_1 \left(\frac{3}{2}l_1 r + l_2\right)\|x_{\hat{k}} - x^*\| \\
&\quad \le c\|x_{\hat{k}} - x^*\|.
\end{aligned}$$

This implies $\bar{x}_{\hat{k}+1} \in S(x^*, r)$.

Furthermore,

$$\begin{aligned}
\|F(\bar{x}_{\hat{k}+1})\| &= \|F(\bar{x}_{\hat{k}+1}) - f'_{\hat{k}}(x_{\hat{k}})(\bar{x}_{\hat{k}+1} - x_{\hat{k}}) - F(x_{\hat{k}})\| \\
&\le \left\| \int_0^1 (f'_{\hat{k}}(x_{\hat{k}} + t(\bar{x}_{\hat{k}+1} - x_{\hat{k}})) - f'_{\hat{k}}(x^*))(\bar{x}_{\hat{k}+1} - x_{\hat{k}})\, dt \right\| \\
&\quad + \|(f'_{\hat{k}}(x^*) - f'_{\hat{k}}(x_{\hat{k}}))(\bar{x}_{\hat{k}+1} - x_{\hat{k}})\| + \|g_{\hat{k}}(\bar{x}_{\hat{k}+1}) - g_{\hat{k}}(x_{\hat{k}})\| \\
&\le 2l_1 r \|\bar{x}_{\hat{k}+1} - x_{\hat{k}}\| + l_2 \|\bar{x}_{\hat{k}+1} - x_{\hat{k}}\| \\
&= (2l_1 r + l_2)\|f'_{\hat{k}}(x_{\hat{k}})^{-1} F(x_{\hat{k}})\| \\
&\le c\|F(x_{\hat{k}})\|.
\end{aligned}$$

Therefore, we have $x_{\hat{k}+1} = \bar{x}_{\hat{k}+1}$. So that $f_{\hat{k}+1} = f_{\hat{k}}$ and $g_{\hat{k}+1} = g_{\hat{k}}$. Repeating the proof with $\hat{k} := \hat{k} + 1$, we obtain (15) and (16). □

Remark 3. If F is smooth, then $g = 0$ and $l_2 = 0$. Then the quadratic rate of convergence of the Damped Newton method is recovered by (17).

4. Approximation using convolution. In this and the next two sections, we use x_i to represent the ith component of a vector x and x^k to represent a vector.

Without loss of generality, we may assume that $F : R^n \to R^n$ is bounded and uniformly continuous. If F is continuous but not bounded and not uniformly continuous, let the level set D_0 be defined as in §2. By Assumption 1, there is $r_0 > 0$ such that $D_0 \subseteq S(0, r_0)$. Define $F_0 : R^n \to R^n$ by

$$F_0(x) = \begin{cases} F(x) & \text{if } ||x|| \le r_0, \\ F(r_0 \frac{x}{||x||}) & \text{if } ||x|| > r_0. \end{cases}$$

Then (1) is equivalent to

$$F_0(x) = 0,$$

while F_0 is bounded and uniformly continuous. Hence, we assume that F is bounded and uniformly continuous. Let M be a bound of $||F||$.

Let $\omega : R_+ \to R_+$ be the modulus of continuity of F defined by

$$\omega(t) = \sup\{||F(x) - F(y)|| \mid ||x - y|| \le t\}.$$

Then ω is a continuous nondecreasing function [9], $\omega(0) = 0$, and for any x and y in R^n we have

$$||F(x) - F(y)|| \le \omega(||x - y||).$$

We call $\Phi : R^n \to R_+$ a kernel function if

$$\int_{R^n} \Phi(x)dx = 1.$$

If Φ is a kernel function, then $\Phi_\lambda : R^n \to R_+$, defined by

$$\Phi_\lambda(x) = \lambda^n \Phi(\lambda x),$$

where λ is a positive number, is also a kernel function. If Φ is smooth, then Φ_λ is also smooth. If $\phi : R \to R_+$ is a one-dimensional (smooth) kernel function, then $\Phi : R^n \to R_+$, defined by

$$\Phi(x) = \phi(x_1)\phi(x_2)\ldots\phi(x_n), \tag{18}$$

is an n-dimensional (smooth) kernel function. Two famous one-dimensional smooth kernel functions are

$$\phi(t) = \frac{1}{\pi} \cdot \frac{1}{1+t^2} \quad \text{(Cauchy kernel)}$$

and

$$\phi(t) = \frac{1}{\sqrt{\pi}} \cdot e^{-t^2} \quad \text{(Weierstrass kernel)}$$

(see Shapiro [24]).

Suppose now that $\Phi : R^n \to R_+$ is a smooth kernel function. For any $\lambda > 0$, define $F_\lambda : R^n \to R^n$ by

$$F_\lambda(x) = \int_{R^n} F(x-y)\Phi_\lambda(y)dy = \int_{R^n} F(y)\Phi_\lambda(x-y)dy.$$

According to [5], [23], [25], F_λ is a smooth function and

$$\nabla F_\lambda(x) = \int_{R^n} F(y) \nabla \Phi_\lambda(x-y)dy.$$

Furthermore, for any x in R^n,

$$\begin{aligned}||F(x) - F_\lambda(x)|| &= ||\int_{R^n} [F(x) - F(x-y)]\Phi_\lambda(y)dy|| \\ &\leq \int_{R^n} ||F(x) - F(x-y)||\Phi_\lambda(y)dy.\end{aligned}$$

For any $\epsilon > 0$, let $\delta > 0$ and $r > 0$ be such that

$$\int_{||x||>r} \Phi(x)dx \leq \frac{\epsilon}{4M} \tag{19}$$

and

$$\omega(\delta) \leq \frac{\epsilon}{2}. \tag{20}$$

For any $\lambda > \frac{r}{\delta}$, we have

$$\begin{aligned}||F(x) - F_\lambda(x)|| &\leq \int_{R^n} ||F(x) - F(x-y)||\Phi_\lambda(y)dy \\ &\leq \int_{||y||\leq\delta} ||F(x) - F(x-y)||\Phi_\lambda(y)dy + \int_{||y||>\delta} ||F(x) \\ &\qquad - F(x-y)||\Phi_\lambda(y)dy\end{aligned}$$

$$
\begin{aligned}
&\leq \omega(\delta)\int_{||y||\leq\delta}\Phi_\lambda(y)dy+2M\int_{||y||>\delta}\Phi_\lambda(y)dy\\
&\leq \frac{\epsilon}{2}+2M\int_{||y||>\delta}\lambda^n\Phi(\lambda y)dy\\
&= \frac{\epsilon}{2}+2M\int_{||z||>\lambda\delta}\Phi(z)dz \qquad (21)\\
&\leq \frac{\epsilon}{2}+2M\int_{||z||>r}\Phi(z)dz\leq\epsilon.
\end{aligned}
$$

Therefore, in §2, if $F(x_k)\neq 0$, we may choose $\epsilon=\alpha||F(x_k)||$ and construct $f_k\equiv F_\lambda$ for $\lambda>r/\delta$. Then $||g_k||=||F-F_\lambda||\leq\epsilon=\alpha||F(x_k)||$, i.e., we have the normal decomposition required in §2.

To construct F_λ satisfying (21), we need to know $r>0$ and $\delta>0$ such that (19) and (20) hold. If Φ is constructed by (18), then it is not very difficult to choose r. Actually, if (18) holds, then we only need r to satisfy

$$\int_{|t|\geq\frac{r}{\sqrt{n}}}\phi(t)dt\leq(\frac{\epsilon}{4M})^{\frac{1}{n}}.$$

On the other hand, if F is globally Lipschitzian with constant L_0, then $\omega(\delta)\leq L_0\delta$. We may let $\delta=\frac{\epsilon}{2L_0}$. Then (20) will be satisfied. In §5, we will give examples of such applications.

If F is locally Lipschitzian, by the construction of F_0 at the beginning of this section, F_0 is always globally Lipschitzian. Hence our method can be implemented as long as F is locally Lipschitzian.

5. Applications of the successive approximation method.

In this section, we discuss some applications of the successive approximation method. The first two examples have appeared in the literature such as [13].

5.1. The variational inequality problem.

Let C be a closed convex subset of R^n and $\phi: C\to R^n$ be a once continuously differentiable function defined on the open set $D\subseteq R^n$ containing C. This problem, which we denote VI(C,ϕ), is to find a vector $x^*\in C$ such that

$$(x-x^*)^T\phi(x^*)\geq 0,\quad \text{for all } x\in C.$$

The system is equivalent to a system of nonsmooth equations in R^n

$$F(x)\equiv x-\Pi_C(x-\phi(x))=0, \qquad (22)$$

where $\Pi_C(y)$ denotes the projection of y on C. The nonsmoothness of the function F is the consequence of the projection operator $\Pi_C(\cdot)$ (see [13]). When C is a polyhedral set, this operator possesses some B-differentiability properties that can be put to use algorithmically (see [10]). However, it is not easy to establish these properties when C is a general convex set.

Since the projection operator is Lipschitzian with modulus 1, we can use the tool of integration convolution stated in §4 to solve the nonsmooth equations (22) by the successive approximation method.

5.2. The maximal monotone operator problem. Let $T : R^n \to R^n$ be a set-valued maximal monotone operator. An important problem is to find $x \in R^n$ such that

$$0 \in T(x). \tag{23}$$

According to the theory of the maximal monotone operator, the resolvent of T, namely, $P_\mu = (I + \mu T)^{-1}$, where I is the identity operator and μ is a positive number, is always single-valued and nonexpansive (hence globally Lipschitzian) [13]. Moreover, the solution of (23) is equivalent to that of the nonsmooth equation (1) where

$$F(x) = x - P_\mu(x).$$

Since P_μ, and therefore F, is globally Lipschitzian, we can use the tool of integration convolution to approximate F and solve the equation by the successive approximation method.

5.3. LC^1 optimization. Consider

$$\min_x \psi(x), \tag{24}$$

where $\psi : R^n \to R$ is a continuously differentiable function. Let $F = \psi'$. Then we may solve

$$F(x) = 0 \tag{25}$$

to find the stationary points of (24). If F is locally Lipschitzian, then ψ is called an LC^1 function and (24) is called an LC^1 optimization problem. There are many examples of LC^1 optimization problems [16], [18]. For example, if $\psi = \psi_1 - \psi_2$, ψ_1, and ψ_2 are conjugate functions of extended-valued strongly convex functions φ_1 and φ_2, then ψ_1', ψ_2', and F are globally Lipschitzian; see Theorem 2.5 of [16]. Actually, we have

$$\begin{aligned} \psi_i(x) = {} & \max_z\{x^T z - \varphi_i(z)\}, \\ Q_i(x) \equiv {} & \psi_i'(x) = \operatorname{argmax}_z\{x^T z - \varphi_i(z)\} \end{aligned} \tag{26}$$

for $i = 1, 2$ [22] and

$$F(x) \equiv Q_1(x) - Q_2(x). \tag{27}$$

If φ_1 and φ_2 are extended-valued convex quadratic functions, then we can rewrite (26) as

$$Q_i(x) = \operatorname{argmax}_z\{x^T z - \frac{1}{2} z^T H_i z : A_i z \leq b^i\}, \tag{28}$$

where $H_i \in R^{n\times n}$ is symmetric and positive definite, $A_i \in R^{m\times n}$, and $b^i \in R^m$ for $i = 1, 2$. In §6, we will give a numerical example where F is defined by (27) and (28).

5.4. The nonlinear complementarity problem. We consider the nonlinear complementarity problem of finding x such that

$$p(x) \geq 0, \qquad q(x) \geq 0, \qquad p(x)^T q(x) = 0,$$

where $p, q : R^n \to R^n$ are continuously differentiable. This problem can be formulated as a system of nonsmooth equations (1) with

$$F(x) = \min(p(x), q(x)) \tag{29}$$

(see [13]). Now, we give a normal decomposition of F defined by (29). This is simpler than the convolution approximation proposed in the general case.

Let $\alpha \in (0,1)$ be a constant and

$$\epsilon_k = \frac{4\alpha}{\sqrt{n}} \|F(x^k)\| \neq 0.$$

Let

$$(\hat{f}_k(x))_i = \frac{q_i(x) - p_i(x) + 2\epsilon_k}{4\epsilon_k} p_i(x) + \frac{p_i(x) - q_i(x) + 2\epsilon_k}{4\epsilon_k} q_i(x) - \frac{\epsilon_k}{4},$$

$$(\hat{g}_k(x))_i = \frac{1}{4\epsilon_k}(\mid p_i(x) - q_i(x) \mid - \epsilon_k)^2,$$

$$(f_k(x))_i = \begin{cases} F_i(x) & \text{if } \mid p_i(x) - q_i(x) \mid > \epsilon_k, \\ (\hat{f}_k(x))_i & \text{if } \mid p_i(x) - q_i(x) \mid \leq \epsilon_k, \end{cases}$$

$$(g_k(x))_i = \begin{cases} 0 & \text{if } \mid p_i(x) - q_i(x) \mid > \epsilon_k, \\ (\hat{g}_k(x))_i & \text{if } \mid p_i(x) - q_i(x) \mid \leq \epsilon_k, \end{cases}$$

$$\text{for } i = 1, 2, \ldots, n.$$

Then it is easy to verify that

$$F(x) = \min(p(x), q(x)) = f_k(x) + g_k(x), \qquad x \in R^n;$$

f_k is continuously differentiable, g_k is continuous and $\|g_k\| \leq \sqrt{n}\|g_k\|_\infty \leq \sqrt{n}\frac{\epsilon_k}{4} = \alpha\|F(x^k)\|$.

5.5. A piecewise smooth function. Consider

$$F(x) = Ax + \psi(x), \tag{30}$$

where $A : R^n \to R^n$ is a matrix and ψ is a diagonal continuous function, that is, $\psi_i(x) = \phi_i(x_i)$, $i = 1, 2, \ldots, n$. See [9]. Such a system arises from nonsmooth partial differential equations. In a general case, ϕ_i is defined as a piecewise smooth function:

$$\phi_i(x_i) = \begin{cases} u_i(x_i) & \text{if } x_i \geq a_i, \\ v_i(x_i) & \text{if } x_i < a_i, \end{cases}$$

where u_i and v_i are smooth functions with $u_i(a_i) = v_i(a_i)$ and a_i are constants, $i = 1, 2, \ldots, n$.

If $u_i'(a_i) > v_i'(a_i)$, we may replace F_i by $-F_i$. Hence, we may assume that for all i, $u_i'(a_i) \leq v_i'(a_i)$. Let

$$p_i(x) = \begin{cases} (Ax)_i + u_i(x_i) & \text{if } x_i \geq a_i, \\ (Ax)_i + v_i(x_i) + (v_i'(a_i) - u_i'(a_i))(a_i - x_i) & \text{if } x_i < a_i \end{cases}$$

and

$$q_i(x) = \begin{cases} (Ax)_i + u_i(x_i) + (v_i'(a_i) - u_i'(a_i))(x_i - a_i) & \text{if } x_i \geq a_i, \\ (Ax)_i + v_i(x_i) & \text{if } x_i < a_i. \end{cases}$$

Then p, q are smooth functions and

$$F(x) = \min(p(x), q(x))$$

is equivalent to (30).

Now, we can give a normal decomposition of F as in §5.4.

6. Numerical experiments. In this section, we give some computation results to illustrate the SAM and MSAM. The first example is from §5.3.

Example 1. Let

$$Z_1 = \{z \in R^n \ : \ A_1 z \leq b_1\} \quad \text{and} \quad Z_2 = \{z \in R^n \ : \ A_2 z \leq b_2\}$$

be nonempty convex polyhedra in R^n. Let

$$Q_1(x) = \arg\max_{z \in Z_1} \ x^T z - \frac{1}{2} z^T H_1 z$$

and

$$Q_2(x) = \arg\max_{z \in Z_2} \ x^T z - \frac{1}{2} z^T H_2 z,$$

where $H_i \in R^{n \times n}$ are symmetric and positive definite, $A_i \in R^{m \times n}$, and $b_i \in R^m$ for $i = 1, 2$.

We consider the nonsmooth equations

$$F(x) \equiv Q_1(x) - Q_2(x) + Px + c_0 = 0,$$

where P is an $n \times n$ nonsingular matrix and c_0 is a fixed vector in R^n.

PROPOSITION 1. *Let*

$$\psi(x) = \max_z \left\{ x^T z - \frac{1}{2} z^T H z : Az \leq b \right\},$$

where $H \in R^{n \times n}$ is symmetric and positive definite, $A \in R^{m \times n}$, and $b \in R^m$. Assume that $\{z : Az \leq b\}$ is not empty. Then

$$||\psi'(x) - \psi'(y)|| \leq ||H^{-1}|| ||x - y||.$$

Proof. By Theorem 26.3 of [22], ψ is continuously differentiable and $\psi'(x) = Q(x)$ =argmax$_z\{x^T z - \frac{1}{2} z^T H z : Az \leq b\}$. Moreover

$$\varphi(z) = \begin{cases} \frac{1}{2} z^T H z & \text{if } Az \leq b, \\ +\infty & \text{otherwise} \end{cases}$$

is the conjugate function of ψ and φ is a strongly convex function, that is,

$$t\varphi(\bar{z}) + (1-t)\varphi(\hat{z}) - \varphi(t\bar{z} + (1-t)\hat{z}) \geq ct(1-t)||\bar{z} - \hat{z}||^2,$$

TABLE 1

Example 1. *For* $\sigma = 0.7, \rho = 0.625, \alpha = 0.625$; *stopping criterion*: $||F(x_k)|| \leq 10^{-6}$

Test Problem 1	
x^*	(4.7545, 1.0537, 1.3704, 0.5054)
$Q_1(x^*)$	(2.3746, 0.0000, 0.0000, 0.0000)
$Q_2(x^*)$	(1.0000, 1.0000, 1.0000, 1.0000)
c_0	(−36.7090, −13.6718, −12.8669, −6.6836)
x_0	(7.0764, 4.4550, 7.1401, 2.1182)
$\|\|x_{68} - x^*\|\|$	1.0383882×10^{-7}
$\|\|F(x_{68})\|\|$	7.5659260×10^{-7}
Test Problem 2	
x^*	(4.7545, 10.5370, 1.3704, 5.0524)
$Q_1(x^*)$	(0.7643, 2.9602, 0.0000, 1.0998)
$Q_2(x^*)$	(1.0000, 1.5613, 1.0000, 1.0000)
c_0	(−45.9436, −96.0082, −26.2708, −47.1997)
x_0	(12.8565, 11.5656, 8.4323, 14.7801)
$\|\|x_{81} - x^*\|\|$	1.0267013×10^{-7}
$\|\|F(x_{81})\|\|$	7.1972663×10^{-7}
Test Problem 3	
x^*	(−4.7545, −1.0537, −1.3704, −0.5052)
$Q_1(x^*)$	(0.0000, 0.0000, 0.0000, 0.0000)
$Q_2(x^*)$	(1.0000, 1.0000, 1.0000, 1.0000)
c_0	(36.3344, 15.6718, 14.8669, 8.6836)
x_0	(−2.0265, 7.9727, 5.7637, 3.0629)
$\|\|x_{144} - x^*\|\|$	1.3155275×10^{-7}
$\|\|F(x_{144})\|\|$	8.8361463×10^{-7}

where $c = \frac{1}{2||H^{-1}||}$. Let $\hat{z} = \psi'(x)$ and $\bar{z} = \psi'(y)$. By Theorem 23.5 of [22], $x \in \partial\varphi(\hat{z})$ and $y \in \partial\varphi(\bar{z})$. Hence, for any $t \in (0,1)$,

$$\begin{aligned} \varphi(\bar{z}) - \varphi(\hat{z}) &\geq [\varphi(\hat{z} + t(\bar{z} - \hat{z})) - \varphi(\hat{z})]/t + c(1-t)||\bar{z} - \hat{z}||^2 \\ &\geq x^T(\bar{z} - \hat{z}) + c(1-t)||\bar{z} - \hat{z}||^2, \end{aligned}$$

where the second inequality holds because of the basic property of subgradients. Let $t \to 0$; we have

$$\varphi(\bar{z}) - \varphi(\hat{z}) \geq x^T(\bar{z} - \hat{z}) + c||\bar{z} - \hat{z}||^2. \tag{31}$$

Similarly, we have

$$\varphi(\hat{z}) - \varphi(\bar{z}) \geq y^T(\hat{z} - \bar{z}) + c||\hat{z} - \bar{z}||^2. \tag{32}$$

Addition of (31) and (32) shows that

$$(y - x)^T(\bar{z} - \hat{z}) \geq 2c||\bar{z} - \hat{z}||^2.$$

Hence,

$$2c||\bar{z} - \hat{z}||^2 \leq ||\bar{z} - \hat{z}|| \cdot ||y - x||.$$

Therefore,

$$||\psi'(x) - \psi'(y)|| = ||\hat{z} - \bar{z}|| \leq \frac{1}{2c}||x - y||. \qquad \square$$

TABLE 2

The iteration number k and $||x^k - x^* ||_1$ (*or* $|| x^k - x^{**} ||_1$) *for* $\sigma = 0.7, \rho = 0.8, \alpha = 0.285$(SAM), $\alpha = 0.000625$(MSAM); *stopping criterion*: $||x^k - x^{k-1}||_1 \leq 10^{-6}$.

Initial Data	(1,0,0,0)	(1,1,1,1)	(1,0,1,0)	(1,0,0,1)
IF	$25(x^*)$ 11×10^{-7}	$23(x^*)$ 52×10^{-7}	$21(x^*)$ 2×10^{-7}	$18(x^{**})$ 7×10^{-7}
SAM	$130(x^{**})$ 15×10^{-7}	$14(x^*)$ 13×10^{-7}	$29(x^{**})$ 25×10^{-7}	$18(x^{**})$ 7×10^{-7}
MSAM	$4(x^{**})$ 17×10^{-7}	$5(x^*)$ 6×10^{-7}	$8(x^{**})$ 0.1×10^{-7}	$13(x^{**})$ 5×10^{-7}

Proposition 1 can be generalized to the case that φ is locally strongly convex. See [16]. By Proposition 1, F is globally Lipschitzian with modulus $L_0 = ||H_1^{-1}|| + ||H_2^{-1}|| + ||P||$ and $||F(x)|| \leq L_0||x|| + ||F(0)||$. If we take the transformation mentioned in the beginning of §4, then $||F|| \leq L_0 r_0 + ||F(0)|| =: M$. Hence we can give a normal decomposition of F using the convolution discussed in §4. Let

$$f_k(x) = \int_{R^n} F(y)\Phi_{\lambda_k}(x-y)dy,$$

where $\Phi_{\lambda_k}(x) = \lambda_k^n \Phi(\lambda_k x)$

$$\Phi(x) = \phi(x_1)\dots\phi(x_n).$$

We use the Cauchy kernel function stated in §4. At the kth step, let $\epsilon = \alpha||F(x^k)||$ and $\lambda_k \geq \frac{2L_0\sqrt{n}}{\epsilon}\tan(\frac{\pi}{2}(1-(\frac{\epsilon}{4M})^{\frac{1}{n}}))$. Then we have $||F(x) - f_k(x)|| \leq \epsilon = \alpha||F(x^k)||$. We solve the system of linear equations

$$\nabla f_k(x^k)d = -F(x^k),$$

where

$$\nabla f_k(x^k) = \int_{R^n} F(y) \nabla \Phi_{\lambda_k}(x^k - y)dy.$$

Let $F(x) = (F_1(x), \dots, F_n(x))^T, \beta_{k,i}(x) = \lambda_k^{n+1}\phi(\lambda_k x_1)\dots\phi'(\lambda_k x_i)\dots\phi(\lambda_k x_n)$, and $\nabla f_k(x^k) = (J_{ij}(x^k))$. Then

$$J_{ij}(x^k) = \int_{R^n} F_i(y)\beta_{k,j}(x^k - y)dy.$$

Obviously, F is nondifferentiable at x if for one $i = 1$ or 2, $Q_i(x)$ is on the boundary of Z_i. We test the algorithm SAM with a four-dimensional problem where

$$P = \begin{pmatrix} 7.0022 & 0.9018 & 0.6111 & 0.5042 \\ 0.9018 & 7.9745 & 1.0961 & 0.9506 \\ 0.6111 & 1.0961 & 6.9120 & 0.6618 \\ 0.5042 & 0.9506 & 0.6618 & 6.6859 \end{pmatrix},$$

$$H_1 = \begin{pmatrix} 2.0022 & 0.9018 & 0.6111 & 0.5042 \\ 0.9018 & 0.2974 & 1.0961 & 0.9506 \\ 0.6111 & 1.0961 & 1.9120 & 0.6618 \\ 0.5042 & 0.9506 & 0.6618 & 1.6850 \end{pmatrix},$$

$$H_2 = \begin{pmatrix} 2.9174 & 1.4182 & 0.4576 & 1.0221 \\ 1.4182 & 4.4486 & 0.7656 & 1.4075 \\ 0.4576 & 0.7656 & 3.0682 & 0.8975 \\ 1.0221 & 1.4075 & 0.8975 & 3.6760 \end{pmatrix}$$

are randomly generated. Let

$$Z_1 = \{z \in R^4 : z_i \geq 0, i = 1, 2, 3, 4\}$$

and

$$Z_2 = \{z \in R^4 : z_i \geq 1, i = 1, 2, 3, 4\}.$$

We randomly generate x^* and choose c_0 such that x^* is a solution of $F(x) = 0$ and F is nondifferentiable at the solution x^*, i.e., $Q_1(x^*)$ or $Q_2(x^*)$ is on the boundary of Z_1 or Z_2, respectively.

The Monte-Carlo method is used to calculate the integral numerically. Numerical results are shown in Table 1 with random initial points.

Example 2. We consider the following degenerate nonlinear complementarity problem [8]:

$$x \geq 0, \quad p(x) \geq 0, \quad x^T p(x) = 0, \quad x \in R^4$$

where $p : R^4 \to R^4$ is given by

$$p(x) = \begin{pmatrix} 3x_1^2 + 2x_1x_2 + 2x_2^2 + x_3 + 3x_4 - 6 \\ 2x_1^2 + x_1 + x_2^2 + 10x_3 + 2x_4 - 2 \\ 3x_1^2 + x_1x_2 + 2x_2^2 + 2x_3 + 9x_4 - 9 \\ x_1^2 + 3x_2^2 + 2x_3 + 3x_4 - 3 \end{pmatrix}.$$

This problem has two solutions:

$$x^* = (1, 0, 3, 0) \quad \text{and} \quad x^{**} = (\sqrt{6}/2, 0, 0, 0.5) \simeq (1.224745, 0, 0, 0.5).$$

Formulate this problem as $F(x) = 0$ with F defined by (29); then $F(x)$ is differentiable at x^* but nondifferentiable at x^{**}.

Using the Newton line-search method with iteration function method (IF), we quote the particular definition of iteration function $G(\cdot, \cdot)$ given in [6]:

$$G_i(x, d) = \begin{cases} d_i & \text{if } x_i < p_i(x), p_i(x) \geq 0, \\ \nabla p_i(x)^T d & \text{if } x_i > p_i(x), x_i \geq 0, \\ \min(d_i, \nabla p_i^T d) & \text{otherwise.} \end{cases}$$

The computational results by the IF, the SAM, and the MSAM are shown in Table 2. We used single precision. We choose $c = \frac{1}{1+\alpha}$ in the MSAM.

We see that these three methods are globally convergent. The final iteration numbers of the SAM and MSAM are comparable with those of the IF. The SAM and MSAM are further featured by less work at each iteration (the SAM and MSAM only needs to solve a linear system of equations at each step). We can construct approximation functions for any locally Lipschitzian function, but until now generally we do not know how to construct iteration functions in this case. Therefore, the successive approximation method is more general.

Acknowledgments. We are thankful to Professor Olvi Mangasarian, Dr. Rob Womersley, three referees and Mr. Houyuan Jiang for their comments. We are grateful to Professor Carl de Boor for his suggestion on convolution, which upgraded this paper.

REFERENCES

[1] X. Chen, *On the convergence of Broyden-like methods for nonlinear equations with nondifferentiable terms*, Ann. Inst. Statist. Math., 42 (1990), pp. 387–401.

[2] X. Chen and L. Qi, *Parameterized Newton method and Broyden-like method for solving nonsmooth equations*, Comput. Optim. Appl. 3 (1994), pp. 157–179.

[3] X. Chen and T. Yamamoto, *On the convergence of some quasi-Newton methods for nonlinear equations with nondifferentiable operators*, Computing, 48 (1992), pp. 87–94.

[4] J. E. Dennis, Jr., and R. B. Schnabel, *Numerical Methods for Unconstrained Optimization and Nonlinear Equations*, Prentice-Hall, Inc., Englewood Cliffs, 1983.

[5] Y. M. Ermoliev, V. I. Norkin, and R. J-B. Wets, *The minimization of discontinuous functions: mollifier subgradients*, SIAM J. Control Optim., 33 (1995), pp. 149–167.

[6] S. P. Han, J. S. Pang, and N. Rangaraj, *Globally convergent Newton methods for nonsmooth equations*, Math. Oper. Res., 17 (1992), pp. 586–607.

[7] S. A. Gabriel and J. S. Pang, *A trust region method for constrained nonsmooth equations*, in Large Scale Optimization: State of the Art, W. W. Hager, D. W. Hearn, and P. M. Pardalos, eds., Kluwer Academic Publishers, Boston, 1994, pp. 159–186.

[8] M. Kojima and S. Shindo, *Extensions of Newton and quasi-Newton methods to systems of PC^1 equations*, J. Oper. Res. Soc. Japan, 29 (1986), pp. 352–374.

[9] J. M. Ortega and W. C. Rheinboldt, *Iterative Solution of Nonlinear Equations in Several Variables*, Academic Press, New York, 1970.

[10] J. S. Pang, *Newton's method for B-differentiable equations*, Math. Oper. Res., 15 (1990), pp. 311–341.

[11] J. S. Pang, *A B-differentiable equation based, globally, and locally quadratically convergent algorithm for nonlinear programs, complementarity and variational inequality problems*, Math. Programming, 51 (1991), pp. 101–131.

[12] J. S. Pang and S. A. Gabriel, *NE/SQP: A robust algorithm for the nonlinear complementarity problem*, Math. Programming, 60 (1993), pp. 295–337.

[13] J. S. Pang and L. Qi, *Nonsmooth equations: motivation and algorithms*, SIAM J. Optim., 3 (1993), pp. 443–465.

[14] R. Poliquin and L. Qi, *Iteration function in some nonsmooth optimization algorithms*, Math. Oper. Res., to appear.

[15] L. Qi, *Convergence analysis of some algorithms for solving nonsmooth equations*, Math. Oper. Res., 18 (1993), pp. 227–244.

[16] L. Qi, *LC^1 functions and LC^1 optimization problems*, Appl. Math. Preprint, AM 91/21, The University of New South Wales, Sydney, Australia, 1991.

[17] L. Qi, *Trust region algorithms for solving nonsmooth equations*, SIAM J. Optim., 5 (1995), pp. 218–229.

[18] L. Qi, *Superlinear convergent approximate Newton methods for LC^1 optimization problems*, Math. Programming, 64 (1994), pp. 277–294.

[19] L. Qi and J. Sun, *A nonsmooth version of Newton's method*, Math. Programming, 58 (1993), pp. 353–367.

[20] D. Ralph, *Global convergence of damped Newton's method for nonsmooth equations via the path search*, Math. Oper. Res., 19 (1994), pp. 352–389.

[21] S. M. Robinson, *Newton's method for a class of nonsmooth functions*, Industrial Engineering Working Paper, University of Wisconsin, Madison, Wisconsin, 1988.

[22] R. T. Rockafellar, *Convex Analysis*, Princeton University Press, Princeton, NJ, 1970.

[23] L. Schwartz, *Théorie des Distributions*, Hermann, Paris, 1966.

[24] H. S. Shapiro, *Smoothing and Approximation of Functions*, Van Nostrand Reinhold Company, New York, 1969.

[25] S. L. Sobolev, *Some Applications of Functional Analysis in Mathematical Physics*, Nauka, Moscow, 1988 (3rd edition, in Russian).

SIAM J. CONTROL AND OPTIMIZATION
Vol. 33, No. 2, pp. 419–439, March 1995

© 1995 Society for Industrial and Applied Mathematics
005

OPTIMAL SUPERVISORY CONTROL OF DISCRETE EVENT DYNAMICAL SYSTEMS*

RATNESH KUMAR† AND VIJAY K. GARG‡

Abstract. The notion of optimal supervisory control of discrete event dynamical systems (DEDSs) is formalized in the framework of Ramadge and Wonham. A DEDS is modeled as a state machine and is controlled by disabling some of its transitions. Two types of cost functions are defined: a cost of control function corresponding to disabling transitions in the state machine, and a penalty of control function corresponding to reaching some undesired states or not reaching some desired states in the controlled system. The control objective is to design an *optimal* control mechanism, if it exists, so that the *net* cost is minimized. Since a DEDS is represented as a state machine—a directed graph—network flow techniques are naturally applied for designing optimal supervisors. It is also shown that our techniques can be used to solve supervisory control problems under complete as well as partial observation. In particular, for the first time, techniques for computing the supremal controllable and normal sublanguage and the infimal controllable and normal/observable superlanguage without having to perform alternate computations of controllable and normal/observable languages are obtained.

Key words. discrete event dynamical systems, supervisory control, automata theory, optimal control, max-flow min-cut

AMS subject classifications. 68Q75, 93B25, 93C83

1. Introduction.

Research of the supervisory control of a discrete event dynamical system (DEDS) was pioneered by Ramadge and Wonham [23]. In [23], a DEDS, also called plant, is modeled as a state machine (SM) and the behavior of a DEDS is described by the language accepted by the corresponding SM. A controller or a supervisor, based on its observation of the past behavior of the plant, determines the transitions to be disabled in the plant, so that some desired qualitative control objective is achieved. Usually, the control objective is to restrict the plant behavior such that it remains confined within a specific range [24]. In some other cases, the control objective is to design a supervisor so that the closed-loop behavior *eventually* remains confined to a prescribed range [3], [18], [12]. Recently, there has also been some work in which the control objective is to restrict the plant behavior so that a certain cost function defined along the trajectory of the system is optimized [20], [19], [2], [26], [10].

In [20], [19], a cost function is defined on the set of transitions and the control objective is to restrict the plant behavior in such a way that after starting from a *given* initial state the plant reaches one of the accepting states along a trajectory of optimal cost. Authors provide an efficient heuristic search-based algorithm to solve the problem. In [2] also, a cost function is defined on the set of transitions and the control objective is to restrict the plant behavior so that after starting from *any* state the plant reaches one of the accepting states along a trajectory of optimal cost. In

* Received by the editors August 3, 1992; accepted for publication (in revised form) October 22, 1993. This research was supported in part by the Center for Robotics and Manufacturing, University of Kentucky, National Science Foundation grants NSF-ECS-9409712 and NSF-CCR-9110605, and a TRW faculty assistantship award. A preliminary version of this paper appeared in R. Kumar and V. K. Garg, *Optimal control of discrete event dynamical systems using network flow techniques*, in Proceedings of 1991 Annual Allerton Conference, Urbana, Illinois, October 1991, pp. 705–714.

† Department of Electrical Engineering, University of Kentucky, Lexington, Kentucky 40506-0046 (kumar@engr.uky.edu).

‡ Department of Electrical and Computer Engineering, University of Texas at Austin, Austin, Texas 78712-1084 (vijay@pine.ece.utexas.edu).

[26], two types of costs, a control cost and a path cost, are defined on the graph representing a plant. The control objective is to determine that subgraph of the plant graph for which the maximum of the total cost along all its trajectories is minimal. The notions of cost of control and penalty of control considered in our paper are somewhat similar to that of control cost and path cost in [26]. However, our control objective is to determine a *state-feedback* supervisor so that the net cost of disabling transitions, that of reaching undesired states, and that of not reaching desired states is minimized. Thus our control objective is different from that considered in [26].

In this paper, we consider two types of cost functions: (i) A positive *cost of control* function is defined on the set of transitions corresponding to the cost of disabling a transition; if a certain transition such as arrival of a customer in a queue is disabled by a supervisor at a certain point, then its cost of control is added to the net cost, otherwise—if the transition is not disabled—no cost is added to the net cost. (ii) A *penalty of control* function is defined on the set of states corresponding to their reachability in the controlled system. The penalty of control takes a negative or positive value depending on whether the state is desired or undesired. If a state is desired—e.g., working or idle state of a machine—then a negative penalty is associated with it. If such a state remains unreachable in the controlled plant, then a positive cost equal in magnitude to its penalty of control is added to the net cost, otherwise—if the state is reachable—no cost is added to the net cost. On the other hand, if a state is undesired—e.g., over/underflow of a buffer—then a positive penalty is associated with it. If such a state can be reached in the controlled plant, then a cost equal to its penalty of control is added to the net cost, otherwise—if it remains unreachable—no cost is added to the net cost. The optimal control problem is to determine a state-feedback supervisor for which the net cost is minimized. State-feedback supervisors [8], [21], [13], [6] exercise control based on the state of the plant rather than the sequence of events executed by it. However, this does not result in any loss of generality, since a "string-feedback" supervision is equivalent to a state-feedback supervision on a suitably refined [7] model of a plant.

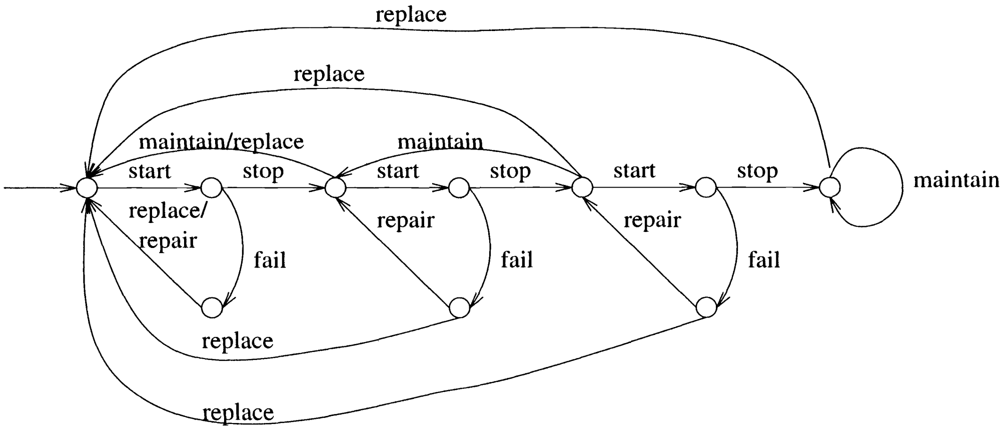


FIG. 1. *Diagram for machine* P *of Example* 1 *with* $N = 4$.

Example 1. Consider for example a machine P shown in Fig. 1. (For clarity, the state labels have been omitted.) Initially P is in the "first idle" state (i.e., idle and unused state). When event "start" is executed, it goes to the "first working" state. In the "kth working" state, where $1 \leq k < N$, the machine may either "fail," in which case it goes to the "kth broken" state; or it may complete its operation, execute "stop," and go to the "$(k+1)$th idle" state. In the "kth broken" state, the machine is either "repaired," in which case it goes back to the "kth idle" state; or it is "replaced," in which case it goes to the "first idle" state. In the "kth idle" state, either the event "start" is executed, which sends the machine to the "kth working" state; or "replace" is executed, which sends the machine to the "first idle" state; or "maintain" is executed, which sends the machine to the "$(\max\{1, 2k - N\})$th idle" state. Thus, in the "Nth idle" state, execution of the "maintain" event does not result in a change in state. Note that the function $\max\{1, 2k - N\}$ is chosen only for an illustration; in general, it is an increasing function of k taking values smaller than k except at $k = 1$ and $k = N$, where it equals k. An optimal control policy is needed to decide (i) whether to repair or replace the machine in a broken state and (ii) whether to operate or maintain or replace the machine in an idle state. An optimal control policy evidently depends on the cost of replacement of machine, cost of repair in the kth broken state, cost of maintenance in the kth idle state, payoff of operating the machine in the kth idle state, penalty of being in a broken state, payoff of being in an idle or a working state, etc.

Our setting is similar to that considered in [27], in which appropriate cost and penalty functions were defined on a suitably refined model of a given plant and the dynamic programming algorithm was used to determine the *existence* of a supervisor for a given control problem. The computational complexity of the dynamic programming algorithm was used to determine the computational complexity of the supervisory control problem thus solved. However, no technique for the *synthesis* of a supervisor, if it exists, was given in that reference. Our approach to optimal supervisory control differs from that considered in [27] in two ways. First, we show that an optimal supervisory control problem of the type described above can be solved using network flow algorithms. Thus a more general algorithm such as dynamic programming can also be used, although this will result in an *increase* in computational complexity. Second, we show that our techniques are equally applicable for the synthesis of supervisors, whenever they exist.

The motivation of formulating an optimal supervisory control problem is twofold: first, to introduce a formal framework for optimal supervisory control in which the control objective is to optimize a suitably defined cost and penalty of exercising controls; and second, to present a unified technique for supervisory synthesis under complete as well as partial observation. In particular, we obtain techniques for computing the supremal controllable and normal sublanguage [22], [16], [1], [11], [4], the infimal controllable and normal superlanguage [14], and the infimal controllable and observable superlanguage [16], [5], [25], [9] *without* having to perform alternate computations of controllable and normal/observable languages as is done in [4] (also refer to Remark 3). Our techniques also illustrate that a supervisory control problem under complete or partial observation can be solved using a state-feedback type of control on a suitably refined state machine representation of the plant. We provide techniques to obtain the appropriate refinements.

In §2, we introduce our notation and formally describe the optimal supervisory control problem under complete as well as partial state observation. In §3, we show how the network flow algorithms can be used to solve the optimal supervisory control problems. In §4, we show that by appropriately defining the cost and penalty functions, our techniques can be used to solve the supervisory control problem under complete as well as partial observation.

2. Notation and problem formulation. A discrete event dynamical system to be controlled, called plant, is modeled as a state machine [7] and denoted as a 4-tuple $G := (X, \Sigma, \delta, x_0)$, where X denotes the set of states, Σ denotes the finite set of *events*, $\delta : X \times \Sigma \rightarrow X$ denotes the partial deterministic state transition function, and $x_0 \in X$ denotes the initial state. A triple $(x_1, \sigma, x_2) \in X \times \Sigma \times X$, such that $\delta(x_1, \sigma) = x_2$, is called a *transition* in G. The behavior of G is described by the language $L(G)$:

$$L(G) := \{s \in \Sigma^\star \mid \delta(x_0, s) \text{ is defined}\},$$

where $\Sigma^\star$ denotes the set of finite sequences of events belonging to Σ, including the zero-length sequence ϵ; the transition function is extended in a natural way to $\delta : X \times \Sigma^\star \rightarrow X$.

In general, a supervisor or a controller determines the set of events to be disabled after each transition, based on the record of observed states and events. We consider a supervisor, denoted S, to be a map $S : X \rightarrow 2^\Sigma$ that determines the set of events $S(x) \subseteq \Sigma$ to be disabled at each state $x \in X$. Events not belonging to the set $S(x)$ remain enabled at x. A supervisor as defined above is called a state-feedback [21], [13], [12], as it exercises control based on the state of G (and not based on the record of observed states and events). As is shown below, this does not result in any loss of generality, as a more general supervisor, which exercises control based on the observed sequence of events, can equivalently be viewed as a state-feedback supervisor on a suitably refined model of the plant. Readers are referred to [24] for a more general definition of a supervisor. The controlled plant, denoted G_S, is another state machine given as the 4-tuple $G_S := (X, \Sigma, \delta_S, x_0)$, where X, Σ, x_0 are as defined above and δ_S denotes the state transition function of the controlled plant G_S:

$$\forall x \in X, \sigma \in \Sigma : \delta_S(x, \sigma) := \begin{cases} \delta(x, \sigma) & \text{if } \sigma \notin S(x), \\ \text{undefined} & \text{otherwise.} \end{cases}$$

The behavior of the closed-loop system is described by the language $L(G_S)$ generated by the controlled plant. It is clear that $L(G_S) \subseteq L(G)$.

Next we formally describe the problem of optimal supervisory control. Let $c : X \times \Sigma \rightarrow \mathcal{R}^+$ denote a cost of control function, where $\mathcal{R}^+$ denotes the set of strictly positive reals, including infinity. The cost $c(x, \sigma)$ represents the cost of disabling the event $\sigma \in \Sigma$ at the state $x \in X$.

We assume for simplicity that the cost of control is a "one-time" cost. A justification for this simplifying assumption is the following: In G_S, a certain state may be visited either only once (if it is not a node on any cycle in the graph of G_S) or an unbounded number of times (otherwise). If a state is visited only once, then the corresponding transitions are controlled only once. Hence, in this case, the cost of control ought to be one-time. On the other hand, if a state is visited an unbounded number of times, then the corresponding transitions are controlled each time the state is visited. However, due to the state-feedback nature of supervision, the *same* control is exercised on each occasion. Hence, in this case, the cost of controlling such transitions can be associated with the first time the control is exercised. The one-time cost of control assumption can also be interpreted as follows: A transition once disabled/enabled at the corresponding state remains disabled/enabled in that state so that on subsequent visits to that state no cost is incurred in disabling/enabling that transition. Alternatively, the cost of control is primarily of setting up a control mechanism—a switch, for example—and the cost of engaging or disengaging the switch is small compared to its one-time setup cost. In each case, the assumption of state-feedback supervision is crucial.

Next, a penalty of control function $p : X \to \mathcal{R}$ is defined on the state set, where $\mathcal{R}$ denotes the set of reals, including positive and negative infinity. It corresponds to the penalty associated with reachability of a certain state in the controlled plant. The penalty function may take a positive or negative value depending on whether the corresponding state is undesired or desired. Given a state $x \in X$, if $p(x) < 0$, then x is a desired state and it should remain reachable in the controlled plant. In case x is unreachable in the controlled plant, a positive cost equal to $-p(x)$ is added to the net cost as a penalty, else no cost is added to the net cost. Similarly, if $p(x) > 0$ for some $x \in X$, then x is an undesired state and it should remain unreachable in the controlled plant. In case x is reachable in the controlled plant, a cost equal to $p(x)$ is added to the net cost as a penalty, else no cost is added to the net cost.

We assume for simplicity, as above for the cost of control, that the penalty of control is a one-time penalty; i.e., the penalty of reaching an undesired state does not depend on the number of times that state is visited. The justification for this simplified assumption is similar to that given above for the one-time cost of control assumption.

Remark 1. If $c(\cdot, \sigma) = \infty$ for an event $\sigma \in \Sigma$, then σ should not be disabled by any supervisor at any state of G. Thus an infinite cost of control of an event captures the notion of *uncontrollable* events [23], [24]. The notion of desired or *target* behavior can be captured by defining the penalty function to be infinity for those states that are reachable by the strings in the undesired behavior, and any fixed negative real for those states that are reachable by strings in the desired behavior. However, this requires that the state machine G be refined [7], [11] with respect to the given target behavior, so that the states corresponding to the desired and undesired behavior can be uniquely identified, and the penalty function is unambiguously defined. This is explained formally in §4.

With the above definitions of the cost and penalty of control functions we can define the optimal supervisory control problem.

DEFINITION 1. For any supervisor $S : X \to 2^{\Sigma}$, the *net cost* of using S, denoted $C(S)$, is defined to be

$$C(S) := \sum_{x \in \mathrm{Re}(G_S)} \left[\sum_{\sigma \in S(x)} c(x, \sigma) \right] + \sum_{\substack{x \in \mathrm{Re}(G_S), \\ p(x) > 0}} p(x) + \sum_{\substack{x \notin \mathrm{Re}(G_S), \\ p(x) < 0}} -p(x),$$

where $\mathrm{Re}(G_S)$ is the set of reachable states[1] in G_S.

Thus the net cost of control of using S consists of the sum of three terms: (i) The first term corresponds to the cost of disabling the events by S. $\sum_{\sigma \in S(x)} c(x, \sigma)$ is the total cost of disabling events at state $x \in X$. Thus $\sum_{x \in \mathrm{Re}(G_S)} \sum_{\sigma \in S(x)} c(x, \sigma)$ is the total cost of disabling the events by S. (ii) The second term, $\sum_{x \in \mathrm{Re}(G_S), p(x) > 0} p(x)$, denotes the penalty of reaching undesired states. (iii) Finally, the third term $\sum_{x \notin \mathrm{Re}(G_S), p(x) < 0} -p(x)$, denotes the penalty of not reaching the desired states.

[1] Given a state machine $V := (Q, \Sigma, \rho, q_0)$, the set of reachable states $\mathrm{Re}(V)$ is recursively defined as: (i) $q_0 \in \mathrm{Re}(V)$; and (ii) $q \in \mathrm{Re}(V), \exists \sigma \in \Sigma : \rho(q, \sigma)$ is defined $\Rightarrow \rho(q, \sigma) \in \mathrm{Re}(V)$.

Optimal supervisory control problem 1 (OSCP1). Let a plant $G := (X, \Sigma, \delta, x_0)$, a cost of control function $c : X \times \Sigma \to \mathcal{R}^+$, and a penalty of control function $p : X \to \mathcal{R}$ be given. Design a supervisor $S : X \to 2^\Sigma$ such that the net cost is minimized, i.e., determine

$$\arg\left\{\min_S C(S)\right\}.$$

2.1. Partial state observation. In OSCP1, a supervisor while deciding its control actions assumes that a complete state information of G is available. We pose another optimal supervisory control problem in which a complete state information is not available, and there exists a mask $\Psi : X \to Y$ defined from the state space X to an observation space Y such that for each $x \in X$, $\Psi(x) \in Y$ is the state value observed by a supervisor. A supervisor in this case is given by a map $S' : Y \to 2^\Sigma$. Since a supervisor takes a control action based on observing a state $y \in Y$, given any $y \in Y$, the same control action is taken at all states in the set $\Psi^{-1}(y) := \{x \in X \mid \Psi(x) = y\}$. Thus corresponding to a supervisor $S' : Y \to 2^\Sigma$, we can equivalently define a supervisor $S : X \to 2^\Sigma$ with the constraint C1:

$$\textbf{C1: } \forall x_1, x_2 \in X : \Psi(x_1) = \Psi(x_2) \Rightarrow S(x_1) = S(x_2).$$

Optimal supervisory control problem 2 (OSCP2). Let a plant $G := (X, \Sigma, \delta, x_0)$, a cost of control function $c : X \times \Sigma \to \mathcal{R}^+$, a penalty of control function $p : X \to \mathcal{R}$, and a mask $\Psi : X \to Y$ be given. Design a supervisor $S' : Y \to 2^\Sigma$ (equivalently a supervisor $S : X \to 2^\Sigma$ satisfying C1) such that the net cost is minimized; i.e., determine

$$\arg\left\{\min_{S:\text{ C1 holds}} C(S)\right\}.$$

Remark 2. The difference between OSCP1 and OSCP2 is that, in OSCP2, the minimization is performed over all supervisors that also satisfy the constraint C1, whereas no such constraint exists in OSCP1. It is easily seen that given an instance of OSCP1, it can be reduced to an instance of OSCP2 by setting the mask function to be the identity function. We show in the next section that given an instance of OSCP2, it can be reduced to an instance of OSCP1 by suitably modifying the cost of control function and the graph representing the plant. Thus the two formulations are reducible to each other.

In the formulation of OSCP2, a state-based mask function is used. This is in contrast to the setting of supervisory control, where usually an event-based mask is used. However, an event-based mask can be used to obtain a state-based mask by first constructing a state estimator, as in [17], and next identifying all the states with identical state estimates to have equal mask value. Thus it is possible to reduce an optimal control problem under partial observation of events to one under partial observation of states.

3. Solution using network flow algorithm. In this section we provide a solution to the optimal supervisory control problems introduced in §2. The problem is to determine for each transition in the state machine G whether to disable or enable it, so that the net cost is minimized. We show that this problem is equivalent to determining an optimal partition of the state space X into the set of states that remain reachable in the controlled plant and the set of remaining unreachable states. The desired optimal partition is determined using the max-flow min-cut theorem [15], a technique for optimal partitioning of directed graphs.

3.1. Max-flow min-cut theorem. Interested readers are referred to [15] for a formal and elaborate description of the max-flow min-cut theorem. Informally described, in its simplest form, a *flow network* is represented as a weighted directed graph having a single *source* node from where the flow starts and a single *terminal* node where the flow terminates. The weights on the directed edges of the graph represent the maximum flow capacities of the corresponding edges (the minimum capacity is zero unless specified). Formally, we have the following definition.

DEFINITION 2. A *flow network* N is a weighted directed graph described by a triple $N = (V, E, u)$, where V denotes the set of vertices or nodes of N, $E \subseteq V^2$—subset of *ordered* pairs of V^2—denotes the set of directed edges or links of N, and $u : E \rightarrow \mathcal{R}^+$ denotes the maximum capacity function of links. V contains two special nodes s and t, the source node and the terminal node respectively.

The basic flow optimization problem is to determine for a given flow network a flow of maximum value between its source and terminal nodes subject to the edge capacity constraints, where a flow and its value are defined as follows.

DEFINITION 3. A *flow* for a network N is a map $f : E \rightarrow \mathcal{R}^+$ such that

1. $\forall e \in E : f(e) \leq u(e)$,
2. $\sum_{v\in V:(s,v)\in E} f((s,v)) = \sum_{v'\in V:(v',t)\in E} f((v',t))$, and
3. $\forall v \in V, v \neq s, v \neq t : \sum_{v'\in V:(v,v')\in E} f((v,v')) = \sum_{v''\in V:(v'',v)\in E} f((v'',v))$.

$\sum_{v\in V:(s,v)\in E} f((s,v)) = \sum_{v'\in V:(v',t)\in E} f((v',t))$ is called the *value* of f. A *max-flow* is a flow of maximum (flow) value. Thus a flow is an assignment of a positive number to each edge in the flow network that corresponds to the amount of flow on that edge, satisfying three constraints. First, the amount of flow through each edge is no greater than its capacity; second, the net flow out of the source node equals the net flow into the terminal node; and finally, the net flow out of the intermediate nodes is zero. The value of a flow equals the net flow out of the source node (equivalently, the net flow into the terminal node).

DEFINITION 4. A *cut* of a network N is a partition of V such that s and t are in different partitions. Let $V_s \subseteq V$ and $V_t := V - V_s$ denote the partitions of a cut such that $s \in V_s$ and $t \in V_t$. Then the *capacity* of this cut is defined as

$$\sum_{(i,j)\in(V_s\times V_t)\cap E} u((i,j)).$$

A *min-cut* is a cut of minimum capacity. Thus any partition of the nodes in N such that the source node and the terminal node belong to different partitions is called a cut. The capacity of a cut equals the sum of capacity of those edges that emerge out of a node contained in the partition containing the source node and terminate at a node contained in the partition containing the terminal node.

THEOREM 1. (max-flow min-cut) [15]. The value of a max-flow of a flow network equals the capacity of a min-cut of that network.

Algorithms for computing a max-flow can be found in [15]. In this paper, we are interested in computing a cost-minimizing supervisor. This is shown to be equivalent to determining a min-cut for a suitably defined flow network, which in view of Theorem 1 can be computed using any of the max-flow computations.

3.2. Solution of OSCP1. In this subsection we provide a solution for the OSCP1 using the network flow technique discussed in the previous subsection. It is clear that each supervisor $S : X \rightarrow 2^{\Sigma}$ partitions the state space X into $\mathrm{Re}(G_S) \cup (X - \mathrm{Re}(G_S))$, the sets of reachable and unreachable states in the controlled plant G_S. We define a supervisor to be parsimonious if and only if it disables those transitions that are defined from a state in $\mathrm{Re}(G_S)$ to a state in $X-$ $\mathrm{Re}(G_S)$. Formally, consider the following definition.

DEFINITION 5. A supervisor $S : X \rightarrow 2^{\Sigma}$ is said to be *parsimonious* if and only if for each state $x \in X$ and event $\sigma \in \Sigma$:

$$\sigma \in S(x) \Leftrightarrow [x \in \mathrm{Re}(G_S),\ \delta(x, \sigma) \notin \mathrm{Re}(G_S)].$$

We prove that parsimonicity is a necessary condition for optimality of a supervisor.

LEMMA 1. If S is an optimal supervisor, then S is parsimonious.

Proof. Assume for the sake of contradiction that S is optimal but not parsimonious. Then there exist an event $\sigma \in \Sigma$ and states $x_1, x_2 \in \mathrm{Re}(G_S)$ such that $\delta(x_1, \sigma) = x_2$ and $\sigma \in S(x_1)$, i.e., σ is disabled at x_1. Consider a supervisor S' that exercises the same control action as S does, except that at state x_1 it does not disable the event σ, i.e., $\sigma \notin S'(x_1)$. Then $C(S') = C(S) - c(x_1, \sigma) < C(S)$, for $c(x_1, \sigma) > 0$. Thus we obtain a contradiction to the optimality of S. □

The following is an immediate corollary of Lemma 1.

COROLLARY 1. OSCP1 is equivalent to determining

$$\arg \left\{ \min_{S:S \text{ parsimonious}} C(S) \right\}.$$

Proof. The proof follows from the fact that parsimonicity is a necessary condition for optimality, and the definition of OSCP1. □

In the next theorem we provide a technique for solving the OSCP1 using the max-flow min-cut theorem. First we define a flow network N_G corresponding to the state machine G, the cost of control function c, and the penalty of control function p as follows.

DEFINITION 6. Given a SM $G := (X, \Sigma, \delta, x_0)$ with cost of control function $c : X \times \Sigma \rightarrow \mathcal{R}^+$ and penalty of control function $p : X \rightarrow \mathcal{R}$, a flow network, denoted N_G, is defined to be $N_G := (V_G, E_G, u_G)$, where

1. $V_G := X \cup \{s, t\}$ with $s, t \notin X$,
2. $E_G := \{(x_1, x_2) \in X \times X \mid \exists \sigma \in \Sigma \text{ s.t. } \delta(x_1, \sigma) = x_2\}$
   $\cup \{(x, t) \in X \times \{t\} \mid p(x) > 0\}$
   $\cup \{(s, x) \in \{s\} \times X \mid p(x) < 0\}$,
3. $\forall (x_1, x_2) \in E_G \cap (X \times X) : u_G((x_1, x_2)) := \sum_{\sigma \in \Sigma : \delta(x_1, \sigma) = x_2} c(x_1, \sigma)$
   $\forall x \in X \text{ s.t. } p(x) > 0 : u_G((x, t)) := p(x)$
   $\forall x \in X \text{ s.t. } p(x) < 0 : u_G((s, x)) := -p(x)$.

Thus the node set of N_G is obtained by adding to the state set X of G two extra nodes s and t—the source and the terminal node, respectively. The edge set of N_G consists of: (i) An edge from $x_1 \in X$ to $x_2 \in X$ if there exists a transition from x_1 to x_2 in G. The capacity of such an edge equals the sum of cost of disabling each transition from x_1 to x_2. (ii) An edge from each $x \in X$ for which $p(x) > 0$ to the terminal node t, with capacity $p(x)$. (iii) An edge from the source node s to each $x \in X$ for which $p(x) < 0$, with capacity $-p(x)$.

Example 2. Consider the plant G shown in Fig. 2, with the state set $X = \{1, 2\}$, the event set $\Sigma = \{a, b\}$, the initial state $x_0 = 1$, and the transition function $\delta(1, a) =$

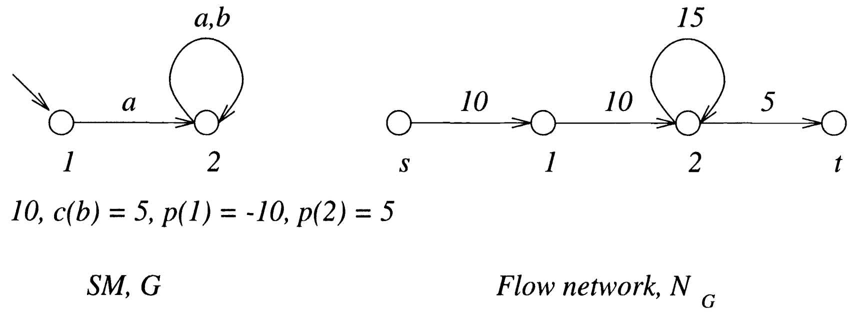


FIG. 2. *Diagram illustrating construction of N_G.*

$2, \delta(2,a) = \delta(2,b) = 2$. Then $L(G) = \overline{a(a+b)^\star}$. Let the cost of control function be defined as $c(a) = 10$ and $c(b) = 5$ and the penalty of control function be defined as $p(1) = -10$ and $p(2) = 5$. Then the flow network N_G, corresponding to the plant G, obtained using Definition 6 is shown in Fig. 2.

DEFINITION 7. Given a supervisor $S : X \to 2^\Sigma$, the cut of flow network N_G induced by S is defined to be $[\mathrm{Re}(G_S) \cup \{s\}] \cup [(X - \mathrm{Re}(G_S)) \cup \{t\}]$. Given a cut $V_s \cup (V_G - V_s)$ of N_G, where $V_s \subseteq V_G$, $s \in V_s$, and $t \notin V_s$, the parsimonious supervisor $S : X \to 2^\Sigma$ induced by the cut is defined to be: for each $x \in X$ and $\sigma \in \Sigma$, $\sigma \in S(x)$ if and only if $x \in V_s$ and $\delta(x,\sigma) \notin V_s$.

THEOREM 2. (Solution of OSCP1). The supervisor induced by a min-cut of N_G is a solution of OSCP1.

We prove a lemma before proving Theorem 2.

LEMMA 2. If S is a parsimonious supervisor, then $C(S)$ equals the capacity of the cut of N_G induced by S.

Proof. Consider the cut $[\mathrm{Re}(G_S) \cup \{s\}] \cup [(X - \mathrm{Re}(G_S)) \cup \{t\})$ induced by S. The capacity of this cut is given by the sum of capacities of all the edges that originate from a node in the set $\mathrm{Re}(G_S) \cup \{s\}$ and terminate at a node in the set $(X - \mathrm{Re}(G_S)) \cup \{t\}$. Let the set of such edges be denoted as E_S, i.e.,

$$E_S := \{(x_1, x_2) \in E_G \mid x_1 \in \mathrm{Re}(G_S) \cup \{s\} \text{ and } x_2 \in (X - \mathrm{Re}(G_S)) \cup \{t\}\}.$$

Then the capacity of the cut of N_G induced by S equals $\sum_{e \in E_S} u_G(e)$. We note that

$$\begin{aligned} E_S = \{(x_1, x_2) &\in \mathrm{Re}(G_S) \times (X - \mathrm{Re}(G_S)) \mid \exists \sigma \in \Sigma \text{ s.t. } \delta(x_1,\sigma) = x_2\} \\ &\cup \{(x,t) \mid x \in \mathrm{Re}(G_S), p(x) > 0\} \\ &\cup \{(s,x) \mid x \notin \mathrm{Re}(G_S), p(x) < 0\}. \end{aligned}$$

Hence

$$\begin{aligned} \sum_{e \in E_S} u_G(e) = &\sum_{x \in \mathrm{Re}(G_S)} \left[\sum_{\sigma \in \Sigma : \delta(x,\sigma) \in X - \mathrm{Re}(G_S)} c(x,\sigma) \right] \\ &+ \sum_{x \in \mathrm{Re}(G_S) : p(x) > 0} p(x) + \sum_{x \notin \mathrm{Re}(G_S) : p(x) < 0} -p(x). \end{aligned}$$

Since S is parsimonious, for each $x \in \mathrm{Re}(G_S)$ the set $\{\sigma \in \Sigma \mid \delta(x,\sigma) \in X - \mathrm{Re}(G_S)\} = \{\sigma \in \Sigma \mid \sigma \in S(x)\}$. Hence $\sum_{e \in E_S} u_G(e) = C(S)$. □

Proof of Theorem 2. Consider the parsimonious supervisor S induced by a min-cut of N_G. Then, from the result of Lemma 2, we obtain that $C(S)$ equals the capacity of min-cut of N_G. In view of Corollary 1, in order to prove the optimality of S it suffices to show that if S' is any other parsimonious supervisor, then $C(S') \geq C(S)$. Since S' is also parsimonious, it follows from Lemma 2 that $C(S')$ equals the capacity of cut of N_G induced by S', which is greater than or equal to the capacity of the min-cut (by definition of min-cut). Since $C(S)$ equals the capacity of the min-cut (by construction of S), we obtain the desired inequality: $C(S') \geq C(S)$. □

3.3. Solution of OSCP2. In this subsection we provide a solution for OSCP2. In this setting, given a plant G, a cost of control function c, a penalty of control function p, and an observation mask Ψ, the control objective is to determine an optimal supervisor $S : X \to 2^{\Sigma}$ satisfying the constraint C1. The constraint C1 can be satisfied by making a few modifications in G and in the cost of control function c as described below. First, we modify the state machine G; the modified state machine is denoted as G'.

DEFINITION 8. Given $G = (X, \Sigma, \delta, x_0)$, the modified state machine G' corresponding to the constraint C1 is the quadruple $G' = (X, \Sigma', \delta', x_0)$, where

1. $\Sigma' := \Sigma \cup \{\theta\}$ with $\theta \notin \Sigma$,
2. $\forall x \in X, \forall \sigma' \in \Sigma' : \delta'(x, \sigma') := \delta(x, \sigma')$ if $\sigma' \neq \theta$, and
3. $\forall \sigma \in \Sigma, x_1, x_2 \in X$ s.t. $\Psi(x_1) = \Psi(x_2), \delta(x_1, \sigma) \neq \delta(x_2, \sigma)$:
$\delta'(\delta(x_1, \sigma), \theta) := \delta(x_2, \sigma)$ and $\delta'(\delta(x_2, \sigma), \theta) := \delta(x_1, \sigma)$.

Thus G' is obtained by adding in G, an oppositely directed pair of transitions labeled θ, between the pair of states reached by executing a common event from a pair of states that look alike under Ψ. No such transition is added if the *same* state is reached after executing a common event from a pair of states that look alike under Ψ.

Next, the cost of control function $c : X \times \Sigma \to \mathcal{R}^+$ is extended to $c' : X \times \Sigma' \to \mathcal{R}^+$ as

$$\forall x \in X, \sigma' \in \Sigma' : c'(x, \sigma') := \begin{cases} c(x, \sigma') & \text{if } \sigma' \in \Sigma, \\ \infty & \text{if } \sigma' = \theta. \end{cases}$$

Thus the cost of control function c' is the extension of c obtained by assigning the cost of disabling the event θ to be infinity. For simplicity of notation, let OSCP1 with respect to G', with cost of control c' and penalty of control p be denoted as OSCP1$'$.

THEOREM 3. (Solution of OSCP2). *OSCP2 is equivalent to OSCP1$'$.*

We prove a few lemmas before proving Theorem 3.

LEMMA 3. If S is parsimonious, then S disables transitions leading from states in Re(G_S) into a state $x \in X$ if and only if it disables all transitions from states in Re(G_S) leading into x.

Proof. if and only if it disables a transition that the set $X - Re(G_S)$. Hence, S does not disable a self-loop. Also, if If S disables some transitions but not all transitions from states in Re(G_S) leading into a state $x \in X$, then x remains reachable in G_S, i.e., $x \in$ Re(G_S), which contradicts that S is parsimonious. □

LEMMA 4. There exists a solution $S : X \to 2^{\Sigma'}$ of OSCP1$'$ such that

1. S never disables the event θ, and
2. S satisfies constraint C1.

Proof. 1. Let $S : X \to 2^{\Sigma'}$ be a solution of OSCP1$'$. If $C(S) < \infty$, then it is clear that S never disables the event θ. Next consider the case when $C(S) = \infty$. If S ever disables the event θ, then consider a supervisor S' that takes the same control action as S does, except that it never disables the event θ. Then by optimality of S, $C(S') \geq C(S) = \infty$, which implies that $C(S') = \infty$. Hence S' is optimal, and it never disables the event θ.

2. Let $S : X \to 2^{\Sigma'}$ be a solution of OSCP1$'$. We show that if $x_1, x_2 \in \text{Re}(G_S)$ are such that $\Psi(x_1) = \Psi(x_2)$, then $S(x_1) = S(x_2)$. In other words, if an event $\sigma \in \Sigma$ is defined at such a pair of states $x_1, x_2 \in \text{Re}(G_S)$, then either σ is disabled at both x_1 and x_2 or it is disabled at neither of x_1 and x_2. Consider such a pair of states $x_1, x_2 \in \text{Re}(G_S)$ and an event $\sigma \in \Sigma$. Then either $\delta(x_1, \sigma) = \delta(x_2, \sigma)$ or $\delta(x_1, \sigma) \neq \delta(x_2, \sigma)$. If $\delta(x_1, \sigma) = \delta(x_2, \sigma)$, then S disables the event σ at both the states or at neither of the states. This follows from the fact that any optimal supervisor is also parsimonious (Lemma 1), and any parsimonious supervisor disables either all transitions leading into a state or none of them (Lemma 3). Thus constraint C1 is satisfied in this case. If $\delta(x_1, \sigma) \neq \delta(x_2, \sigma)$, then according to the construction of G' these states are connected by a pair of oppositely directed transitions labeled θ. Since S never disables the event θ (from part 1), the states $\delta(x_1, \sigma)$ and $\delta(x_2, \sigma)$ do not belong to separate partitions induced by S. Suppose that they both are in $\text{Re}(G_S)$; then due to parsimonicity of S, σ is enabled at both x_1 and x_2. On the other hand, if they both are in $X-$ $\text{Re}(G_S)$, then clearly, σ is disabled at x_1 and x_2. □

Proof of Theorem 3. We first show that if $S : X \to 2^{\Sigma'}$ is a solution of OSCP1$'$, then it is also a solution for OSCP2. In view of Lemma 4, it can be assumed, without loss of generality, that S never disables the event θ and satisfies C1. Since S never disables the event θ and satisfies C1, it can be viewed as a map $S : X \to 2^{\Sigma}$ satisfying C1; and hence it can also be used as a supervisor under partial state observation. Assume for the sake of contradiction that S is not a solution of OSCP2. Let $S' : X \to 2^{\Sigma}$ with $S' \neq S$ be a solution of OSCP2; then we must have $C(S') < C(S)$, where S, S' are treated as *feasible* solutions of OSCP2. Let $C'(S), C'(S')$ denote the net costs of using S and S' respectively, when S, S' are treated as *feasible* solutions of OSCP1$'$. Since (i) S and S' do not disable the event θ and (ii) for each $x \in X$ and $\sigma \in \Sigma$, $c'(x, \sigma) = c(x, \sigma)$, we have $C'(S) = C(S)$ and $C'(S') = C(S')$. Since $C(S') < C(S)$, we obtain $C'(S') < C'(S)$. This contradicts the optimality of S (treated as a solution of OSCP1$'$).

Next we show that if $S : X \to 2^{\Sigma}$ is a solution of OSCP2, then it is also a solution of OSCP1$'$. It is clear that S can also be viewed as a map $S : X \to 2^{\Sigma'}$. Assume for the sake of contradiction that S is not a solution for OSCP1$'$. Let $S' : X \to 2^{\Sigma'}$ with $S' \neq S$ be a solution of OSCP1$'$; then we must have $C'(S') < C'(S)$. Also, from Lemma 4, S' never disables the event θ and satisfies C1. Thus S' can be used as a supervisor under partial state observation. As above, $C'(S) = C(S)$ and $C'(S') = C(S')$. However, since $C'(S') < C'(S)$, we obtain $C(S') < C(S)$. This is a contradiction to the optimality of S (treated as a solution of OSCP2). □

4. Applications to supervisory control.

It is clear from Theorem 3 that an instance of OSCP2 can be reduced to an instance of OSCP1 by suitably modifying the graph of the plant and the cost of control function. We show in this section that supervisory control problems under complete as well as partial observation can also be reduced to instances of OSCP1. We begin with the problem of computing the supervisors under complete observation, which requires computation of supremal controllable sublanguage and infimal controllable superlanguage.

4.1. Computations related to controllability of DEDSs.

We first consider the computation of supremal controllable sublanguage: Given a desired prefix closed behavior $K \subseteq L(G)$, compute the *supremal* sublanguage $K^{\uparrow} \subseteq K$ such that it is *controllable* [23], i.e., $K^{\uparrow}\Sigma_u \cap L(G) \subseteq K^{\uparrow}$, where $\Sigma_u \subseteq \Sigma$ denotes the set of uncontrollable events. A closed-form expression for $K^{\uparrow}$ is given in [1], and an optimal algorithm for computing $K^{\uparrow}$ is given in [11].

In order to reduce the problem of computing $K^{\uparrow}$ to an instance of OSCP1, we refine G with respect to K so that the states corresponding to strings in K are uniquely identified. This is done as follows. Let $V := (Q, \Sigma, \rho, q_0)$ be a trim deterministic state machine that generates K. The graph of V is made "complete" by adding a dump state d to its state set. If a certain event σ is not defined at some state $q \in Q$, then a transition labeled σ from the state q to the dump state d is added. Also, execution of any event in the dump state leaves the system in that state. Formally, the *completion* of V is another state machine $V' := (Q', \Sigma, \rho', q_0)$, where $Q' := Q \cup \{d\}$ with $d \notin Q$ and

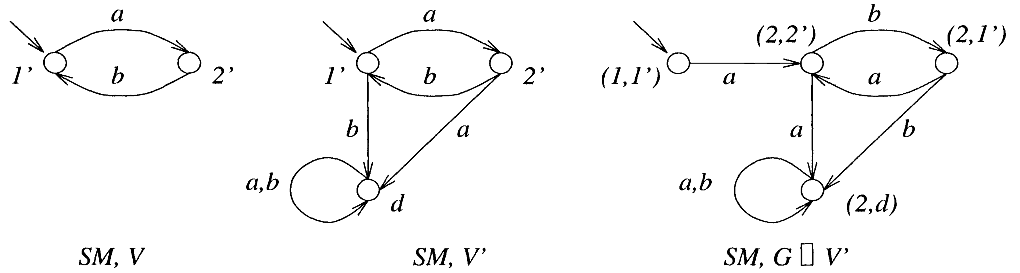
$$\forall q' \in Q', \sigma \in \Sigma : \rho'(q', \sigma) := \begin{cases} \rho(q', \sigma) & \text{if } q' \in Q \text{ and } \rho(q', \sigma) \text{ is defined,} \\ d & \text{otherwise.} \end{cases}$$

It is clear that $L(V') = \Sigma^\star$. Consider the synchronous composition [8], [11] of G and V': $G \Box V' := (X \times Q', \Sigma, \alpha, (x_0, q_0))$, where

$$\forall x \in X, q' \in Q', \sigma \in \Sigma : \alpha((x, q'), \sigma) := \begin{cases} (\delta(x, \sigma), \rho'(q', \sigma)) & \text{if } \delta(x, \sigma), \rho'(q', \sigma) \text{ ,} \\ & \text{are defined,} \\ \text{undefined} & \text{otherwise.} \end{cases}$$

It is easily shown that $L(G \Box V') = L(G) \cap L(V') = L(G) \cap \Sigma^\star = L(G)$. $G \Box V'$ is called *refinement* of G with respect to K. Note that the first coordinate of a state in $G \Box V'$ corresponds to a state of G and the second coordinate to a state of V'.

Example 3. Let G be the plant as in Example 2 and the target language K be given by $K = \overline{(ab)^\star} \subseteq L(G) = \overline{a(a+b)^\star}$. The generator V for the language K is shown in Fig. 3, with the state set $Q = \{1', 2'\}$, the initial state $q_0 = 1'$, and the transition function $\rho(1', a) = 2', \rho(2', b) = a$. The state machine V' obtained by



SM, V SM, V' SM, G □ V'

FIG. 3. *Diagram illustrating construction of* $G \Box V'$.

completing the graph of V and the state machine $G \Box V'$ obtained by synchronous composition of G and V' are both shown in Fig. 3. Note that $L(V') = (a+b)^\star = \Sigma^*$ and $L(G \Box V') = \overline{a(a+b)^\star} = L(G)$.

LEMMA 5. Given a string $s \in \Sigma^\star$, $s \in L(G) - K$ if and only if the second coordinate of the state reached by executing s in $G \Box V'$ is d.

Proof. The proof is straightforward. □

Example 4. Consider the state machine $G \Box V'$ of Example 3. Then the strings, the execution of which take to the state $(2, d)$, belong to $L(G) - K = \overline{a(a+b)^*} - \overline{(ab)^*}$.

Also, the strings, the execution of which take to states $(1, 1')$ or $(2, 2')$ or $(2, 1')$, belong to the language $K = \overline{(ab)^*}$.

The result of Lemma 5 can be used to identify unambiguously the desired and undesired states in $G\Box V'$. The next lemma proves that it is also possible to obtain the generator for $K^\uparrow$ by partitioning the graph of $G\Box V'$ into sets of reachable and unreachable states.

LEMMA 6. [11, Prop. 3.6]. Let $s, t \in K$ be such that $\alpha((x_0, q_0), s) = \alpha((x_0, q_)), t)$; then $s \in K^\uparrow$ if and only if $t \in K^\uparrow$.

Based on the result of Lemma 5, we define a penalty of control function $p : X \times Q' \to \mathcal{R}$ for $G\Box V'$ as

$$(1) \qquad \forall (x, q') \in X \times Q' : p((x, q')) := \begin{cases} \infty & \text{if } q' = d, \\ -p_0 & \text{otherwise,} \end{cases}$$

where $p_0 \in \mathcal{R}^+$ is any positive real. Since the penalty of control of a state in $G\Box V'$ with second coordinate d is infinity, it should remain unreachable in an optimally controlled plant; and since the penalty of control of all other states is $-p_0$, as many such states as possible should remain reachable in an optimally controlled plant.

Example 5. Consider the state machine $G\Box V'$ of Example 3. Then in order to compute the supremal controllable sublanguage of $K = \overline{(ab)^*}$ with respect to G of Example 2, we define the penalty of control function as $p[(2, d)] = \infty, p[(1, 1')] = p[(2, 2')] = p[(2, 1')] = -p_0$.

Next we define a cost of control function $c : X \times \Sigma \to \mathcal{R}^+$ as

$$(2) \qquad \forall x \in X, \sigma \in \Sigma : c(x, \sigma) := \begin{cases} \infty & \text{if } \sigma \in \Sigma_u, \\ \frac{p_0}{|e|+1} & \text{otherwise,} \end{cases}$$

where $|e|$ denotes the total number of transitions in the graph of $G\Box V'$. Since the cost of control of an uncontrollable event is infinity, it should not be disabled by an optimal supervisor; and since the cost of control of a controllable event is $\frac{p_0}{|e|+1}$, as few such events as possible should be disabled. With the above definitions of cost and penalty of control functions, we prove in the following theorem that the computation of $K^\uparrow$ can be posed as an instance of OSCP1.

THEOREM 4. If $K^\uparrow \neq \emptyset$, then $K^\uparrow$ equals the language generated by $G\Box V'$ under the control of a solution of OSCP1 with respect to $G\Box V'$, with cost of control of function as in equation 2 and penalty of control function as in equation 1.

Proof. Let $S : X \times Q' \to 2^\Sigma$ be the parsimonious supervisor induced by the partition of $G\Box V'$ into the set of states that correspond to the supremal controllable sublanguage and the set of remaining states, i.e., S disables those transitions that are defined from states corresponding to the supremal controllable sublanguage to the set of remaining states. That such a partition exists follows from Lemma 6 and the fact that $K^\uparrow \neq \emptyset$. It is clear that S does not disable any uncontrollable events (otherwise the controlled system behavior is not a controllable language), and no state with second coordinate d remains reachable under the control of S (otherwise the controlled system behavior is not a sublanguage of K). Hence $C(S) < \infty$. Let $S' : X \times Q' \to 2^\Sigma$ be a solution of OSCP1. Then it follows from the optimality of S' that $C(S') \leq C(S) < \infty$. Thus S' does not disable any uncontrollable event and no state with second coordinate d remains reachable in the controlled system (otherwise $C(S') = \infty$). Since S' does not disable any of the uncontrollable events, the controlled system behavior under its control is controllable [11, Lemma 2.7]. Also, since no state with second coordinate d remains reachable under the control of S', the controlled system behavior under the control of S' is a sublanguage of K. Assume for contradiction that the controlled system behavior under the control of S' is not the supremal controllable sublanguage of K. So there exists at least one state with second coordinate unequal to d such that it is reachable under the control of S and unreachable under the control of S'. Hence $C(S') - C(S) \geq p_0 - n\frac{p_0}{|e|+1}$, where n is the difference between the number of controllable transitions disabled by S and the number of controllable transitions disabled by S'. Since $n \leq |e|$, the total number of transitions in $G\Box V'$, $n\frac{p_0}{|e|+1} < p_0$. In other words, $C(S') - C(S) > 0$. This is a contradiction to the optimality of S'. □

Next we consider the problem of computing the infimal controllable superlanguage: Given a desired prefix closed behavior $K \subseteq L(G)$, compute the *infimal* superlanguage $K^{\downarrow} \supseteq K$ such that it is *controllable*. A closed-form expression for $K^{\downarrow}$ and an algorithm for computing it is given in [14]. We use the same notation as above. The following modification is made to the penalty of control function:

$$(3) \qquad \forall (x,q') \in X \times Q' : p((x,q')) := \begin{cases} p_0 & \text{if } q' = d, \\ -\infty & \text{otherwise} \end{cases}$$

Since penalty of control of a state with second coordinate d is p_0, as few such states as possible should remain reachable in the controlled system; and since the penalty of control of a state with second coordinate unequal to d is $-\infty$, all such states should remain reachable in the controlled system.

Example 6. Consider the state machine $G\Box V'$ of Example 3. Then in order to compute the infimal controllable superlanguage of $K = \overline{(ab)^*}$ with respect to G of Example 2, we define the penalty of control function as $p[(2,d)] = p_0, p[(1,1')] = p[(2,2')] = p[(2,1')] = -\infty$.

With the above modification in the penalty function, we prove in the following theorem that the computation of $K^{\downarrow}$ can be posed as an instance of OSCP1.

THEOREM 5. $K^{\downarrow}$ equals the language generated by $G\Box V'$ under the control of a solution of OSCP1 with respect to $G\Box V'$, with cost of control function as in equation 2 and penalty of control function as in equation 3.

Proof. The proof is similar to that of Theorem 4. □

4.2. Computations related to observability of DEDSs. Suppose that the supervisor's observation of events is filtered through a mask of the type $M : \Sigma \rightarrow \Lambda \cup \{\epsilon\}$. Computation of a supervisor under such a partial observation requires computation of languages such as supremal normal sublanguage, infimal normal/observable superlanguage, supremal controllable and normal sublanguage, infimal controllable and normal/observable superlanguage, etc. We show that each of these computations can be reduced to instances of OSCP1. Our techniques illustrate how supervisory control under partial observation can be solved using a state-feedback type control on a suitably refined state machine representation of the plant.

We first consider the problem of computing the supremal normal sublanguage: Given a desired prefix closed language $K \subseteq L(G)$, compute the *supremal* sublanguage $K^{\circ} \subseteq K$, such that it is *normal*, i.e., $M^{-1}(M(K^{\circ})) \cap L(G) \subseteq K^{\circ}$. The existence of K° is shown in [16], and a closed form expression for K° is given in [1], [11]. We first refine the state machine G with respect to V (the generator for K) and mask function M so that the states corresponding to K° are uniquely identified. We begin by constructing the machine $G_1 := G\Box V'$. Recall that $L(G_1) = L(G)$. Using G_1 we construct a machine that generates the language $M^{-1}M(L(G_1))$ by employing the following algorithm:

ALGORITHM 1.

1. Replace each transition $\sigma \in \Sigma$ in G_1 by the transition $M(\sigma) \in \Lambda \cup \{\epsilon\}$. Call this machine G_2; clearly, $L(G_2) = M(L(G_1))$.
2. Construct a deterministic machine G_3 that is language equivalent to G_2 [7]. Then the state space of G_3 is $2^{X\times Q'}$ and $L(G_3) = L(G_2) = M(L(G_1))$.
3. Replace each transition $\lambda \in \Lambda$ in machine G_3 by the events in the set $M^{-1}(\lambda) := \{\sigma \in \Sigma \mid M(\sigma) = \lambda\}$. Also, at each state in G_3, add self-loops corresponding to the events in the set $M^{-1}(\epsilon) = \{\sigma \in \Sigma \mid M(\sigma) = \epsilon\}$. Call this machine G_4; clearly, $L(G_4) = M^{-1}(L(G_3)) = M^{-1}(L(G_2)) = M^{-1}M(L(G_1))$.

G_4 has a nice property that if two strings $s, t \in L(G_4)$ are such that $M(s) = M(t)$, then the state reached by executing them are the same. This follows from the observations that (i) the state reached by executing s in G_4 is the same as that reached by executing $M(s)$ in G_3 (by construction); (ii) since $M(s) = M(t)$ and G_3 is deterministic, the same state is reached by executing $M(t)$ in G_3; and (iii) by construction, this is the state reached by t in G_4. We exploit the above property of G_4 in identifying the states corresponding to the strings in the language K°. This, however, requires the construction of machine $G_5 := G_1 \Box G_4$, for which it is clear that $L(G_5) = L(G_1) \cap L(G_4) = L(G) \cap M^{-1}M(L(G)) = L(G)$ and the state space of G_5 equals $X \times Q' \times 2^{X\times Q'}$. Construction of state machines G_1 through G_5, their state spaces, and their languages are summarized in Table 1.

TABLE 1
Various machines used for computation of K°

SM	Construction	State space	Language
G_1	$G\Box V'$	$X \times Q'$	$L(G)$
G_2	$M(G_1)$	$X \times Q'$	$M(L(G))$
G_3	$det(G_2)$	$2^{X\times Q'}$	$M(L(G))$
G_4	$M^{-1}(G_3)$	$2^{X\times Q'}$	$M^{-1}(M(L(G))$
G_5	$G_1\Box G_4$	$X \times Q' \times 2^{X\times Q'}$	$L(G)$

Note that in Table 1 and Fig. 4, we have used the notation (i) $M(G_1)$ to represent that G_2 is obtained by "masking" the transitions of G_1, (ii) $\det(G_2)$ to represent that G_3 is obtained by "determinizing" G_2, and (iii) $M^{-1}(G_3)$ to represent that G_4 is obtained by "unmasking" the transitions of G_3.

Example 7. Consider the state machine $G_1 = G\Box V'$ of Example 3. Let the mask $M : \Sigma = \{a, b\} \to \{\lambda\}$ be defined as $M(a) = M(b) = \lambda$. Then the state machines G_2, G_3, G_4, and G_5 obtained by using Algorithm 1 are shown in Fig. 4.

We use $r = ((x, q'), \{(x_1, q_1'), (x_2, q_2'), \ldots, (x_r, q_r')\}) \in X \times Q' \times 2^{X\times Q'}$ to denote a typical state of G_5, where $(x, q') \in X\times Q'$ and $\{(x_1, q_1'), (x_2, q_2'), \ldots, (x_r, q_r')\} \in 2^{X\times Q'}$. We call $r' := (x, q')$ the G_1 part of r and $R := \{(x_1, q_1'), (x_2, q_2'), \ldots, (x_r, q_r')\}$ the G_4 part of r. We prove in the next lemma that if r_1 and r_2 are two states of G_5 with identical G_4 part, then corresponding to each string, the execution of which takes to the state r_1, there exists another string, the execution of which takes to the state r_2, such that it looks like the former string. We call such a pair of states to be a matching pair. Formally, consider the following definition.

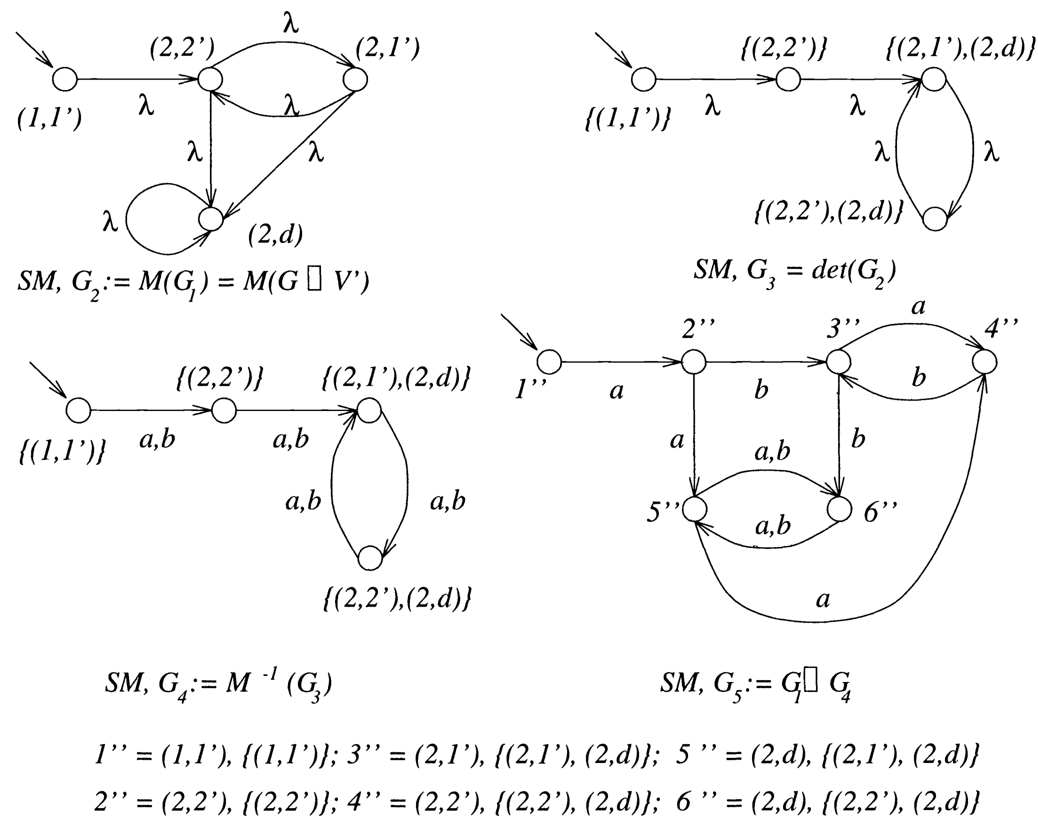

(1,1')
(2,2')
(2,1')
(2,d)
λ
SM, G2:= M(G1) = M(G □ V')
{(1,1')}
{(2,2')}
{(2,1'),(2,d)}
{(2,2'),(2,d)}
SM, G3 = det(G2)
a,b
SM, G4:= M -1 (G3)
1''
2''
3''
4''
5''
6''
a
b
a,b
SM, G5:= G1□ G4
1'' = (1,1'), {(1,1')}; 3'' = (2,1'), {(2,1'), (2,d)}; 5 '' = (2,d), {(2,1'), (2,d)}
2'' = (2,2'), {(2,2')}; 4'' = (2,2'), {(2,2'), (2,d)}; 6 '' = (2,d), {(2,2'), (2,d)}

FIG. 4. *Diagram illustrating construction of* G_2, G_3, G_4, G_5.

DEFINITION 9. Let $r_1 = (r_1', R_1)$ and $r_2 = (r_2', R_2) \in X \times Q' \times 2^{X \times Q'}$ be such that $R_1 = R_2$. Then the pair r_1 and r_2 of states is called a *matching* pair of states.

Example 8. Consider the SM G_5 of Example 7. The states $3'' = (2,1'), \{(2,1'), (2,d)\}$ and $5'' = (2,d), \{(2,1'), (2,d)\}$ constitute a matching pair of states. Similarly, the pair of states $4'' = (2,2'), \{(2,2'), (2,d)\}$ and $6'' = (2,d), \{(2,2'), (2,d)\}$ is another matching pair of states.

LEMMA 7. Let $r_1, r_2 \in X \times Q' \times 2^{X \times Q'}$ be a matching pair of states. Then given a string s, the execution of which takes to r_1, there exists a string t, the execution of which takes to r_2, such that $M(s) = M(t)$.

Proof. First note that $r = (r', R)$ is a reachable state of G_5 if and only if $r' \in R$, i.e., if and only if the G_1 part of r is an element of the G_4 part of r. This follows from (i) if $s \in L(G_5)$ is a string, the execution of which takes to r, then the execution of s takes to the state r' in G_1 and to the state R in G_4, and (ii) the execution of s takes to the state $R = \{(x_1, q_1'), (x_2, q_2'), \ldots, (x_r, q_r')\}$ in G_4 implies that there exists a state $(x_j, q_j') \in R$ such that the execution of s takes to the state (x_j, q_j') in G_1.

Consider then the states $r_1, r_2 \in X \times Q' \times 2^{X \times Q'}$ such that $R_1 = R_2 := R$. Then $r_1 = (r_1', R)$ and $r_2 = (r_2', R)$. It follows from the discussion in the preceding paragraph that $r_1' \in R$ and $r_2' \in R$. Let $s \in L(G_5)$ be a string, the execution of which takes to state r_1 in G_5; then the execution of s takes to the state r_1' in G_1 and to the state R in G_4. Since both the states $r_1', r_2' \in R$, there exists at least one string t, the execution of which takes to state r_2' in G_1 and to state R in G_4, such that

$M(t) = M(s)$. This follows from: (i) states r'_1 and r'_2 of G_2 belong to the same state R of G_3 if and only if there exists a string in $L(G_3) = L(G_2) \subseteq \Lambda^\star$, the execution of which takes to state R in G_3 and the execution of it takes to both states r'_1, r'_2 in G_2 (note that G_2 is a nondeterministic machine in general); and (ii) a string, the execution of which takes to the states r'_1, r'_2 in G_2, corresponds to two different strings in $L(G_1)$ having the same mask value. □

The result of Lemma 7 can be used to identify the strings in K°. Note that a string $s \in K - K^\circ$ if and only if there exists a string $t \in L(G) - K$ such that $M(t) = M(s)$. A state $r = (r', R)$ in G_5 corresponds to strings in $L(G) - K$ if and only if $r' = (x, d)$ for some $x \in X$. Thus strings in K and those in $L(G) - K$ can easily be identified in G_5. Let r_1 and r_2 be a matching pair of states in G_5, with $R_1 = R_2 := R$, such that the second coordinate of r'_1 does not equal d whereas the second coordinate of r'_2 equals d. Then as discussed above, strings leading to r_1 belong to K and those leading to r_2 are in $L(G) - K$. Moreover, strings leading to r_1 are in $K - K^\circ$. This follows from Lemma 7, which asserts that corresponding to each string that leads to r_1 (i.e., the string is in K), there exists a string leading to r_2 (i.e., this string is in $L(G) - K$) such that it looks like the former string. Thus states corresponding to $K - K^\circ$ can also be identified in G_5 by first identifying all those matching pair of states for which exactly one of the states in each pair has its second coordinate of the G_1 part equal to d and then, among these matching pair of states, determining those states for which the second coordinate of the G_1 part does not equal d. Hence it is possible to obtain the generator for K° by partitioning the graph of G_5 into the set of reachable and unreachable states. Finally, we pose the problem of computing the supremal normal sublanguage as an instance of OSCP1 with respect to the machine obtained by adding an equally directed pair of transitions labeled θ between each matching pair of states in G_5. We call the machine thus obtained G'_5.

Define the following cost of control function for G'_5:

$$(4) \qquad \forall r \in X \times Q' \times 2^{X \times Q'}, \sigma' \in \Sigma \cup \{\theta\} : c(r, \sigma') := \begin{cases} \infty & \text{if } \sigma' = \theta, \\ \frac{p_0}{|e|+1} & \text{otherwise,} \end{cases}$$

where $|e|$ denotes the total number of transitions in G'_5. The cost of control function for the event θ is infinity. Such a cost of control function ensures that the matching pair of states remains in the same partition of states induced by an optimal supervisor. Define the following penalty of control function on G'_5:

$$(5) \qquad \forall r = (r', R) \in X \times Q' \times 2^{X \times Q'} : p(r) := \begin{cases} \infty & \text{if } r' \in X \times \{d\}, \\ -p_0 & \text{otherwise.} \end{cases}$$

Thus the penalty of control is positive infinity whenever a state corresponds to strings in $L(G) - K$. Such states are undesired and should remain unreachable in the controlled plant under an optimal supervision. Other states have a negative penalty of control, $-p_0$, implying that such states are desired, and as many of them as possible should remain reachable in the controlled plant.

Example 9. Consider the state machine G_5 of Example 7. Then in order to compute the supremal normal sublanguage of the language $K = \overline{(ab)^*}$ with respect to plant G of Example 2 and mask $M(a) = M(b) = \lambda$, G'_5 is constructed by adding, in G_5, a pair of oppositely directed transitions labeled θ between both the pair of matching states, namely, between $3''$ and $5''$, and between $4''$ and $6''$. The cost of disabling θ is assigned to be infinity. Finally, the penalty of control function is defined to be infinity for the states $5''$ and $6''$ and $-p_0$ for the remaining states, $1'', 2'', 3''$, and $4''$. Note that the states $5''$ and $6''$ are such that for them the second part of the G_1 part equals d.

THEOREM 6. If $K^\circ \neq \emptyset$, then K° equals the controlled plant behavior of G'_5 under the control of a solution of OSCP1 with respect to G'_5, with a cost of control function as in equation 4 and penalty of control function as in equation 5.

Proof. The proof proceeds similarly to that of Theorem 4. The key to the proof is that each matching pair of states belong to the same partition induced by an optimal supervisor. □

It can be shown that the cost of control function in equation 4 can be modified slightly to compute the supremal controllable and normal sublanguage of K:

$$(6)\quad \forall r \in X \times Q' \times 2^{X \times Q'}, \sigma' \in \Sigma \cup \{\theta\} : c(r, \sigma') := \begin{cases} \infty & \text{if } \sigma' \in \Sigma_u \cup \{\theta\}, \\ \frac{p_0}{|e|+1} & \text{otherwise.} \end{cases}$$

Also, the penalty of control function in equation 5 can be replaced by the following penalty of control function to compute the infimal controllable and normal superlanguage of K:

$$(7)\qquad \forall r = (r', R) \in X \times Q' \times 2^{X \times Q'} : p(r) := \begin{cases} p_0 & \text{if } r' \in X \times \{d\}, \\ -\infty & \text{otherwise.} \end{cases}$$

THEOREM 7. If the supremal controllable and normal sublanguage of K is nonempty, then it equals the controlled plant behavior of G'_5 under the control of a solution of OSCP1 with respect to G'_5, with cost of control function as in equation 6 and penalty of control function as in equation 5. Furthermore, if instead penalty of control function as in equation 7 is used, then the controlled plant behavior equals the infimal controllable and normal superlanguage of K.

Proof. The proof is similar to that of Theorem 4. □

Finally we show that the computation of the *infimal observable* superlanguage of K can be posed as an instance of OSCP2. Refer to [16] for a detailed discussion of observable languages and their properties. It is shown in [16] that the infimal observable superlanguage of a prefix closed language exists, and a closed-form expression for computing it is obtained in [25], [9]. Observability of a language K requires that whenever a pair of strings belonging to K and having the same mask value are extended by a common event, either both the resulting strings belong to K or do not belong to K. This condition is needed so that the supervisor can take the same control action after execution of a pair of strings that have the same mask value. If K does not satisfy this property, then the infimal observable superlanguage of K is computed, which satisfies such a property.

The notion of pairs of strings in K with the same mask value is captured by the matching pair of states having the second coordinate of the G_1 part unequal to d. Let $r_1, r_2 \in X \times Q' \times 2^{X \times Q'}$ be a matching pair of states so that the second coordinate of the G_1 part of both the states is unequal to d. Since r_1 and r_2 is a matching pair of states, Lemma 7 implies that, corresponding to each string that leads to the state r_1, there exists a string having the same mask value as of the former string, such that it leads to the state r_2, and vice versa. Since the supervisor must take the same control action after the execution of a pair of strings that have the same mask value, an event is enabled at state r_1 if and only if it is enabled at r_2. This constraint is similar to the constraint C1 and can be captured by defining a mask function Ψ on the state space of G_5 as follows:

$$(8)\quad \forall r_1 = (r'_1, R_1), r_2 = (r'_2, R_2) \in X \times Q' \times 2^{X \times Q'} : \Psi(r_1) = \Psi(r_2) \Leftrightarrow R_1 = R_2.$$

Thus two states in state space of G_5 have the same mask value if and only if they constitute a matching pair. Hence, according to constraint C1 of OSCP2, it is ensured that the same control action is taken at any matching pair of states. Next the cost of control function is defined as

$$(9) \qquad \forall r \in X \times Q' \times 2^{X \times Q'}, \sigma \in \Sigma : c(r, \sigma) := \frac{p_0}{|e| + 1}$$

where $|e|$ denotes the number of transitions in G_5.

THEOREM 8. The infimal observable superlanguage of K equals the controlled plant behavior of G_5 under the control of a solution of OSCP2 with respect to G_5, with cost of control function as in equation 9, penalty of control function as in equation 7, and state mask function Ψ as in equation 8. Furthermore, if the cost of control function is modified so that $c(\cdot, \sigma) = \infty$ whenever $\sigma \in \Sigma_u$, then the controlled plant behavior equals the infimal controllable and observable superlanguage of K.

Proof. The proof is similar to that of Theorem 4. □

Example 10. Consider the state machine G_5 of Example 7. Then in order to compute the infimal observable superlanguage of the language $K = \overline{(ab)^*}$ with respect to the plant G of Example 2 and mask $M(a) = M(b) = \lambda$, we define the mask Ψ on the state space G_5 such that $\Psi(3'') = \Psi(5'')$ and $\Psi(4'') = \Psi(6'')$. Next the penalty of control function is defined to be negative infinity for the states 1", 2", 3", and 4". The penalty of control for the states 5" and 6" is defined to be p_0.

Remark 3. An advantage of using the above techniques to compute controllable and normal/observable sublanguages/superlanguages is that they do not require alternate computations of controllable and normal/observable sublanguages/superlanguages as they do in [4]. Note that if M is a *projection* type mask, then the formula in [1] can be used to compute the supremal controllable and normal sublanguage without having to perform alternate computations of supremal controllable and supremal normal sublanguages. However, if the mask is nonprojection type, then no such formula is known, and techniques developed above can be used. reduction of the overall computational complexity.

In case M is a nonprojection type mask, the fact that a computation of supremal controllable sublanguage followed by a computation of supremal normal sublanguage does not necessarily yield the supremal controllable and normal sublanguage can be illustrated as follows. Suppose $\Sigma = \{a, b, c, u\}, \Sigma_u = \{u\}, M(b) = M(c) = M(u) \neq \epsilon$, $K = \{\epsilon, a, au\}$, and $L(G) = \{\epsilon, a, au, ab, ac, acu\}$. Clearly K is controllable, i.e., $K^\uparrow = K$; and K is not normal, as $au \in K, ab \in L(G) - K$ and $M(au) = M(ab)$, i.e., $K^\circ \neq K$. It is easily seen that $K^\circ = \{\epsilon, a\}$. Then K° is not controllable, as $a \in K^\circ, u \in \Sigma_u$, and $au \in L(P) - K^\circ$. Thus $(K^\uparrow)^\circ = K^\circ$ does not equal the supremal controllable and normal sublanguage of K. On the other hand, if we let $\hat{K} = \{\epsilon, a, au, ab, ac\}$, then $\hat{K}$ is normal, i.e., $\hat{K}^\circ = \hat{K}$; and $\hat{K}$ is not controllable, i.e., $\hat{K}^\uparrow \neq \hat{K}$, as $ac \in \hat{K}, u \in \Sigma_u$, and $acu \in L(G) - \hat{K}$. It is easily seen that $\hat{K}^\uparrow = \{\epsilon, a, au, ab\}$. Then $\hat{K}^\uparrow$ is not normal, as $ab \in \hat{K}^\uparrow, ac \in L(G) - \hat{K}^\uparrow$, and $M(ab) = M(ac)$. Thus $(\hat{K}^\circ)^\uparrow = \hat{K}^\uparrow$ does not equal the supremal controllable and normal sublanguage of $\hat{K}$. Also, note that the formula for computing the supremal controllable and normal sublanguage given in [1, Thm. 4] is only applicable in a setting where, whenever a controllable and an uncontrollable events have "nonepsilon" mask values, then their mask values are different; so that the set of "masked uncontrollable" events is unambiguously identified. Since $u \in \Sigma_u$ and $b, c \in \Sigma - \Sigma_u$ are such that $M(b) = M(c) = M(u) \neq \epsilon$, the formula of [1, Theorem 4] is not applicable here.

5. Conclusion. We introduced the problem of optimal supervisory control for DEDS by introducing the notions of cost and penalty of using a controller. Cost of control is incurred when an event is disabled by a controller, and penalty of control is incurred whenever undesired states remain reachable or desired states remain unreachable in the controlled plant. The control objective is to optimize the net cost of control. This is formulated as OSCP1 for the case of complete state observation and OSCP2 for the case of incomplete state observation. We show that a solution to OSCP1 can be obtained as a min-cut of an associated flow network and a solution for OSCP2 is obtained by reducing an instance of OSCP2 to an instance of OSCP1. We show that supervisory control problems under complete as well as partial observations can be reduced to instances of OSCP1. In particular, we provide techniques for the computation of supremal controllable and normal sublanguage and infimal controllable and normal/observable superlanguage without having to perform alternate computations of controllable and normal/observable languages until a fixed point is reached. Thus, the above theory serves as a unified computational framework for supervisory control problems.

We did not comment on the computational complexity of any of the algorithms derived in this paper. However, since (i) all the algorithms developed in this paper are instances of OSCP1 and (ii) OSCP1 is solved using the max-flow min-cut computation, the computational complexity of any of the algorithms presented in this paper can be obtained from that of the max-flow min-cut computation, which is $O(|v| \cdot |e| \log(|v|^2/|e|))$, where $|v|$ denotes the number of vertices and $|e|$ denotes the number of edges in the underlying flow network. Note that we are not suggesting that the computation of a supervisor under partial observation can be performed in a polynomial time; as in case of partial observation, OSCP1 is solved with respect to a state machine having its state space as the *power set* of the state space of the plant composed with the generator of the desired behavior. like to thank anonymous reviewers for their helpful comments.

REFERENCES

[1] R. D. Brandt, V. K. Garg, R. Kumar, F. Lin, S. I. Marcus, and W. M. Wonham, *Formulas for calculating supremal controllable and normal sublanguages*, Systems Control Lett., 15 (1990), pp. 111–117.

[2] Y. Brave and M. Heymann, *On optimal attraction in discrete event processes*, Tech. report CIS 9010, Technion—Israel Institute of Technology, Haifa, Israel 32000, 1990.

[3] ———, *On stabilization of discrete event processes*, Internat. J. Control, 51 (1990), pp. 1101–1117.

[4] H. Cho and S. I. Marcus, *On supremal languages of class of sublanguages that arise in supervisor synthesis problems with partial observations*, Math. Control Signals Systems, 2 (1989), pp. 47–69.

[5] R. Cieslak, C. Desclaux, A. Fawaz, and P. Varaiya, *Supervisory control of discrete event processes with partial observation*, IEEE Trans. Automat. Control, 33 (1988), pp. 249–260.

[6] V. K. Garg and R. Kumar, *State-variable approach for controlling discrete event systems with infinite states*, in Proceedings of 1992 American Control Conference, Chicago, IL, July 1992, pp. 2809–2813.

[7] J. E. Hopcroft and J. D. Ullman, *Introduction to Automata Theory, Languages and Computation*, Addison-Wesley, Reading, MA, 1979.

[8] R. Kumar, *Supervisory Synthesis Techniques for Discrete Event Dynamical Systems: Transition Model Based Approach*, PhD thesis, Department of Electrical and Computer Engineering, University of Texas at Austin, 1991.

[9] ———, *Formulas for observability of discrete event dynamical systems*, in Proceedings of 1993 Conference on Information Sciences and Systems, Johns Hopkins University, Baltimore, MD, March 1993, pp. 581–586.

[10] R. KUMAR AND V. K. GARG, *Optimal control of discrete event dynamical systems using network flow techniques*, in Proceedings of 1991 Annual Allerton Conference, Urbana, IL, October 1991, pp. 705–714.

[11] R. KUMAR, V. K. GARG, AND S. I. MARCUS, *On controllability and normality of discrete event dynamical systems*, Systems Control Lett., 17 (1991), pp. 157–168.

[12] ———, *Language stability and stabilizability of discrete event dynamical systems*, SIAM J. Control Optim., 31 (1993), pp. 1294–1320.

[13] ———, *Predicates and predicate transformers for supervisory control of discrete event systems*, IEEE Trans. Automat. Control, 38 (1993), pp. 232–247.

[14] S. LAFORTUNE AND E. CHEN, *On the infimal closed and controllable superlanguage of a given language*, IEEE Trans. Automat. Control, 35 (1990), pp. 398–404.

[15] E. LAWLER, *Combinatorial Optimization—Networks and Matroids*, Holt, Rinehart, and Winston, New York, 1976.

[16] F. LIN AND W. M. WONHAM, *On observability of discrete-event systems*, Inform. Sci., 44 (1988), pp. 173–198.

[17] C. M. OZVEREN AND A. S. WILLSKY, *Observability of discrete event dynamical systems*, IEEE Trans. Automat. Control, 35 (1990), pp. 797–806.

[18] C. M. OZVEREN, A. S. WILLSKY, AND P. J. ANTSAKLIS, *Stability and stabilizability of discrete event dynamical systems*, J. Assoc. Comput. Mach., 38 (1991), pp. 730–752.

[19] K. M. PASSINO AND P. J. ANTSAKLIS, *Near-optimal control of discrete event systems*, in Proceedings of 1989 Allerton Conference, Allerton, IL, September 1989, pp. 915–924.

[20] ———, *On the optimal control of discrete event systems*, in Proceedings of 1989 IEEE Conference on Decision and Control, Tampa, FL, December 1989, pp. 2713–2718.

[21] P. J. RAMADGE AND W. M. WONHAM, *Modular feedback logic for discrete event systems*, SIAM J. Control Optim., 25 (1987), pp. 1202–1218.

[22] ———, *On the supremal controllable sublanguage of a given language*, SIAM J. Control Optim., 25 (1987), pp. 637–659.

[23] ———, *Supervisory control of a class of discrete event processes*, SIAM J. Control Optim., 25 (1987), pp. 206–230.

[24] ———, *The control of discrete event systems*, Proc. IEEE: Special Issue on Discrete Event Systems, 77 (1989), pp. 81–98.

[25] K. RUDIE AND W. M. WONHAM, *The infimal prefix closed and observable superlanguage of a given language*, Systems Control Lett., 15 (1990), pp. 361–371.

[26] R. SENGUPTA AND S. LAFORTUNE, *A graph-theortic optimal control problem for terminating discrete event processes*, Discrete Event Dynamic Systems: Theory Appl. 2, 1992, pp. 139–172.

[27] J. N. TSITSIKLIS, *On the control of discrete event dynamical systems*, Math. Control Signals Systems, 2 (1989), pp. 95–107.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 33, No. 2, pp.440–454, March 1995

© 1995 Society for Industrial and Applied Mathematics

006

UNIFORM STABILIZATION OF A HYBRID SYSTEM OF ELASTICITY*

BOPENG RAO†

Abstract. The problem of boundary feedback stabilization of a Euler–Bernoulli beam with an endmass is considered. Using a method of compact perturbation, the lack of uniform stabilization is proved in the case of a clamped beam with the usual boundary feedbacks applied to the end with the mass. Next, the uniform stabilization when the usual boundary feedbacks are applied to the end without the mass is proved. Also the uniform decay of energy by means of higher-order feedbacks applied to the end with the mass is established.

Key words. hybrid system, lack of uniform stabilization, compact perturbation, higher-order feedbacks, boundary multipliers, uniform decay of energy

AMS subject classifications. 35B40, 35M10, 93C15, 93C20, 93D15, 93D30

1. Introduction. The purpose of this work (the results of which were announced in Rao [14]) is to study the boundary feedback stabilization of the well-known SCOLE model. Consisting of an elastic beam, linked to a rigid antenna, this dynamical system can be described by the Euler–Bernoulli equation for the vibrations of the elastic beam and the Newton–Euler rigid-body equations for the oscillations of the antenna

$$\begin{cases} y_{tt} + y_{xxxx} = 0, \\ \mu_1 y_{tt}(1,t) - y_{xxx}(1,t) = L_1(y, y_t), \\ \mu_2 y_{xtt}(1,t) + y_{xx}(1,t) = L_2(y, y_t) \end{cases}$$

where L_1, L_2 denote linear boundary feedbacks acting on the antenna and where μ_1, μ_2 are positive constants. This system is composed of one partial differential equation and two ordinary differential equations and called a hybrid system. For further descriptions concerning the physical structure of the system, we refer to Littman–Markus [10]. Our goal is to choose suitable boundary damping at the end $x = 0$ and boundary feedbacks L_1, L_2 applied to the end $x = 1$ such that the hybrid system can be stabilized uniformly.

In one specific case, Littman–Markus [9] proved the strong stabilization together with the lack of uniform stabilization. Lack of uniform stabilization for hybrid systems was also observed in the string/mass model (cf. Lee–You [7]). However, to the author's knowledge, there is no positive result in the literature concerning uniform stabilization of hybrid systems.

Using an energy multipliers method (cf. Rao [13]), we recently obtained the uniform stabilization for a string/mass model. The idea of the proof is to apply the usual boundary feedbacks to the end of the string without the mass.

Let us outline briefly the content of this work. In §2, using a method of compact perturbation, we prove the lack of uniform stabilization in the case of a clamped beam with the usual boundary feedbacks L_1, L_2. Thus we generalize the result of Littman–Markus mentioned above. In §3, we consider a beam with usual boundary feedbacks applied to the end without the mass. We first formulate the problem by means of the semigroup approach. Next we prove that, in that case, the usual boundary feedbacks are sufficient to obtain uniform exponential decay. Section 4 is devoted to the study of the clamped beam. We establish the uniform exponential decay of energy by higher-order feedbacks applied to the end with the mass.

* Received by the editors July 12, 1993; accepted for publication (in revised form) November 8, 1993.

† Institut de Recherche Mathématique Avancée, Université Louis Pasteur et Centre National de la Recherche Scientifique, 7, Rue René Descartes, 67084 Strasbourg, France.

2. Lack of uniform stabilization. We first consider the dynamical system of the SCOLE model in the case of a clamped beam

$$(2.1)\qquad \begin{cases} y_{tt} + y_{xxxx} = 0, & t > 0, \quad 0 < x < 1, \\ y(0,t) = y_x(0,t) = 0, & t > 0, \\ \mu_1 y_{tt}(1,t) - y_{xxx}(1,t) = L_1(y, y_t), & t > 0, \\ \mu_2 y_{xtt}(1,t) + y_{xx}(1,t) = L_2(y, y_t), & t > 0. \end{cases}$$

We prove the lack of uniform stabilization for the usual boundary feedbacks

$$(2.2)\qquad \begin{cases} L_1(y, y_t) = -\delta_{11} y(1,t) - \delta_{12} y_x(1,t) - \nu_{11} y_t(1,t) - \nu_{12} y_{xt}(1,t), \\ L_2(y, y_t) = -\delta_{21} y(1,t) - \delta_{22} y_x(1,t) - \nu_{21} y_t(1,t) - \nu_{22} y_{xt}(1,t), \end{cases}$$

where the coefficients δ_{ij}, ν_{ij} are real numbers.

Let y be a smooth solution of the system (2.1) – (2.2). After a procedure of Slemrod [18], we introduce the auxiliary functions:

$$z(x,t) = y_t(x,t), \quad \xi(t) = y_t(1,t), \quad \eta(t) = y_{xt}(1,t), \quad u(t) = (y(t), z(t), \xi(t), \eta(t)).$$

Then we can write formally the hybrid system (2.1) – (2.2) into the following form:

$$(2.3)\qquad \begin{aligned} &(y_t(t),\, z_t(t),\, \xi_t(t),\, \eta_t(t)) + \left(-z(t),\, y_{xxxx}(t), -\frac{1}{\mu_1} y_{xxx}(1,t),\, \frac{1}{\mu_2} y_{xx}(1,t)\right) \\ &\quad = (0,\, 0,\, L_1(y, y_t),\, L_2(y, y_t)). \end{aligned}$$

Now let us introduce the energy space E

$$E = \{(y, z, \xi, \eta) \in H^2(0,1) \times L^2(0,1) \times \mathbb{R} \times \mathbb{R} \text{ such that } y(0) = y_x(0) = 0\}.$$

For any $u = (y, z, \xi, \eta) \in E$ and any $\tilde{u} = (\tilde{y}, \tilde{z}, \tilde{\xi}, \tilde{\eta}) \in E$, we define the inner product

$$\langle u, \tilde{u} \rangle = \langle (y, z, \xi, \eta), (\tilde{y}, \tilde{z}, \tilde{\xi}, \tilde{\eta}) \rangle = \int_0^1 (y_{xx}\tilde{y}_{xx} + z\tilde{z})\, dx + \mu_1 \xi\tilde{\xi} + \mu_2 \eta\tilde{\eta}.$$

Next we define the unbounded operator A and the linear operators B, $\tilde{B}$:

$$(2.4)\quad D(A) = \left(\begin{array}{l} u = (y, z, \xi, \eta) \in H^4(0,1) \times H^2(0,1) \times \mathbb{R} \times \mathbb{R} \text{ such that} \\ y(0) = y_x(0) = z(0) = z_x(0) = 0 \quad \text{and} \quad \xi = z(1),\, \eta = z_x(1) \end{array} \right),$$

$$(2.5)\qquad Au = (-z, y_{xxxx}, -\frac{1}{\mu_1} y_{xxx}(1), \frac{1}{\mu_2} y_{xx}(1)) \quad \forall\, u = (y, z, \xi, \eta) \in D(A),$$

$$(2.6)\qquad Bu = (0, 0, b_{11} + b_{12}, b_{21} + b_{22}) \quad \forall\, u = (y, z, \xi, \eta) \in E,$$

$$(2.7)\qquad \tilde{B}u = (0, 0, b_{11} - b_{12}, b_{21} - b_{22}) \quad \forall\, u = (y, z, \xi, \eta) \in E,$$

where we have posed

$$(2.8)\qquad \begin{cases} b_{11} = \dfrac{1}{\mu_1}(\delta_{11} y(1) + \delta_{12} y_x(1)), & b_{12} = \dfrac{1}{\mu_1}(\nu_{11}\xi + \nu_{12}\eta), \\ b_{21} = \dfrac{1}{\mu_2}(\delta_{21} y(1) + \delta_{22} y_x(1)), & b_{22} = \dfrac{1}{\mu_2}(\nu_{21}\xi + \nu_{22}\eta). \end{cases}$$

With these notations we can formulate the system (2.3) into the following abstract form:

$$\frac{d}{dt}u(t) + (A+B)u(t) = 0, \qquad u(0) = u_0. \tag{2.9}$$

THEOREM 2.1. *For any real constants δ_{ij} and ν_{ij}, the equation* (2.9) *is not uniformly stable on the energy space E.*

Proof. It has been shown in Littman–Markus [9] that the operator A defined by (2.4)–(2.5) is maximal monotone and satisfies the property $A^* = -A$; hence it generates a group $S_0(t)$ of isometries on the energy space E.

Let $S(t)$ and $\tilde{S}(t)$ denote the groups generated, respectively, by $A+B$ and $A+\tilde{B}$. From the definitions (2.6)–(2.8) and Sobolev's embedding, it follows that the operators B and $\tilde{B}$ are continuous and of finite rank, therefore compact. Using a classical result of compact perturbation (cf. Russell [17, Prop. 1.1]), we deduce that there exist no positive constants $0 \le \gamma < 1$ and $t_0 > 0$ such that

$$\|S(t_0)\| \le \gamma \quad \text{and} \quad \|\tilde{S}(-t_0)\| \le \gamma. \tag{2.10}$$

Given $\mu_0 \in D(A)$, $u(t) = S(t)u_0 = (y(t), z(t), \xi(t), \eta(t))$ is the solution of the equation (2.9). We define

$$\tilde{y}(x,t) = y(x,-t), \quad \tilde{z}(x,t) = -z(x,-t), \quad \tilde{\xi}(t) = -\xi(-t),$$
$$\tilde{\eta}(t) = -\eta(-t), \quad \tilde{u}(t) = (\tilde{y}(t), \tilde{z}(t), \tilde{\xi}(t), \tilde{\eta}(t)).$$

Then we have $\|\tilde{u}(-t)\| = \|u(t)\|$ for all $t \in \mathbb{R}$.

On the other hand, a straightforward computation shows that $\tilde{u}(t) = \tilde{S}(t)\tilde{u}(0)$. It follows that

$$\|\tilde{S}(-t)\| = \|S(t)\| \quad \forall t \in \mathbb{R},$$

which together with (2.10) achieves the proof of Theorem 2.1. □

Remark 2.1. The proof of Theorem 2.1 is a direct application of the classical result of compact perturbation. In fact, there are several recent more general results that allow us to conclude that the semigroup $S(t)$ is not exponentially stable (cf. Curtain [4], Jacobson [5], and Rebarber [16]).

Remark 2.2. In the specific case of boundary feedbacks L_1, L_2 of the form

$$L_1(y, y_t) = -y_t(1,t), \qquad L_2(y, y_t) = -y_{xt}(1,t),$$

Theorem 2.1 was proved by Littman–Markus [9]. Their method is constructive and based on the asymptotic estimation of the eigenvalues of the operator $A+B$. Here our method is very simple and can be easily applied to other cases, in particular to some plate models (cf. Markus–You [11]) where the estimation of the eigenvalues seems to be impossible.

3. Uniform stabilization by passive boundary damping. In this section, we consider an elastic beam with usual boundary feedbacks applied to the end without the mass. By means of multipliers method, we establish the uniform decay of energy for that case. Let us consider the following hybrid system:

$$y_{tt} + y_{xxxx} = 0, \tag{3.1}$$

$$\alpha_1 y(0,t) + y_{xxx}(0,t) + \beta_{11} y_t(0,t) + 2\beta_{12} y_{xt}(0,t) = 0, \tag{3.2}$$

$$\alpha_2 y_x(0,t) - y_{xx}(0,t) + 2\beta_{21} y_t(0,t) + \beta_{22} y_{xt}(0,t) = 0, \tag{3.3}$$

$$\mu_1 y_{tt}(1,t) - y_{xxx}(1,t) = 0, \tag{3.4}$$

$$\mu_2 y_{xtt}(1,t) + y_{xx}(1,t) = 0, \tag{3.5}$$

where α_1, α_2 are positive constants.

Let y be a smooth solution of the system (3.1)–(3.5). We define the associated energy $E(t)$ by

(3.6)

$$E(t) = \frac{1}{2}\left\{ \int_0^1 (y_t^2 + y_{xx}^2)\,dx + \alpha_1 y^2(0,t) + \alpha_2 y_x^2(0,t) + \mu_1 y_t^2(1,t) + \mu_2 y_{xt}^2(1,t) \right\}.$$

A straightforward computation gives (at least formally)

$$\frac{d}{dt}E(t) = -(\beta_{11} y_t^2(0,t) + 2(\beta_{12} + \beta_{21}) y_t(0,t) y_{xt}(0,t) + \beta_{22} y_{xt}^2(0,t)). \tag{3.7}$$

Assume that the boundary damping coefficients β_{ij} satisfy the following conditions:

$$\beta_{11} > 0, \quad \beta_{22} > 0, \quad \beta_{11}\beta_{22} > (\beta_{12} + \beta_{21})^2. \tag{3.8}$$

Then there exists a constant $\beta_0 > 0$ such that

$$\frac{d}{dt}E(t) \geq -\beta_0 (y_t^2(0,t) + y_{xt}^2(0,t)). \tag{3.9}$$

Now let us introduce the energy space

$$E = \{(y, z, \xi, \eta) \in H^2(0,1) \times L^2(0,1) \times \mathbb{R} \times \mathbb{R}\}. \tag{3.10}$$

For all $u = (y, z, \xi, \eta) \in E$ and all $\tilde{u} = (\tilde{y}, \tilde{z}, \tilde{\xi}, \tilde{\eta}) \in E$, we define the inner product

$$\langle u, \tilde{u} \rangle = \int_0^1 (y_{xx}\tilde{y}_{xx} + z\tilde{z})\,dx + \alpha_1 y(0)\tilde{y}(0) + \alpha_2 y_x(0)\tilde{y}_x(0) + \mu_1 \xi\tilde{\xi} + \mu_2 \eta\tilde{\eta}. \tag{3.11}$$

Next we define the unbounded operator A as follows:

(3.12)

$$D(A) = \left(\begin{array}{c} (y, z, \xi, \eta) \in H^4(0,1) \times H^2(0,1) \times \mathbb{R}^2 \text{ such that } \xi = z(1), \quad \eta = z_x(1) \\ \alpha_1 y(0) + y_{xxx}(0) + \beta_{11} z(0) + 2\beta_{12} z_x(0) = 0 \\ \alpha_2 y_x(0) - y_{xx}(0) + 2\beta_{21} z(0) + \beta_{22} z_x(0) = 0 \end{array} \right),$$

$$Au = \left(-z, y_{xxxx}, -\frac{1}{\mu_1} y_{xxx}(1), \frac{1}{\mu_2} y_{xx}(1) \right) \quad \forall\, u = (y, z, \xi, \eta) \in D(A). \tag{3.13}$$

Then by setting

$$z(x,t) = y_t(x,t), \quad \xi(t) = y_t(1,t), \quad \eta(t) = y_{xt}(1,t), \quad u(t) = (y(t), z(t), \xi(t), \eta(t)),$$

we can write the hybrid system (3.1) – (3.5) into the following abstract form:

$$\frac{d}{dt}u(t) + Au(t) = 0, \qquad u(0) = u_0. \tag{3.14}$$

Under the conditions (3.8), it is easy to prove that A is a maximal monotone operator, hence it generates a semigroup of contractions $S(t)$ on the space E. Moreover by the classical Hille–Yosida theorem (cf. Brezis [1, p. 105], and Pazy [12, p. 14]), we have the following results.

PROPOSITION 3.1. *Assume that the conditions* (3.8) *hold.*

(i) *For any initial data* $u_0 \in D(A)$, *the equation* (3.14) *admits a unique strong solution* $u(t) = (y(t), z(t), \xi(t), \eta(t)) \in D(A)$ *such that*

$$y(t) \in C^0([0,+\infty[; H^4(0,1)) \cap C^1([0,+\infty[; H^2(0,1)) \cap C^2([0,+\infty[; L^2(0,1)), \tag{3.15}$$

$$y(1,t) \in C^2([0,+\infty[; \mathbb{R}), \qquad y_x(1,t) \in C^2([0,+\infty[; \mathbb{R}). \tag{3.16}$$

(ii) *For any initial data* $u_0 \in E$, *the equation* (3.14) *admits a unique weak solution* $u(t) = S(t)u_0 = (y(t), z(t), \xi(t), \eta(t)) \in E$ *such that*

$$y(t) \in C^0([0,+\infty[; H^2(0,1)) \cap C^1([0,+\infty[; L^2(0,1)). \tag{3.17}$$

$$y(1,t) \in C^1([0,+\infty[; \mathbb{R}), \qquad y_x(1,t) \in C^1([0,+\infty[; \mathbb{R}). \tag{3.18}$$

Notice that for a general function $y(x,t)$ possessing only the smoothness property (3.15), the trace functions $y_{tt}(1,t)$ and $y_{xtt}(1,t)$ don't make sense. Here they are defined by means of the equations

$$y_{tt}(1,t) = \frac{1}{\mu_1} y_{xxx}(1,t) \qquad y_{xtt}(1,t) = -\frac{1}{\mu_2} y_{xx}(1,t).$$

From (3.15), we see that the right-hand sides of these equations are continuous functions.

Because the domain $D(A)$ is dense in E and because $S(t)$ is a semigroup of contractions on the space E, we assume in this section systematically the smoothness properties (3.15) – (3.16). All the results of decay estimates established for strong solutions can be extended to the case of weak solutions by a standard argument of density and contraction property.

Now let y be a smooth solution of (3.1) – (3.5). We define the following functionals:

$$\rho_1(t) = 4\int_0^1 (x-1) y_t y_x - 5\int_0^1 y_t y \, dx, \tag{3.19}$$

$$\rho_2(t) = y_x(0,t)\left\{\int_0^1 y_t(C_0 x + 4) + \mu_1(C_0 + 4) y_t(1,t) + \mu_2 C_0 y_{xt}(1,t)\right\}, \tag{3.20}$$

$$\rho_3(t) = 5y(0,t)\left\{2\int_0^1 y_t \, dx + 2\mu_1 y_t(1,t) + \beta_{11} y(0,t) + 2\beta_{12} y_x(0,t)\right\}, \tag{3.21}$$

$$\rho_4(t) = -5\mu_1 y_t(1,t) y(1,t) - \mu_2 y_{xt}(1,t) y_x(1,t), \tag{3.22}$$

$$\rho(t) = \rho_1(t) + \rho_2(t) + \rho_3(t) + \rho_4(t). \tag{3.23}$$

The constant C_0 will be determined later (see (3.39)).

PROPOSITION 3.2. *Assume that the conditions* (3.8) *hold. Then there exist positive constants* C_1, C_2, *and* θ *such that the following estimates hold*

$$|\rho(t)| \leq C_1 E(t) \quad \forall t \geq 0, \tag{3.24}$$

$$\frac{d}{dt}\rho(t) \leq -\theta E(t) + C_2(y_t^2(0,t) + y_{xt}^2(0,t)) \quad \forall t \geq 0 \tag{3.25}$$

for all solutions y *of the system* (3.1) – (3.5).

Proof. First using the Cauchy–Schwartz inequality, a straightforward computation gives

$$\begin{aligned}
&|\rho_1(t)| \leq 5\int_0^1 (y_t^2 + y^2 + y_x^2)\,dx,\\
&|\rho_2(t)| \leq (C_0+4)(1+\mu_1+\mu_2)\left(y_x^2(0,t) + y_t^2(1,t) + y_{xt}^2(1,t) + \int_0^1 y_t^2\,dx\right),\\
&|\rho_3(t)| \leq 10(1+\mu_1+\beta_{11}+|\beta_{12}|)\left(y^2(0,t) + y_x^2(0,t) + y_t^2(1,t) + \int_0^1 y_t^2\,dx\right),\\
&|\rho_4(t)| \leq 3(\mu_1+\mu_2)(y_t^2(1,t) + y_{xt}^2(1,t) + y^2(1,t) + y_x^2(1,t)).
\end{aligned}$$

Now let $\gamma > 0$ denote the largest constant such that

$$y^2(1) + y_x^2(1) + \int_0^1 (y^2 + y_x^2)\,dx \leq \gamma\left(y^2(0) + y_x^2(0) + \int_0^1 y_{xx}^2\,dx\right)$$

for all functions $y \in H^2(0,1)$. Combining the above estimates gives the estimate (3.24).

The estimate (3.25) is more difficult to establish. For the sake of clarity, we divide the proof into five steps.

Step 1. *Calculation of the derivative of* $\rho_1(t)$. Using the equation (3.1), a simple calculation gives

$$\begin{aligned}
\frac{d}{dt}\rho_1(t) = &-\int_0^1 \{7y_t^2 + y_{xx}^2\}\,dx + 2y_t^2(0,t) + 2y_{xx}^2(0,t)\\
&- 4y_{xxx}(0,t)y_x(0,t) - y_{xx}(1,t)y_x(1,t) + y_{xx}(0,t)y_x(0,t)\\
&+ 5y_{xxx}(1,t)y(1,t) - 5y_{xxx}(0,t)y(0,t).
\end{aligned} \tag{3.26}$$

Step 2. *Estimate for the derivative of* $\rho_2(t)$. It follows from the equation (3.1) that

(3.27)

$$\begin{aligned}
\frac{d}{dt}\rho_2(t) = &\; y_{xt}(0,t)\left\{\int_0^1 (C_0x+4)y_t\,dx + \mu_1(C_0+4)y_t(1,t) + \mu_2 C_0 y_{xt}(1,t)\right\}\\
&+ y_x(0,t)\left\{-\int_0^1 (C_0x+4)y_{xxxx}\,dx + \mu_1(C_0+4)y_{tt}(1,t) + \mu_2 C_0 y_{xtt}(1,t)\right\}.
\end{aligned}$$

Using the boundary conditions (3.4) – (3.5), we calculate

$$
\begin{aligned}
&-\int_0^1 (C_0x+4)y_{xxxx}\,dx\\
&= 4y_{xxx}(0,t)-(C_0+4)y_{xxx}(1,t)+C_0y_{xx}(1,t)-C_0y_{xx}(0,t)\\
&= 4y_{xxx}(0,t)-(C_0+4)\mu_1y_{tt}(1,t)-C_0\mu_2y_{xtt}(1,t)-C_0y_{xx}(0,t).
\end{aligned}
\tag{3.28}
$$

Inserting (3.28) into (3.27) gives

$$
\begin{aligned}
\frac{d}{dt}\rho_2(t) &= 4y_x(0,t)y_{xxx}(0,t)-C_0y_x(0,t)y_{xx}(0,t)\\
&\quad + y_{xt}(0,t)\left\{\int_0^1 (C_0x+4)y_t\,dx+\mu_1(C_0+4)y_t(1,t)+\mu_2C_0y_{xt}(1,t)\right\}.
\end{aligned}
\tag{3.29}
$$

Using the Cauchy–Schwartz inequality, we get

$$
\begin{aligned}
&y_{xt}(0,t)\left\{\int_0^1 (C_0x+4)y_t\,dx+\mu_1(C_0+4)y_t(1,t)+\mu_2C_0y_{xt}(1,t)\right\}\\
&\le Cy_{xt}^2(0,t)+\frac{1}{4}\left\{\int_0^1 y_t^2\,dx+\mu_1y_t^2(1,t)+\mu_2y_{xt}^2(1,t)\right\}.
\end{aligned}
\tag{3.30}
$$

Inserting (3.30) into (3.29) gives

$$
\begin{aligned}
\frac{d}{dt}\rho_2(t) &\le 4y_{xxx}(0,t)y_x(0,t)-C_0y_{xx}(0,t)y_x(0,t)\\
&\quad + Cy_{xt}^2(0,t)+\frac{1}{4}\left\{\int_0^1 y_t^2\,dx+\mu_1y_t^2(1,t)+\mu_2y_{xt}^2(1,t)\right\}.
\end{aligned}
\tag{3.31}
$$

Step 3. *Estimate for the derivative of* $\rho_3(t)$. Using the equation (3.1), we have

$$
\begin{aligned}
\frac{d}{dt}\rho_3(t) &= 5y_t(0,t)\left\{2\int_0^1 y_t\,dx+2\mu_1y_t(1,t)+\beta_{11}y(0,t)+2\beta_{12}y_x(0,t)\right\}\\
&\quad + 5y(0,t)\left\{-2\int_0^1 y_{xxxx}\,dx+2\mu_1y_{tt}(1,t)+\beta_{11}y_t(0,t)+2\beta_{12}y_{xt}(0,t)\right\}.
\end{aligned}
\tag{3.32}
$$

By the boundary conditions (3.2) and (3.4), we obtain

$$
\begin{aligned}
-2\int_0^1 y_{xxxx}\,dx &= 2y_{xxx}(0,t)-2y_{xxx}(1,t)\\
&= y_{xxx}(0,t)-\alpha_1y(0,t)-\beta_{11}y_t(0,t)-\beta_{12}y_{xt}(0,t)-2\mu_1y_{tt}(1,t).
\end{aligned}
$$

Inserting the above relation into (3.32), we get

(3.33)

$$\frac{d}{dt}\rho_3(t) = 5y(0,t)y_{xxx}(0,t) - 5\alpha_1 y^2(0,t) + 5y_t(0,t)\left\{2\int_0^1 y_t\,dx + 2\mu_1 y_t(1,t) + \beta_{11}y(0,t) + 2\beta_{12}y_x(0,t)\right\}.$$

Using the Cauchy–Schwartz inequality, we can find a positive constant C such that

$$5y_t(0,t)\left\{2\int_0^1 y_t\,dx + 2\mu_1 y_t(1,t) + \beta_{11}y(0,t) + 2\beta_{12}y_x(0,t)\right\} \le \frac{9\alpha_1}{2}y^2(0,t) + \frac{3\alpha_2^2}{2}y_x^2(0,t) + Cy_t^2(0,t) + \frac{1}{4}\left(\int_0^1 y_t^2\,dx + \mu_1 y_t^2(1,t)\right).$$

Inserting the above estimate into (3.33) gives

$$\frac{d}{dt}\rho_3(t) \le \frac{1}{4}\left(\int_0^1 y_t^2\,dx + \mu_1 y_t^2(1,t)\right) + 5(y_{xxx}y)(0,t) - \frac{\alpha_1}{2}y^2(0,t) + \frac{3\alpha_2^2}{2}y_x^2(0,t) + Cy_t^2(0,t). \tag{3.34}$$

Step 4. *Calculation of the derivative of* $\rho_4(t)$. Using the boundary conditions (3.4) – (3.5) we calculate

$$\begin{aligned}\frac{d}{dt}\rho_4(t) &= -5\mu_1 y_t^2(1,t) - 5\mu_1 y_{tt}(1,t)y(1,t) - \mu_2 y_{xt}^2(1,t) - \mu_2 y_{xtt}(1,t)y_x(1,t)\\ &= -5\mu_1 y_t^2(1,t) - 5y_{xxx}(1,t)y(1,t) - \mu_2 y_{xt}^2(1,t) + y_{xx}(1,t)y_x(1,t).\end{aligned} \tag{3.35}$$

Step 5. *Estimate for the derivative of* $\rho(t)$. Combining (3.26), (3.31), (3.34), and (3.35), we get

$$\begin{aligned}\frac{d}{dt}\rho(t) \le{}& -\frac{1}{2}\left\{\int_0^1 (y_t^2 + y_{xx}^2)\,dx + \alpha_1 y^2(0,t) + \mu_1 y_t^2(1,t) + \mu_2 y_{xt}^2(1,t)\right\}\\ &+ C(y_t^2(0,t) + y_{xt}^2(0,t)) + \frac{3\alpha_2^2}{2}y_x^2(0,t) + 2y_{xx}^2(0,t)\\ &- (C_0 - 1)y_{xx}(0,t)y_x(0,t).\end{aligned} \tag{3.36}$$

On the other hand, by the boundary condition (3.3), we have

$$\alpha_2^2 y_x^2(0,t) - 2\alpha_2(y_x y_{xx})(0,t) + y_{xx}^2(0,t) \le 5\beta_{21}^2 y_t^2(0,t) + 5\beta_{22}^2 y_{xt}^2(0,t). \tag{3.37}$$

Plugging (3.37) into (3.36) gives

(3.38)

$$\frac{d}{dt}\rho(t) \le C_2(y_t^2(0,t) + y_{xt}^2(0,t)) - (C_0 - 4\alpha_2 - 1)y_{xx}(0,t)y_x(0,t) -\frac{1}{2}\left\{\int_0^1 (y_t^2 + y_{xx}^2)\,dx + \alpha_1 y^2(0,t) + \alpha_2^2 y_x^2(0,t) + \mu_1 y_t^2(1,t) + \mu_2 y_{xt}^2(1,t)\right\}.$$

Now, by choosing

$$C_0 = 4\alpha_2 + 1, \qquad \theta = \min\{1,\, \alpha_2\}, \tag{3.39}$$

in (3.38), we obtain

$$\frac{d}{dt}\rho(t) \leq -\theta E(t) + C_2(y_t^2(0,t) + y_{xt}^2(0,t)).$$

The proof of Proposition 3.1 is thus complete. □

Now we proceed as in Komornik–Zuazua [6]. Starting from the estimates (3.24) – (3.25), we can establish the following uniform decay of energy.

THEOREM 3.3. *Assume that the conditions* (3.8) *hold. Then given any* $M > 1$, *there exists a positive constant* ω *such that*

$$E(t) \leq ME(0)\exp(-\omega t) \quad \forall t \geq 0, \tag{3.40}$$

for all solutions y *of the system* (3.1) – (3.5).

Proof. First for $\varepsilon > 0$, we introduce the perturbed energy

$$E_\varepsilon(t) = E(t) + \varepsilon\rho(t).$$

Then given any $M > 1$, using the estimate (3.24), we prove the following inequalities:

$$M^{-1/2}E_\varepsilon(t) \leq E(t) \leq M^{1/2}E_\varepsilon(t) \quad \forall t \geq 0, \tag{3.41}$$

provided $\varepsilon > 0$ is small enough.

Next using (3.9) and (3.25), we have

$$\frac{d}{dt}E_\varepsilon(t) = \frac{d}{dt}E(t) + \varepsilon\frac{d}{dt}\rho(t) \leq -\theta\varepsilon E(t) + (\varepsilon C_2 - \beta_0)(y_{xt}^2(0,t) + y_{xxt}^2(0,t)).$$

Then for $\varepsilon > 0$ small enough, we deduce that

$$\frac{d}{dt}E_\varepsilon(t) \leq -\theta\varepsilon E(t) \leq -\theta\varepsilon M^{-1/2}E_\varepsilon(t).$$

Solving the above differential inequality, we obtain

$$E_\varepsilon(t) \leq E_\varepsilon(0)\exp(-\varepsilon M^{-1/2}t) = E_\varepsilon(0)\exp(-\omega t) \quad \forall t \geq 0,$$

which together with (3.41) implies that

$$E(t) \leq ME(0)\exp(-\omega t) \quad \forall t \geq 0.$$

The proof of Theorem 3.3 is complete. □

Remark 3.1. The conditions (3.8) are by no means optimal in the sense that only one moment control $y_{xt}(0,t)$ or only one force control $y_t(0,t)$ suffices for obtaining the uniform decay of energy. In fact, instead of the boundary feedbacks (3.2) – (3.3), we can consider the following ones:

$$y(0,t) = 0, \qquad y_x(0,t) - y_{xx}(0,t) + \beta y_{xt}(0,t) = 0, \tag{3.42}$$

$$y(0,t) + y_{xxx}(0,t) + \beta y_t(0,t) = 0, \qquad y_x(0,t) = 0. \tag{3.43}$$

For these two cases, the uniform exponential decay of (3.40) remains true for $\beta > 0$ (cf. Rao [15]).

Notice that the boundary feedbacks (3.2) – (3.3), (3.42), and (3.43) can be realized by means of passive mechanical systems of springs-dampers similar to those used in Chen et al. [3].

4. Uniform stabilization by higher-order feedbacks. In this section, we carry out a study of uniform stabilization for a clamped beam. In view of the negative result of Theorem 2.1, we have to choose suitable boundary feedbacks L_1, L_2 and energy space such that not only the boundary feedbacks L_1, L_2 are noncompact but also the problem is well posed. Here, as an example, we propose the following boundary feedbacks:

$$L_1(y, y_t) = y_{xxxt}(1,t), \qquad L_2(y, y_t) = -y_{xxt}(1,t).$$

To make our computations as clear as possible, we assume, without loss of generality, the constants μ_1 and μ_2 are equal to one (because $\mu_1 > 0$, $\mu_2 > 0$), and then we obtain the following hybrid system:

$$(4.1) \qquad \begin{cases} y_{tt} + y_{xxxx} = 0, & t > 0, \quad 0 < x < 1, \\ y(0,t) = y_x(0,t) = 0, & t > 0, \\ y_{tt}(1,t) - y_{xxx}(1,t) = y_{xxxt}(1,t), & t > 0, \\ y_{xtt}(1,t) + y_{xx}(1,t) = -y_{xxt}(1,t), & t > 0. \end{cases}$$

Because of the presence of the higher-order feedbacks such as $y_{xxxt}(1,t)$ and $-y_{xxt}(1,t)$ in the system (4.1), for the well-posedness of the problem, we choose the following more smooth energy space:

$$(4.2)\ E = \{(y,z) \in H^4(0,1) \times H^2(0,1) \text{ such that } y(0) = y_x(0) = z(0) = z_x(0) = 0\}.$$

For any $u = (y,z) \in E$ and any $\tilde{u} = (\tilde{y}, \tilde{z}) \in E$, we define the inner product by

$$(4.3) \qquad \langle u, \tilde{u} \rangle = \int_0^1 \{y_{xxxx}\tilde{y}_{xxxx} + z_{xx}\tilde{z}_{xx}\}\, dx + y_{xx}(1)\tilde{y}_{xx}(1) + y_{xxx}(1)\tilde{y}_{xxx}(1).$$

Next we define the unbounded operator A:

$$(4.4) \quad D(A) = \begin{pmatrix} u = (y,z) \in H^6(0,1) \times H^4(0,1) \text{ such that} \\ y(0) = y_x(0) = z(0) = z_x(0) = 0,\ y_{xxxx}(0) = y_{xxxxx}(0) = 0 \\ y_{xxxxx}(1) + y_{xxx}(1) = -z_{xxx}(1),\ y_{xxxxxx}(1) - y_{xx}(1) = z_{xx}(1) \end{pmatrix},$$

$$(4.5) \qquad Au = (-z,\ y_{xxxx}) \quad \forall\, u = (y,z) \in D(A).$$

PROPOSITION 4.1. *The operator A defined by* (4.4)–(4.5) *is maximal monotone on the Hilbert space E defined by* (4.2)–(4.3).

Proof. First for any $u = (y,z) \in D(A)$, by the definitions (4.3) and (4.5) and integration by parts, we have

$$(4.6) \qquad \begin{aligned} \langle Au, u \rangle &= z_{xx}(1)(y_{xxxxxx}(1) - y_{xx}(1)) - z_{xxx}(1)(y_{xxxxx}(1) + y_{xxx}(1)) \\ &= z_{xx}^2(1) + z_{xxx}^2(1) \geq 0. \end{aligned}$$

The last equality follows from the boundary conditions at $x = 1$ for $u \in D(A)$.

We next verify the range condition

$$(4.7) \qquad R(I + A) = E.$$

Let $u_0 = (y_0, z_0) \in E$, we have to solve the equation

$$u \in D(A) \quad \text{such that } u + Au = u_0,$$

which means that

$$(4.8)\qquad \begin{cases} y - z = y_0, \quad z - y_{xxxx} = z_0, \\ y(0) = y_x(0) = z(0) = z_x(0) = 0, \\ y_{xxxx}(0) = y_{xxxxx}(0) = 0, \\ y_{xxxx}(1) + y_{xxx}(1) = -z_{xxx}(1), \\ y_{xxxxx}(1) - y_{xx}(1) = z_{xx}(1). \end{cases}$$

Eliminating the unknown z in (4.8), we obtain

$$(4.9)\qquad \begin{cases} y + y_{xxxx} = y_0 + z_0 \in H^2(0,1), \\ y(0) = y_x(0) = 0, \\ 2y_{xxx}(1) + y_{xxxx}(1) = y_{0xxx}(1), \\ 2y_{xx}(1) - y_{xxxxx}(1) = y_{0xx}(1). \end{cases}$$

Next, we replace $y_{xxxx}(1)$ and $y_{xxxxx}(1)$ in (4.9) by their expressions

$$y_{xxxx}(1) = y_0(1) + z_0(1) - y(1), \qquad y_{xxxxx}(1) = y_{0x}(1) + z_{0x}(1) - y_x(1)$$

so that we get

$$(4.10)\qquad \begin{cases} y + y_{xxxx} = y_0 + z_0 \in H^2(0,1), \\ y(0) = y_x(0) = 0, \\ y(1) - 2y_{xxx}(1) = y_0(1) - y_{0xxx}(1) + z_0(1), \\ y_x(1) + 2y_{xx}(1) = y_{0x}(1) + y_{0xx}(1) + z_{0x}(1). \end{cases}$$

Now let us denote by V the subspace $V = \{y \in H^2(\omega) : y(0) = y_x(0) = 0\}$. We introduce the bilinear continuous operator $a(\cdot,\cdot)$ in $V \times V$

$$a(y,\varphi) = \int_0^1 (y\varphi + y_{xx}\varphi_{xx})\,dx + \frac{1}{2}y(1)\varphi(1) + \frac{1}{2}y_x(1)\varphi_x(1) \quad \forall y, \varphi \in V.$$

For any $\varphi \in V$, we define the linear continuous operator f by the following way

$$\begin{aligned} f(\varphi) &= \int_0^1 (y_0 + z_0)\varphi\,dx + \frac{1}{2}(y_0(1) - y_{0xxx}(1) + z_0(1))\varphi(1) \\ &\quad + \frac{1}{2}(y_{0x}(1) + y_{0xx}(1) + z_{0x}(1))\varphi_x(1). \end{aligned}$$

It is clear that $a(\cdot,\cdot)$ is V-elliptic in $V \times V$. We therefore deduce from the Lax–Milgram theorem (cf. Lions–Magenes [8, p. 216]) that there exists a unique function $y \in V$ such that

$$(4.11)\qquad a(y,\varphi) = f(\varphi) \quad \forall \varphi \in V,$$

which implies in particular that

$$y + y_{xxxx} = y_0 + z_0 \quad \text{in } \mathcal{D}'.$$

Because $y_0 + z_0 \in V$, we obtain that

$$y_{xxxx} = y_0 + z_0 - y \in V,$$

which means precisely that $y \in H^6(0,1)$. Now, given this regularity, when interpreting the equation (4.11) we find that y is the solution of the equation (4.10). Moreover, we have

$$\begin{cases} y_{xxxx}(0) = y_0(0) + z_0(0) - y(0) = 0, \\ y_{xxxxx}(0) = y_{0x}(0) + z_{0x}(0) - y_x(0) = 0. \end{cases}$$

This proves the range condition (4.7). Proposition 4.1 is thus proved. □

Now letting $z = y_t$, we can write the hybrid system (4.1) in the following abstract form:

$$\frac{d}{dt}u(t) + Au(t) = 0, \qquad u(0) = u_0. \tag{4.12}$$

We have the following well-posed theorem.

THEOREM 4.2. (i) *For any initial data* $u_0 = (y_0, z_0) \in D(A)$, *the equation* (4.12) *admits a unique strong solution* u *such that*

$$u(t) = (y(t), z(t)) \in D(A) \quad \forall t \geq 0, \tag{4.13}$$

$$y(t) \in C^2([0,+\infty[; H^2(0,1)) \cap C^1([0,+\infty[; H^4(0,1)) \cap C^0([0,+\infty[; H^6(0,1)). \tag{4.14}$$

(ii) *For any initial data* $u_0 = (y_0, z_0) \in E$, *the equation* (4.12) *admits a unique weak solution* $u(t) = (y(t), z(t))$ *such that*

$$u(t) = S(t)u_0 = (y(t), z(t)) \in E \quad \forall t \geq 0, \tag{4.15}$$

$$y(t) \in C^1([0,+\infty[; H^2(0,1)) \cap C^0([0,+\infty[; H^4(0,1)). \tag{4.16}$$

where $S(t)$ *denotes the strongly continuous semigroup of contractions on* E *generated by the maximal monotone operator* A.

Proof. Because the operator A is maximal monotone on the Hilbert space E, by the classical linear semigroup theory (cf. Brezis [1, p. 105] and Pazy [12, p. 14]), we know that for any $u_0 \in D(A)$ the equation (4.15) has a unique strong solution u such that

$$u(t) = (y(t), z(t)) \in C^1([0,+\infty[; E) \cap C^0([0,+\infty[; D(A)).$$

This means that

$$y(t) \in C^1([0,+\infty[; H^4(0,1)) \cap C^0([0,+\infty[; H^6(0,1)),$$

$$y_t(t) = z(t) \in C^1([0,+\infty[; H^2(0,1)) \cap C^0([0,+\infty[; H^4(0,1)).$$

The last two conditions, together with Sobolev's embeddings imply that

$$y(t) \in C^2([0,+\infty[; H^2(0,1)).$$

This proves the condition (4.14).

Now let $u_0 \in E$. Then by the denseness of $D(A)$ in E and the contraction of the semigroup $S(t)$, we know that the weak solution u given by $u(t) = S(t)u_0$ satisfies

$$u(t) = (y(t), z(t)) \in C^0([0,+\infty[; E),$$

which means that

$$y(t) \in C^0([0,+\infty[; H^4(0,1)), \qquad y_t(t) = z(t) \in C^0([0,+\infty[; H^2(0,1)).$$

Using Sobolev's embeddings, the above conditions imply that

$$y(t) \in C^1([0,+\infty[; H^2(0,1)).$$

The proof of Theorem 4.2 is thus complete. □

Now let $u(t) = (y(t), z(t))$ be a smooth solution of the system (4.12). We introduce the associated energy

$$E(t) = \frac{1}{2}\|u(t)\|^2 = \frac{1}{2}\left\{\int_0^1 (y_{xxt}^2 + y_{xxxx}^2)\,dx + y_{xx}^2(1,t) + y_{xxx}^2(1,t)\right\}. \tag{4.17}$$

Using (4.6) and (4.12), we calculate the derivative of the energy

$$\frac{d}{dt}E(t) = -\langle Au(t), u(t)\rangle = -(z_{xx}^2(1,t) + z_{xxx}^2(1,t)). \tag{4.18}$$

Using the relation $z = y_t$, it follows from (4.18) that

$$\frac{d}{dt}E(t) = -(y_{xxt}^2(1,t) + y_{xxxt}^2(1,t)) \leq 0. \tag{4.19}$$

Remark 4.1. The expression (4.19) shows that the energy $E(t)$ is nonincreasing and therefore defines a Lyapunov function. This justifies the introduction of the inner product (4.3).

Now let y be a solution of (4.1). We define the following functionals:

$$\begin{cases} \rho_1(t) = 2\int_0^1 x y_{tt} y_{tx}\,dx, \\ \rho_2(t) = 2(y_{xt}(1,t)y_{xx}(1,t) - y_t(1,t)y_{xxx}(1,t)), \\ \rho(t) = \rho_1(t) + \rho_2(t). \end{cases} \tag{4.20}$$

PROPOSITION 4.3. *There exist positive constants* C_1, C_2 *such that*

$$|\rho(t)| \leq C_1 E(t) \quad \forall t \geq 0, \tag{4.21}$$

$$\frac{d}{dt}\rho(t) \leq -E(t) + C_2(y_{xxt}^2(1,t) + y_{xxxt}^2(1,t)) \quad \forall t \geq 0 \tag{4.22}$$

for all solutions y *of the system* (4.1).

Proof. The estimate (4.21) is a direct application of the Cauchy–Schwartz inequality. In fact, we have

$$|\rho(t)| \leq 2(1+\lambda)E(t),$$

where $\lambda > 0$ is the largest constant such that

$$y^2(1) + y_x^2(1) + \int_0^1 y_x^2\,dx \leq \lambda\int_0^1 y_{xx}^2\,dx$$

for all functions $y \in H^2(0,1) : y(0) = y_x(0) = 0$.

The proof of the estimate (4.22) is more complicated. First by a straightforward computation, we have

$$\begin{aligned} \frac{d}{dt}\rho_1(t) = {} & -3\int_0^1 y_{xx}^2\,dx - \int_0^1 y_{xxxx}^2\,dx + y_{xxt}^2(1,t) + y_{tt}^2(1,t) \\ & - 2y_{xt}(1,t)y_{xxxt}(1,t) + 2y_{xt}(1,t)y_{xxt}(1,t), \end{aligned} \tag{4.23}$$

$$\begin{aligned}\frac{d}{dt}\rho_2(t) &= 2y_{xtt}(1,t)y_{xx}(1,t) + 2y_{xt}(1,t)y_{xxt}(1,t)\\ &\quad - 2y_{tt}(1,t)y_{xxx}(1,t) - 2y_t(1,t)y_{xxxt}(1,t).\end{aligned} \tag{4.24}$$

Combining (4.23), (4.24), and the boundary conditions at the end $x = 1$, we obtain

$$\begin{aligned}\frac{d}{dt}\rho(t) &\leq -2E(t) + (y_{xtt}(1,t) + y_{xx}(1,t))^2 + (y_{tt}(1,t) - y_{xxx}(1,t))^2\\ &\quad + y_{xxt}^2(1,t) - 2y_t(1,t)y_{xxxt}(1,t) + 2y_{xt}(1,t)(2y_{xxt}(1,t) - y_{xxxt}(1,t))\\ &= -2E(t) + 2y_{xxt}^2(1,t) + y_{xxxt}^2(1,t)\\ &\quad - 2y_t(1,t)y_{xxxt}(1,t) + 2y_{xt}(1,t)(2y_{xxt}(1,t) - y_{xxxt}(1,t)).\end{aligned} \tag{4.25}$$

Let $\sigma > 0$ denote the largest constant such that

$$y^2(1) + y_x^2(1) \leq \frac{\sigma}{2}\int_0^1 y_{xx}^2\,dx,$$

for all functions $y \in H^2(0,1) : y(0) = y_x(0) = 0$. It follows that

$$y_t^2(1,t) + y_{xt}^2(1,t) \leq \frac{\sigma}{2}\int_0^1 y_{txx}^2\,dx \leq \sigma E(t) \quad \forall t \geq 0.$$

Applying Young's inequality, it follows that

$$\begin{aligned}&-2y_t(1,t)y_{xxxt}(1,t) + 2y_{xt}(1,t)(2y_{xxt}(1,t) - y_{xxxt}(1,t))\\ &\quad \leq E(t) + 6\sigma(y_{xxt}^2(1,t) + y_{xxxt}^2(1,t)).\end{aligned} \tag{4.26}$$

Inserting (4.26) into (4.25) gives

$$\frac{d}{dt}\rho(t) \leq -E(t) + 2(1 + 3\sigma)(y_{xxt}^2(1,t) + y_{xxxt}^2(1,t)).$$

The proof of Proposition 4.3 is thus complete. □

Remark 4.2. We see that the first functional $\rho_1(t)$ is defined by means of the well-known classical multiplier xy_{tx} (cf. Chen et al. [2]) and the second one $\rho_2(t)$ designates the two *boundary multipliers*: $y_{xx}(1,t)$ and $y_{xxx}(1,t)$. The idea of the proof of the estimate (4.22) is to construct a suitable functional $\rho_2(t)$ such that all the boundary terms can be absorbed by the dissipative boundary damping.

THEOREM 4.4. *Given any* $M > 1$, *there exists a positive constant* ω *such that*

$$E(t) \leq ME(0)\exp(-\omega t) \quad \forall t \geq 0, \tag{4.27}$$

for all solutions y *of the hybrid system* (4.1).

The proof is a slight modification of that one of Theorem 3.3. Therefore, we omit the details.

Acknowledgment. The author would like to thank Professor F. Conrad for helpful discussions and the anonymous referees for the corrections and suggestions.

REFERENCES

[1] H. BREZIS, *Analyse Fonctionelle, Théorie et Applications*, Masson, Paris (1992).

[2] G. CHEN, M. C. DELFOUR, A. M. KRALL, AND G. PAYRE, *Modeling, stabilization and control of serially connected beams*, SIAM J. Control Optim., 25 (1987), pp. 526–546.

[3] G. Chen, S. G. Krantz, D. W. Ma, C. E. Wayne, and H. H. West, *The Euler–Bernoulli beam equation with boundary energy dissipation*, in Operator Methods for Optimal Control Problems, S. J. Lee, ed., Lecture Notes in Pure and Appl. Math. no. 108, Marcel Dekker, New York, 1987, pp. 67–96.

[4] R. Curtain, *Equivalence of input-output stability and exponential stability for infinite dimensional system*, Math. Systems Theory, 21 (1988), pp. 19–48.

[5] C. A. Jacobson, *Linear state-space systems in infinite-dimensional space: The role and characterization of joint stabilizability/detectability*, IEEE Trans. Automat. Control, 33 (1988), pp. 541–549.

[6] V. Komornik and E. Zuazua, *A direct method for the boundary stabilization of the wave equation*. J. Math. Pures Appl., 69 (1990), pp. 33–54.

[7] E. B. Lee and Y. C. You, *Stabilization of a hybrid (string/point mass) system*, in Proc. Fifth Int. Conf. on System Engineering, Dayton, OH, 1987.

[8] J. L. Lions and E. Magenes, *Problèmes aux limites non homogènes*, Vol. I, Dunod, Paris, 1968.

[9] W. Littman and L. Markus, *Stabilization of a hybrid system of elasticity by feedback boundary damping*, Ann. Mat. Pura Appl., 152 (1988), pp. 281–330.

[10] W. Littman and L. Markus, *Exact boundary controllability of a hybrid system of elasticity*, Arch. Rational Mech. Anal., 103 (1988), pp. 193–236.

[11] L. Markus and Y. C. You, *Dynamical boundary control for elastical plates of general shape*, SIAM J. Control Optim., 31 (1993), pp. 983–992.

[12] A. Pazy, *Semigroups of linear operators and applications to partial differential equations*, Springer-Verlag, New York, 1983.

[13] B. P. Rao, *Decay estimates of solutions for a hybrid system of flexible structures*, European Journal of Applied Mathematics, 4 (1993), pp. 303–319.

[14] B. P. Rao, *Stabilisation uniforme d'un système hybride en élasticité*, C. R. Acad. Sci. Paris, 316, Série I, 1993, pp. 261–266.

[15] B. P. Rao, *Recent results in non-uniform and uniform stabilization of the SCOLE model by boundary feedbacks*, in Lecture Notes in Pure and Applied Mathematics 163, J.-P. Zolésio, ed., Marcel Dekker, New York, 1994, pp. 357–365.

[16] R. Rebarber, *Necessary conditions for exponential stabilizability of distributed parameter systems with infinite dimensional unbounded feedback*, Systems Control Lett., 14 (1990), pp. 241–248.

[17] D. L. Russell, *Decay rates for weakly damped systems in Hilbert space obtained with control-theoretic methods*, J. Differential Equations, 19 (1975), pp. 344–370.

[18] M. Slemrod, *Feedback stabilization of a linear system in Hilbert space with an a priori bounded control*, Math. Control Signals Systems, 2 (1989), pp. 265–285.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 33, No. 2, pp. 455–468, March 1995

© 1995 Society for Industrial and Applied Mathematics
007

SYSTEM EQUIVALENCE FOR PERIODIC MODELS AND SYSTEMS*

OSVALDO M. GRASSELLI[†], SAURO LONGHI[‡], AND ANTONIO TORNAMBÈ[†]

Abstract. In this paper, the problem of obtaining a periodic description in state-space form of a linear process which can be modeled by linear difference equations with periodic coefficients is considered. On the basis of a polynomial time-invariant description of a linear periodic process, system equivalence between two such processes is introduced and studied. For a given periodic causal process, under an additional assumption, a periodic state-space description is found which is system equivalent to it. It is shown that the order, the characteristic multipliers, and the stacked transfer matrix at any initial time of the periodic system thus obtained coincide with those of the original periodic process, and that the asymptotic stability, the reachability, the observability, the controllability, the reconstructibility, the stabilizability, the detectability, and even the Jordan form of the monodromy matrix of such a system are determined by the original periodic model, as well as the existence of a solution of the robust tracking and regulation problem.

Key words. periodic systems, discrete-time systems, periodic models, system equivalence, realization theory

AMS subject classifications. 93A10, 93A99, 93B15

1. Introduction.

For processes which can be modeled by linear difference (or differential) equations with constant coefficients, Rosenbrock [28] introduced the polynomial matrix description in the form of the following pair of vector equations:

$$T(\delta)\xi = U(\delta)u, \tag{1}$$

$$y = V(\delta)\xi + W(\delta)u, \tag{2}$$

where $T(\delta), U(\delta), V(\delta)$, and $W(\delta)$ are polynomial matrices in δ, which for difference equations has the meaning of the one-step forward-shift operator. He showed that under the polynomial transformations on (1), (2) that he called strict system equivalence, if $T(\delta)$ is square, $\det T(\delta) \neq 0$ (and has a degree equal to the dimension of $T(\delta)$), and the input-output transfer matrix corresponding to (1), (2) is proper, then it is possible to obtain a description of the same process in state-space form, i.e., in the case of difference equations, of the type

$$x(k+1) = A\,x(k) + B\,u(k), \tag{3}$$

$$y(k) = C\,x(k) + D\,u(k). \tag{4}$$

Since then, several authors have studied the polynomial matrix description (1), (2) and the procedures for the computation of a state-space realization (3), (4) strictly system equivalent to (1), (2) (see, e.g., [4], [5], [8], [18], [19], [24], [26], [30]).

The same kind of problem seems to be of real interest for processes which can be modeled by linear difference equations with periodic coefficients (whose period will be denoted by ω) of the following form:

$$\sum_{i=0}^{r} T_i(k)\xi(k+i) = \sum_{i=0}^{r} U_i(k)u(k+i), \tag{5}$$

* Received by the editors July 20, 1992; accepted for publication (in revised form) November 8, 1993. This work was supported by Ministero dell'Università e della Ricerca Scientifica e Tecnologica and Ministero della Pubblica Istruzione.

[†] Dipartimento di Ingegneria Elettronica, Università di Roma "Tor Vergata," via della Ricerca Scientifica, 00133 Roma, Italy.

[‡] Dipartimento di Elettronica e Automatica, Università di Ancona, via Brecce Bianche, 60131 Ancona, Italy.

$$y(k) = \sum_{i=0}^{r} V_i(k)\xi(k+i) + \sum_{i=0}^{r} W_i(k)u(k+i) \tag{6}$$

for some integer $r \geq 0$, where $k \in \mathbf{Z}, \xi(k+i) \in \mathbf{R}^m$ is the vector of the internal variables or pseudostate, $u(k+i) \in \mathbf{R}^p$ is the input, $y(k) \in \mathbf{R}^q$ is the output, $T_i(k)$, $U_i(k)$, $V_i(k)$, and $W_i(k)$ $(i = 0, \ldots, r)$ are real periodic matrices of period ω (briefly ω-periodic), and the $T_i(k)$, $i = 0, \ldots, r$, are possibly square. Equations (5), (6) are termed the *model* of the process under consideration.

The interest in obtaining a description of such a process in state-space form is motivated by the large variety of processes which can be modeled by linear equations with periodic coefficients and the resulting attention devoted to linear periodic systems [1], [2], [23], [27], specifically to discrete-time ones (see, e.g., [3], [9], [13], [14], [29]), for which a control theory (based on a state-space description) is developing, including eigenvalue assignment, state and output dead-beat control, disturbance localization, model matching, robust tracking and regulation, and block decoupling [6], [10], [11], [12], [15], [16], [21], [22], [25].

In this paper, the problem of obtaining a state-space description of the periodic process (5), (6) is faced. In §2 a polynomial time-invariant characterization of such a process and some related notions are introduced, including the order of such a model (i.e., the number of arbitrary and independent initial conditions needed for uniquely solving (5) for a given function $u(\cdot)$), as well as some conditions which are to be satisfied so that the process is causal. In §3, system equivalence between two models of the form (5), (6) is introduced, and the properties which are invariant under it are analysed. In §4, a periodic system which is system equivalent to a causal periodic model (5), (6) is found under an additional assumption on the model.

Henceforth, the identity matrix of dimension ν will be denoted either by I_ν, or simply by I; Δ will denote the ω-steps forward-shift operator, Δ^{-1} will denote its inverse; the following notation will be used also. Let $R_\nu(\Delta)$, $\nu \in \mathbf{Z}^+$, be defined by

$$R_\nu(\Delta) := \begin{bmatrix} 0 & I_{(\omega-1)\nu} \\ \Delta I_\nu & 0 \end{bmatrix}. \tag{7}$$

Let a vector function $z(t) \in \mathbf{R}^\nu$ be given with $t \in \mathbf{Z}$; for any $k \in \mathbf{Z}$, the *ω-stacked form of $z(t)$ at (the initial) time k* is defined by

$$z_k(h) := \left[z^T(k+h\omega)\ z^T(k+h\omega+1)\ \ldots\ z^T(k+h\omega+\omega-1)\right]^T, \quad h \in \mathbf{Z}.$$

The vector $z_k(h)$ can be considered a function either of k or of h; in the following, whenever the operator $R_\nu(\Delta)$ will be applied to $z_k(h)$, the operator Δ will have the meaning of an ω-steps forward-shift in the k variable, or, equivalently, a one-step forward-shift in the h variable. Let an ω-periodic matrix $F(t) \in \mathbf{R}^{\nu\times\mu}$ be given, with $t \in \mathbf{Z}$, representing the linear map $z(t) = F(t)w(t)$. For any $k \in \mathbf{Z}$, the *ω-stacked form of $F(t)$ at (the initial) time k* is defined by $\mathcal{F}_k := \operatorname{diag}\{F(k), F(k+1), \ldots, F(k+\omega-1)\}$, and represents the induced linear map between the ω-stacked forms at time k of the vector functions $z(t)$ and $w(t)$, i.e., $z_k(h) = \mathcal{F}_k w_k(h), h \in \mathbf{Z}$.

The following lemma can be easily proved, taking into account that $\Delta F(k)\Delta^{-1} = F(k+\omega)$ and $\Delta^{-1}F(k+\omega-1)\Delta = F(k-1)$.

LEMMA 1.1. *For any vector function $z(t) \in \mathbf{R}^\nu$ and ω-periodic matrix $F(t) \in \mathbf{R}^{\nu\times\mu}$ $(t \in \mathbf{Z})$, and for any $k \in \mathbf{Z}$, the following identities hold:*

$$R_\nu^{j\omega+i}(\Delta)z_k(h) = z_{k+j\omega+i}(h) = z_{k+i}(h+j) \quad \forall i,j \in \mathbf{Z}; \tag{8}$$

$$R_\nu^{j\omega+i}(\Delta)\mathcal{F}_k R_\mu^{-(j\omega+i)}(\Delta) = \mathcal{F}_{k+j\omega+i} = \mathcal{F}_{k+i} \quad \forall i,j \in \mathbf{Z}. \tag{9}$$

Identity (9) *still holds with* Δ *replaced by a scalar complex variable.*

2. A time-invariant characterization of ω-periodic models and systems. By introducing the ω-stacked forms $\xi_{k_0}(h)$, $u_{k_0}(h)$, and $y_{k_0}(h)$ at time k_0 of vectors $\xi(k), u(k), y(k)$, and the ω-stacked forms $\mathcal{T}_{i,k_0}, \mathcal{U}_{i,k_0}, \mathcal{V}_{i,k_0}$, and $\mathcal{W}_{i,k_0}$ at time k_0 of matrices $T_i(k), U_i(k)$, $V_i(k), W_i(k), i = 0, \ldots, r$, and taking (8) into account, model (5), (6) can be expressed in the following form, which is called the *ω-stacked form at time k_0 of model* (5), (6) (briefly, the *ω-stacked model at time k_0*):

$$\mathcal{T}_{k_0}(\Delta)\xi_{k_0}(h) = \mathcal{U}_{k_0}(\Delta)u_{k_0}(h), \tag{10}$$

$$y_{k_0}(h) = \mathcal{V}_{k_0}(\Delta)\xi_{k_0}(h) + \mathcal{W}_{k_0}(\Delta)u_{k_0}(h), \tag{11}$$

where $\mathcal{T}_{k_0}(\Delta) := \sum_{i=0}^{r} \mathcal{T}_{i,k_0} R_m^i(\Delta)$, $\mathcal{U}_{k_0}(\Delta) := \sum_{i=0}^{r} \mathcal{U}_{i,k_0} R_p^i(\Delta)$, $\mathcal{V}_{k_0}(\Delta) : = \sum_{i=0}^{r} \mathcal{V}_{i,k_0} R_m^i(\Delta)$, $\mathcal{W}_{k_0}(\Delta) := \sum_{i=0}^{r} \mathcal{W}_{i,k_0} R_p^i(\Delta)$. Since the matrices $\mathcal{T}_{k_0}(\Delta)$, $\mathcal{U}_{k_0}(\Delta), \mathcal{V}_{k_0}(\Delta), \mathcal{W}_{k_0}(\Delta)$ are polynomial in Δ with constant coefficients, it seems natural to associate the following matrix with (10), (11):

$$S_{k_0}^M(\Delta) := \begin{bmatrix} -\mathcal{T}_{k_0}(\Delta) & \mathcal{U}_{k_0}(\Delta) \\ \mathcal{V}_{k_0}(\Delta) & \mathcal{W}_{k_0}(\Delta) \end{bmatrix}, \tag{12}$$

which is termed the *ω-stacked system matrix at time k_0 of model* (5), (6), thus extending the time-invariant Rosenbrock system matrix [28]. For the computation of the solutions of (10), (11), the following three types of elementary operations on the scalar rows of (10) can be useful, as for $\omega = 1$ [28]: (i) multiply any row by a nonzero real constant α; (ii) interchange rows i and j; (iii) add a multiple, by a polynomial $\beta(\Delta)$ in Δ with real coefficients, of row j to row i. On this basis, the next proposition can be easily proved, thus justifying the following assumption.

Assumption 2.1. The polynomial matrix $\mathcal{T}_{k_0}(\Delta)$ is square and nonsingular.

PROPOSITION 2.2. *If Assumption* 2.1 *does not hold, then one* (*or more*) *of the following situations occurs for equation* (10):

(α) *by a finite sequence of elementary operations of the types* (i), (ii), *and* (iii) *on the rows of equation* (10), *one of the scalar rows of the transformed equation of* (10) *can be reduced to the trivial identity* $0 = 0$;

(β) *there exists an ω-stacked input function $u_{k_0}(\cdot)$ for which* (10) *admits no solution for $h \geq 0$*;

(γ) *there exist solutions of* (10) *for $h \geq 0$ and for any $u_{k_0}(\cdot)$, but they depend on an infinite number of arbitrary and independent initial conditions.*

If Assumption 2.1 *holds, then, for each input function $u(\cdot)$, there exist solutions $\xi_{k_0}(\cdot), y_{k_0}(\cdot)$ of* (10), (11) *for $h \geq 0$, and they depend on arbitrary and independent initial conditions whose number is equal to the degree of* $\det\mathcal{T}_{k_0}(\Delta)$.

Relation (9) yields the following lemma about the dependence on k_0 of the matrices in (10), (11).

LEMMA 2.3. *If $\mathcal{T}_{k_0}(\Delta)$ is square, then the following identities hold*:

$$\mathcal{T}_{k_0+1}(\Delta) = R_m(\Delta)\mathcal{T}_{k_0}(\Delta)R_m^{-1}(\Delta), \quad \det\mathcal{T}_{k_0+1}(\Delta) = \det\mathcal{T}_{k_0}(\Delta), \tag{13}$$

$$\mathcal{U}_{k_0+1}(\Delta) = R_m(\Delta)\mathcal{U}_{k_0}(\Delta)R_p^{-1}(\Delta), \quad \mathcal{V}_{k_0+1}(\Delta) = R_q(\Delta)\mathcal{V}_{k_0}(\Delta)R_m^{-1}(\Delta), \tag{14}$$

$$\mathcal{W}_{k_0+1}(\Delta) = R_q(\Delta)\mathcal{W}_{k_0}(\Delta)R_p^{-1}(\Delta). \tag{15}$$

If Assumption 2.1 *holds for $k_0 = \overline{k}_0 \in \mathbf{Z}$, then it holds for any $k_0 \in \mathbf{Z}$, and the degree of* $\det\mathcal{T}_{k_0}(\Delta)$ *is independent of the initial time k_0.*

Hereafter, in view of Proposition 2.2, Assumption 2.1 will be assumed to hold; then, by virtue of Lemma 2.3, the degree of $\det \mathcal{T}_{k_0}(\Delta)$ for an arbitrary $k_0 \in \mathbf{Z}$ will be called the *order of model* (5), (6).

Now, consider a linear ω-periodic *model in state-space form*, i.e., a linear *system* described by:

$$x(k+1) = A(k)\,x(k) + B(k)\,u(k), \tag{16}$$

$$y(k) = C(k)\,x(k) + D(k)\,u(k), \tag{17}$$

where $k \in \mathbf{Z}$, $x(k) \in \mathbf{R}^n$ is the state, $u(k) \in \mathbf{R}^p$, $y(k) \in \mathbf{R}^q$, and $A(\cdot), B(\cdot), C(\cdot), D(\cdot)$ are real ω-periodic matrices. For this special case of an ω-periodic model, equations (10), (11) reduce to

$$R_n(\Delta)\,x_{k_0}(h) = \mathcal{A}_{k_0}\,x_{k_0}(h) + \mathcal{B}_{k_0}\,u_{k_0}(h), \tag{18}$$

$$y_{k_0}(h) = \mathcal{C}_{k_0}\,x_{k_0}(h) + \mathcal{D}_{k_0}\,u_{k_0}(h) \tag{19}$$

(where $x_{k_0}(h)$, $\mathcal{A}_{k_0}, \mathcal{B}_{k_0}, \mathcal{C}_{k_0}$, and $\mathcal{D}_{k_0}$ are the ω-stacked forms at time k_0 of $x(k)$, $A(k), B(k), C(k)$, and $D(k)$, respectively), which are termed the *ω-stacked form at time k_0 of system* (16), (17) (briefly, the *ω-stacked system at time k_0*). The stacked forms $\mathcal{A}_{k_0}$, $\mathcal{B}_{k_0}$, $\mathcal{C}_{k_0}$, $\mathcal{D}_{k_0}$ allow the notions of invariant zero, input (output) decoupling zero, and characteristic multiplier to be characterized and studied in a direct way for the ω-periodic system (16), (17) [13]. This is achieved through the matrix obtained by substituting the operator Δ by the complex variable z in

$$S^S_{k_0}(\Delta) := \begin{bmatrix} \mathcal{A}_{k_0} - R_n(\Delta) & \mathcal{B}_{k_0} \\ \mathcal{C}_{k_0} & \mathcal{D}_{k_0} \end{bmatrix}, \tag{20}$$

which from now on will be called the *ω-stacked system matrix at time k_0 of system* (16), (17).

Notice that, under Assumption 2.1, the roots of the equation $\det\mathcal{T}_{k_0}(z) = 0$ coincide with the values of z, for which there exists a nonzero solution of (5) for $u(\cdot) = 0$, satisfying the following:

$$\xi(k_0 + h\omega + \ell) = z^h \xi(k_0 + \ell) \quad \forall h \in \mathbf{Z}^+, \ell = 0, 1, \dots, \omega - 1; \tag{21}$$

therefore, such values play the same role for model (5), (6) that is played by the roots of the equation $\det[R_n(z) - \mathcal{A}_{k_0}] = 0$ for system (16), (17). In view of the results in [13], it seems natural to call *eigenvalues* or *characteristic multipliers of model* (5), (6) *at time* k_0, under Assumption 2.1, the zeros of the polynomial $\det\mathcal{T}_{k_0}(z)$ (and *characteristic polynomial of model* (5), (6) *at time* k_0 such a polynomial) with corresponding *ordered sets of structural indices* defined at the same time as their nondecreasing sequences of multiplicities as zeros of the invariant polynomials of $\mathcal{T}_{k_0}(z)$. In a similar way, under the same Assumption 2.1, the *invariant zeros*, *input decoupling zeros*, and *output decoupling zeros of model* (5), (6) *at time* k_0 are defined to be the zeros of the invariant polynomials of $S^M_{k_0}(z)$, $[-\mathcal{T}_{k_0}(z) \;\; \mathcal{U}_{k_0}(z)]$, $[-\mathcal{T}^T_{k_0}(z) \;\; \mathcal{V}^T_{k_0}(z)]^T$, respectively, with *ordered sets of structural indices* defined at the same time as their nondecreasing sequences of multiplicities as zeros of such polynomials.

The following proposition extends some results of [13] about system (16), (17) to model (5), (6); it can be proved on the basis of Lemmas 1.1 and 2.3 in a way similar to Theorem 3.4 in [13].

PROPOSITION 2.4. *Identities* (13)–(15) *still hold with* Δ *replaced by a scalar complex variable. Under Assumption* 2.1, *the rank of* $S^M_{k_0}(\Delta)$, *the nonzero invariant* *zeros, the nonzero input (output) decoupling zeros of model* (5), (6) *at time* k_0*, their ordered sets of structural indices, the whole characteristic polynomial of model* (5), (6) *at time* k_0*, and the ordered sets of structural indices of the nonzero characteristic multipliers are independent of* k_0.

Note that, under Assumption 2.1 for a fixed initial time k_0, the application of the z-transform to both sides of (10), (11), with zero initial conditions, yields $y_{k_0}(z) = G^M_{k_0}(z)\, u_{k_0}(z)$, where $G^M_{k_0}(z) := \mathcal{V}_{k_0}(z)\mathcal{T}^{-1}_{k_0}(z)\mathcal{U}_{k_0}(z) + \mathcal{W}_{k_0}(z)$ is called the *ω-stacked transfer matrix of model* (5), (6) *at* (*the initial*) *time* k_0. Proposition 2.4 yields

$$G^M_{k_0+1}(z) = R_q(z)G^M_{k_0}(z)R_p^{-1}(z) \quad \forall z \in \mathbf{C} \quad \forall k_0 \in \mathbf{Z}, \tag{22}$$

thus extending a similar relation for the ω-stacked transfer matrix of system (16), (17) [13]. By (22), if the transmission zeros and the poles of model (5), (6) at time k_0 and their ordered sets of structural indices are defined through the Smith–MacMillan form of $G^M_{k_0}(z)$, as in [13] those of system (16), (17), then by the same proofs as in [13] it is readily shown that the nonzero transmission zeros and poles of model (5), (6) and their ordered sets of structural indices are independent of time.

If $y_a(k)$ and $y_b(k)$ denote the output solutions of (5), (6) at time $k \geq k_0$ corresponding to given input functions $u_a(\cdot)$ and $u_b(\cdot)$, respectively, and to the same initial time k_0 and initial conditions, then model (5), (6) is said to be *causal* if $y_a(k) = y_b(k)$ for all the input functions $u_a(\cdot), u_b(\cdot)$ such that $u_a(t)|_{[k_0,k]} = u_b(t)|_{[k_0,k]}$, for all the initial conditions, for all $k \in \mathbf{Z}$, $k \geq k_0$, and for all $k_0 \in \mathbf{Z}$.

The proof of the following proposition is given in the appendix. A result similar to its first part was given in [20].

PROPOSITION 2.5. *Under Assumption* 2.1, *the ω-periodic model* (5), (6) *is causal only if, for all* $k_0 \in \mathbf{Z}$, *the corresponding ω-stacked model* (10), (11) *at time* k_0 *satisfies the following conditions*:

(i) $G^M_{k_0}(z)$ *is a proper rational matrix*;

(ii) *if* $G^M_{k_0}(z)$ *is rewritten as* $G^M_{k_0}(z) = F_{k_0}(z) + Q_{k_0}$ *with* $F_{k_0}(z)$ *strictly proper and* Q_{k_0} *constant, and* Q_{k_0} *is decomposed into blocks of dimensions* $q \times p$, *then* Q_{k_0} *is lower block triangular.*

If conditions (i) *and* (ii) *hold for* $k_0 = \bar{k}_0, \bar{k}_0 \in \mathbf{Z}$, *then* (i) *and* (ii) *hold for all* $k_0 \in \mathbf{Z}$.

3. System equivalence.

By formally using the same definition as the one introduced by Rosenbrock [28] for the time-invariant case, two $(m\omega + q\omega) \times (m\omega + p\omega)$ polynomial system matrices $S^1(\Delta)$ and $S^2(\Delta)$ with real coefficients are said to be *strictly system equivalent* if a relation of the following form holds:

$$S^2(\Delta) = \begin{bmatrix} M(\Delta) & 0 \\ Y(\Delta) & I_{q\omega} \end{bmatrix} S^1(\Delta) \begin{bmatrix} N(\Delta) & X(\Delta) \\ 0 & I_{p\omega} \end{bmatrix}, \tag{23}$$

where $M(\Delta), N(\Delta), X(\Delta)$, and $Y(\Delta)$ are polynomial matrices in Δ with real coefficients, and $M(\Delta), N(\Delta)$ are square and unimodular. It is stressed that strict system equivalence is an equivalence relation [28]. In addition, for the same reason as in [28] for the case $\omega = 1$, the following extra operations can be considered on the ω-stacked form (10), (11) at time k_0 of model (5), (6):

(a) for each $l = 0, \ldots, \omega - 1$, add to the vector component $\xi(k_0 + h\omega + l)$ of $\xi_{k_0}(h)$, ν scalar components, $\nu \geq 0$, which are defined to be equal to zero for each $h \geq 0$;

(b) for each $l = 0, \ldots, \omega - 1$, remove from the vector component $\xi(k_0 + h\omega + l)$ of $\xi_{k_0}(h)$, ν scalar components, $0 \leq \nu \leq m$, if they are equal to zero for each $\ell = 0, \ldots, \omega - 1$, for each $h \geq 0$, for all input functions $u(\cdot)$, and for all initial conditions.

The ω-stacked system matrix at time k_0, obtained from (10), (11) after an operation of type (a) has been carried out, is strictly system equivalent to the following one:

$$S_{k_0}^{ME}(\Delta) = \left[\begin{array}{cc|c} -I_{\nu\omega} & 0 & 0 \\ 0 & -\mathcal{T}_{k_0}(\Delta) & \mathcal{U}_{k_0}(\Delta) \\ \hline 0 & \mathcal{V}_{k_0}(\Delta) & \mathcal{W}_{k_0}(\Delta) \end{array}\right]. \tag{24}$$

A similar characterization of operation (b) holds. Then, two ω-periodic models of type (5), (6) satisfying Assumption 2.1 and having inputs and outputs of the same dimensions p and q, respectively, pseudostates of dimensions $m_i, i = 1, 2$, and corresponding ω-stacked models $\mathcal{M}_{k_0}^i$ of the form (10), (11), $i = 1, 2$, at the same time k_0, are said to be *system equivalent at time* k_0 if there exist an operation of type (a) or (b) to be carried out on $\mathcal{M}_{k_0}^1$ and an operation of type (a) or (b) to be carried out on $\mathcal{M}_{k_0}^2$ such that the ω-stacked system matrices at time k_0 associated with the resulting ω-stacked models at time k_0 are strictly system equivalent. For the proof of the following proposition see the appendix.

PROPOSITION 3.1. *The relation of system equivalence at time* k_0 *between two* ω*-periodic models of type* (5), (6) *is an equivalence relation.*

Remark 3.2. The solutions for $k \geq k_0$ of two ω-periodic models which are system equivalent at time k_0 are biuniquely related in the pseudostate, and are exactly the same in the output.

Such a statement is more deeply specified and clarified by the following proposition (which follows from well-known time invariant results [28], Lemma 2.3, relation (22), and Proposition 2.4) and by the subsequent remark.

PROPOSITION 3.3. *Given two* ω*-periodic models* $\mathcal{M}_1$ *and* $\mathcal{M}_2$ *of type* (5), (6) *satisfying Assumption* 2.1 *and having inputs and outputs of the same dimensions* p *and* q*, respectively, pseudostates of dimensions* $m_i, i = 1, 2$*, and the following* ω*-stacked system matrices at time* k_0:

$$S_{k_0,i}^{M}(\Delta) = \begin{bmatrix} -\mathcal{T}_{k_0,i}(\Delta) & \mathcal{U}_{k_0,i}(\Delta) \\ \mathcal{V}_{k_0,i}(\Delta) & \mathcal{W}_{k_0,i}(\Delta) \end{bmatrix}, \quad i = 1, 2, \tag{25}$$

if $\mathcal{M}_1$ *and* $\mathcal{M}_2$ *are system equivalent at time* k_0*, then*: (i) *the matrices in each of the pairs* $(S_{k_0,1}^M(z), S_{k_0,2}^M(z))$, $(\mathcal{T}_{k_0,1}(z), \mathcal{T}_{k_0,2}(z))$, $([-\mathcal{T}_{k_0,1}(z) \ \mathcal{U}_{k_0,1}(z)], [-\mathcal{T}_{k_0,2}(z) \ \mathcal{U}_{k_0,2}(z)])$, $([-\mathcal{T}_{k_0,1}^T(z) \ \mathcal{V}_{k_0,1}^T(z)]^T, [-\mathcal{T}_{k_0,2}^T(z) \ \mathcal{V}_{k_0,2}^T(z)]^T)$, *have the same Smith form, apart from some unit invariant polynomials, equal in number to* $\omega|m_1 - m_2|$; (ii) *the orders of* $\mathcal{M}_1$ *and* $\mathcal{M}_2$ *coincide*; (iii) *the* ω*-stacked transfer matrices of* $\mathcal{M}_1$ *and* $\mathcal{M}_2$ *at any initial time coincide*; (iv) $\mathcal{M}_1$ *and* $\mathcal{M}_2$ *have the same nonzero invariant zeros, nonzero input* (*output*) *decoupling zeros at all times, the same corresponding ordered sets of structural indices* (*apart from* $\omega|m_1 - m_2|$ *null structural indices*)*, the same characteristic multipliers at all times, and the same ordered sets of structural indices of their nonzero characteristic multipliers* (*apart from* $\omega|m_1 - m_2|$ *null structural indices*).

Remark 3.4. By Proposition 3.3, if a system $\mathcal{M}_2$ of the form (16), (17) is system equivalent at time k_0 to a given model $\mathcal{M}_1$, then all the features of $\mathcal{M}_2$ that are listed in items (ii), (iii), and (iv) of Proposition 3.3 are specified by the original model $\mathcal{M}_1$. Hence, such a system $\mathcal{M}_2$ is controllable (respectively, reconstructible) if and only if $\mathcal{M}_1$ has no nonzero input (respectively, output) decoupling zeros; it is stabilizable (respectively, detectable), if and only if $\mathcal{M}_1$ has no input (respectively, output) decoupling zeros outside the open disk of unit radius; it is reachable (respectively, observable) at time k_0 if and only if $\mathcal{M}_1$ has no input (respectively, output) decoupling zeros at time k_0 [13]. Moreover, the order, the ω-stacked transfer matrix at any time k_0, the asymptotic stability [7], the rate of convergence of the free motions, all the characteristic multipliers of system $\mathcal{M}_2$, and even the number and the dimensions of the Jordan blocks corresponding to each nonzero characteristic multiplier in the Jordan form of the monodromy matrix of system $\mathcal{M}_2$ (denoted later on by $E^S_{k_0}$) [13] at any time k_0 are determined by the properties of the original model $\mathcal{M}_1$. In addition, $S^M_{k_0,2}(z)$ has full row-rank for any $k_0 \in \mathbf{Z}$ and for any nonzero $z \in \mathbf{C}$ if and only if $S^M_{k_0,1}(z)$ has full row-rank (it is recalled that such a condition on the ω-stacked system matrix $S^M_{k_0,2}(z)$ of the ω-periodic system $\mathcal{M}_2$ described by equations of the form (16), (17) is necessary and sufficient, together with stabilizability and detectability, for the existence of a solution of the robust tracking and regulation problem when the ω-stacked forms of reference signals and disturbance functions have a time dependence of the form z^h [16]).

In view of the above discussion, it seems reasonable to look for an ω-periodic system of the form (16), (17) that is system equivalent at time k_0 to the given ω-periodic model (5), (6). The following section is devoted to this.

4. State-space representation for periodic models. For the ω-periodic model (5), (6) and its corresponding ω-stacked form at time k_0 (10), (11), if Assumption 2.1 holds, its ω-stacked transfer matrix $G^M_{k_0}(z)$ is proper, and its order $\overline{n}$ is less than or equal to $m\omega$, possibly after a preliminary operation of the type (a) has been carried out on (10), (11), then $S^M_{k_0}(\Delta)$ is strictly system equivalent to the following matrix [28]:

$$\left[\begin{array}{cc|c} -I_{m\omega-\overline{n}} & 0 & 0 \\ 0 & E_{k_0} - \Delta I_{\overline{n}} & J_{k_0} \\ \hline 0 & L_{k_0} & P_{k_0} \end{array}\right] =: S^{Ma}_{k_0}(\Delta) \tag{26}$$

(with $E_{k_0}, J_{k_0}, L_{k_0}$, and P_{k_0} being constant matrices), which will be called a *normal system matrix associated with* (10), (11).

The property stated by the following lemma will be used constructively to find a representation in the state-space form (16), (17) that is system equivalent at time k_0 to a given model of a periodic process having the form (5), (6).

LEMMA 4.1. *For the ω-periodic model* (5), (6), *if Assumption* 2.1 *holds, its ω-stacked transfer matrix $G^M_{\overline{k}_0}(z)$ satisfies conditions* (i) *and* (ii) *of Proposition* 2.5, *and* $\det\mathcal{T}_{\overline{k}_0}(z)|_{z=0} \neq 0$ *for some time $\overline{k}_0$, then for a normal system matrix $S^{Ma}_{k_0}(\Delta)$ of the form* (26) *associated with its ω-stacked model* (10), (11) *at any time k_0, the following properties hold:* (i) *matrix E_{k_0} is nonsingular;* (ii) *the matrix $H_{k_0} := P_{k_0} - L_{k_0}(E_{k_0})^{-1}J_{k_0}$ is upper block triangular if it is partitioned into blocks of dimensions $q \times p$.*

Proof. By (13) and Propositions 2.4 and 2.5, the same hypotheses of the lemma hold at all $k_0 \in \mathbf{Z}$. Then, for an arbitrary $k_0 \in \mathbf{Z}$, by (23) and (24) the following relation holds:

$$\det\mathcal{T}_{k_0}(\Delta) = c\det(\Delta I - E_{k_0}) \tag{27}$$

with $c \in \mathbf{R}$ being a nonzero constant, thus yielding (i). In addition, well-known time-invariant results [28] imply that

$$G^M_{k_0}(z) = L_{k_0}(zI - E_{k_0})^{-1}J_{k_0} + P_{k_0}. \tag{28}$$

Property (i) and relation (28) imply that $G^M_{k_0}(z)$ is analytic at $z = 0$; therefore, $G^M_{k_0}(z)$ can be expressed by means of a McLaurin series as $G^M_{k_0}(z) = \sum_{i=0}^{+\infty} G^{M,i}_{k_0} z^i$, whence (28) and the definition of H_{k_0} yield $H_{k_0} = G^M_{k_0}(0) = G^{M,0}_{k_0}$.

Now, let matrices $G^{M,i}_{k_0}$ be partitioned as follows in the same way as H_{k_0}:

(29)

$$G^{M,i}_{k_0} = \begin{bmatrix} G^{M,i}_{k_0,00} & G^{M,i}_{k_0,01} & \cdots & G^{M,i}_{k_0,0,j-1} & \cdots & G^{M,i}_{k_0,0,\omega-2} & G^{M,i}_{k_0,0,\omega-1} \\ G^{M,i}_{k_0,10} & G^{M,i}_{k_0,11} & \cdots & G^{M,i}_{k_0,1,j-1} & \cdots & G^{M,i}_{k_0,1,\omega-2} & G^{M,i}_{k_0,1,\omega-1} \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ G^{M,i}_{k_0,j0} & G^{M,i}_{k_0,j1} & \cdots & G^{M,i}_{k_0,j,j-1} & \cdots & G^{M,i}_{k_0,j,\omega-2} & G^{M,i}_{k_0,j,\omega-1} \\ G^{M,i}_{k_0,j+1,0} & G^{M,i}_{k_0,j+1,1} & \cdots & G^{M,i}_{k_0,j+1,j-1} & \cdots & G^{M,i}_{k_0,j+1,\omega-2} & G^{M,i}_{k_0,j+1,\omega-1} \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ G^{M,i}_{k_0,\omega-1,0} & G^{M,i}_{k_0,\omega-1,1} & \cdots & G^{M,i}_{k_0,\omega-1,j-1} & \cdots & G^{M,i}_{k_0,\omega-1,\omega-2} & G^{M,i}_{k_0,\omega-1,\omega-1} \end{bmatrix}.$$

By the arbitrariness of k_0, relation (22), and the McLaurin series expansion of $G^M_{k_0}(z)$ imply that

(30)

$$G^M_{k_0+1}(z) = \sum_{i=0}^{+\infty} G^{M,i}_{k_0+1} z^i = R_q(z) G^M_{k_0}(z) R_p^{-1}(z) = \sum_{i=0}^{+\infty} R_q(z) G^{M,i}_{k_0} R_p^{-1}(z) z^i$$

$$= \sum_{i=0}^{+\infty} \begin{bmatrix} G^{M,i}_{k_0,11} & \cdots & G^{M,i}_{k_0,1,j-1} & \cdots & G^{M,i}_{k_0,1,\omega-2} & G^{M,i}_{k_0,1,\omega-1} & z^{-1} G^{M,i}_{k_0,10} \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ G^{M,i}_{k_0,j1} & \cdots & G^{M,i}_{k_0,j,j-1} & \cdots & G^{M,i}_{k_0,j,\omega-2} & G^{M,i}_{k_0,j,\omega-1} & z^{-1} G^{M,i}_{k_0,j0} \\ G^{M,i}_{k_0,j+1,1} & \cdots & G^{M,i}_{k_0,j+1,j-1} & \cdots & G^{M,i}_{k_0,j+1,\omega-2} & G^{M,i}_{k_0,j+1,\omega-1} & z^{-1} G^{M,i}_{k_0,j+1,0} \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ G^{M,i}_{k_0,\omega-1,1} & \cdots & G^{M,i}_{k_0,\omega-1,j-1} & \cdots & G^{M,i}_{k_0,\omega-1,\omega-2} & G^{M,i}_{k_0,\omega-1,\omega-1} & z^{-1} G^{M,i}_{k_0,\omega-1,0} \\ z G^{M,i}_{k_0,0,1} & \cdots & z G^{M,i}_{k_0,0,j-1} & \cdots & z G^{M,i}_{k_0,0,\omega-2} & z G^{M,i}_{k_0,0,\omega-1} & G^{M,i}_{k_0,0,0} \end{bmatrix} z^i.$$

Let matrices $G^{M,i}_{k_0+1}$ be partitioned in the same way as $G^{M,i}_{k_0}$ in (29). By computing $G^{M,0}_{k_0+1}$ with the help of (30), we have $G^{M,0}_{k_0+1,\omega-1,\ell} = 0$, $\ell = 0, \dots, \omega - 2$; whence, by the arbitrariness of k_0, we have $G^{M,0}_{k_0,\omega-1,\ell} = 0$, $\ell = 0, \dots, \omega - 2$.

Now, the proof can be carried out by induction by showing that

$$(31) \qquad G^{M,0}_{k_0,j,\ell} = 0,\ j = \bar{j} - 1, \dots, \omega - 1,\ \ell = 0, \dots, j - 1$$

under the inductive assumption that

$$(32) \qquad G^{M,0}_{k_0,j,\ell} = 0,\ j = \bar{j}, \dots, \omega - 1,\ \ell = 0, \dots, j - 1$$

for $\bar{j} \le \omega - 1$, satisfying $\bar{j} \ge 2$. In fact, with the help of (30) it is readily seen that (32) implies (31) written with $k_0 + 1$ instead of k_0, and this proves (31) by the arbitrariness of k_0. Since (31) coincides with property (ii) for $\bar{j} = 2$, this completes the proof. □

Remark 4.2. In the special case of the state-space model (16), (17), $S^S_{k_0}(\Delta)$ is strictly system equivalent to (26) with $m = \bar{n} = n$. It is stressed that in this case the quadruplet $(E_{k_0}, J_{k_0}, L_{k_0}, P_{k_0})$ can be shown to coincide [17] within a coordinate transformation with the quadruplet $(E^S_{k_0}, J^S_{k_0}, L^S_{k_0}, P^S_{k_0})$ characterizing the time-invariant *associated system* (or *lifted representation*) *at time* k_0 *of system* (16), (17) [13], [16], [23]. $E^S_{k_0}$ is the monodromy matrix at time k_0 of system (16), (17), expressed by

$$E^S_{k_0} := A(k_0 + \omega - 1)A(k_0 + \omega - 2) \cdots A(k_0), \tag{33}$$

so that the characteristic multipliers, the invariant zeros, the input (output) decoupling zeros at time k_0, the corresponding ordered sets of structural indices, and the ω-stacked transfer matrix at time k_0 of system (16), (17) can be equivalently defined in terms of these quadruplets instead of the ω-stacked system (18), (19) at time k_0 [13]. Note that in the same case of system (16), (17), the hypothesis of nonsingularity of $R_n(0) - \mathcal{A}_{\overline{k}_0}$ means the nonsingularity of the monodromy matrix $E^S_{k_0}$, i.e., that of $A(k)$ for all $k \in \mathbf{Z}$, which implies the time reversibility of the causal system (16), (17), i.e., its state $x(k)$ and its output $y(k)$ at any time $k < k_0$ can be uniquely computed from the knowledge of the state $x(k_0)$ at time k_0, and of the values $u(j)$ of the input function at times $j < k_0$. This, together with the causality of system (16), (17), is easily seen to imply condition (ii) of Lemma 4.1.

Under the following assumption (which implies Assumption 2.1), the necessary and sufficient conditions will now be given for the existence of a periodic system of the form (16), (17) that is system equivalent at time k_0 to a given periodic model (5), (6).

Assumption 4.3. The polynomial matrix $\mathcal{T}_{k_0}(\Delta)$ is square and $\mathcal{T}_{k_0}(z)|_{z=0}$ is nonsingular.

By Proposition 2.4, if Assumption 4.3 holds for $k_0 = \overline{k}_0 \in \mathbf{Z}$, then it holds for any $k_0 \in \mathbf{Z}$ (as well as the pair of conditions (i) and (ii) of Proposition 2.5).

THEOREM 4.4. *For the ω-periodic model* (5), (6) *and its corresponding ω-stacked form* (10), (11) *at time k_0, under Assumption* 4.3, *there exists an ω-periodic system of the form* (16), (17) *which is system equivalent at time k_0 to the model* (5), (6), *if and only if its ω-stacked transfer matrix $G^M_{k_0}(z)$ is proper and satisfies condition* (ii) *of Proposition* 2.5.

Proof. Necessity. If model (5), (6) is system equivalent at time k_0 to system (16), (17), then it satisfies conditions (i) and (ii) of Proposition 2.5 by Theorem 3.1 in [13] and Proposition 3.3.

Sufficiency. Denote by $\mathcal{M}$ the given ω-periodic model (5), (6), and by n its order. Denote by $\overline{\mathcal{M}}$ the ω-periodic model—which is system equivalent to $\mathcal{M}$ at time k_0—whose ω-stacked form at time k_0 is obtained from the ω-stacked form (10), (11) at time k_0 of $\mathcal{M}$ through an operation of type (a) with $\nu := n - m$ if $m \leq n$, and with $\nu := 0$ if $m > n$. Denote by $\overline{\mathcal{T}}_{k_0}(\Delta)$ the $(\overline{m}\omega) \times (\overline{m}\omega)$ matrix thus obtained from $\mathcal{T}_{k_0}(\Delta)$ with $\overline{m} := m + \nu \geq n$, and denote by $\overline{S}^M_{k_0}(\Delta)$ the ω-stacked system matrix at time k_0 associated with $\overline{\mathcal{M}}$. The definition of operation (a) yields $\det\mathcal{T}_{k_0}(\Delta) = \det\overline{\mathcal{T}}_{k_0}(\Delta)$, thus implying that the degree of $\det\overline{\mathcal{T}}_{k_0}(\Delta)$ is equal to n and that $\overline{\mathcal{T}}_{k_0}(z)|_{z=0}$ is nonsingular; in addition, by Proposition 3.3, the ω-stacked transfer matrix of $\overline{\mathcal{M}}$ at time k_0 coincides with $G^M_{k_0}(z)$.

The hypotheses on $\overline{\mathcal{M}}$ guarantee that $\overline{S}^M_{k_0}(\Delta)$ is strictly system equivalent to a matrix $S^{Ma}_{k_0}(\Delta)$ of the form (26) [28], with n instead of $\overline{n}$, and E_{k_0} nonsingular by Lemma 4.1. Let J_{k_0}, L_{k_0}, and P_{k_0} be partitioned into blocks of dimensions $n \times p$, $q \times n$, and $q \times p$, respectively, as follows:

$$J_{k_0} = [\, J_{k_0,0} \quad J_{k_0,1} \quad \cdots \quad J_{k_0,\omega-1} \,], \tag{34}$$

$$(35)\qquad P_{k_0} = \begin{bmatrix} P_{k_0,0,0} & 0 & 0 & \dots & 0 \\ P_{k_0,1,0} & P_{k_0,1,1} & 0 & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots \\ P_{k_0,\omega-1,0} & P_{k_0,\omega-1,1} & P_{k_0,\omega-1,2} & \dots & P_{k_0,\omega-1,\omega-1} \end{bmatrix},$$

$$(36)\qquad L_{k_0} = \begin{bmatrix} L_{k_0,0} \\ L_{k_0,1} \\ \dots \\ L_{k_0,\omega-1} \end{bmatrix},$$

where the zero blocks in P_{k_0} are yielded by condition (ii) of Proposition 2.5, and by Proposition 3.3.

By virtue of property (ii) of Lemma 4.1, it is easy to check that

$$(37)\qquad \begin{bmatrix} \hat{M}(\Delta) & 0 \\ \hat{Y}(\Delta) & I_{q\omega} \end{bmatrix} S_{k_0}^{Ma}(\Delta) \begin{bmatrix} \hat{N}(\Delta) & \hat{X}(\Delta) \\ 0 & I_{p\omega} \end{bmatrix} = \hat{S}_{k_0}^{M}(\Delta)$$

with

(38)

$$\hat{S}_{k_0}^{M}(\Delta) = \left[\begin{array}{cc|c} -I_{\omega(\overline{m}-n)} & 0 & 0 \\ 0 & \mathrm{diag}\{A_0,\dots,A_{\omega-1}\} - R_n(\Delta) & \mathrm{diag}\{B_0,\dots,B_{\omega-1}\} \\ \hline 0 & \mathrm{diag}\{C_0,\dots,C_{\omega-1}\} & \mathrm{diag}\{D_0,\dots,D_{\omega-1}\} \end{array}\right],$$

$$(39)\quad \hat{M}(\Delta) = \begin{bmatrix} I_{\omega(\overline{m}-n)} & 0 & 0 & \dots & 0 & 0 \\ 0 & -I_n & 0 & \dots & 0 & 0 \\ 0 & 0 & -I_n & \dots & 0 & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & 0 & \dots & -I_n & 0 \\ 0 & E_{k_0} & E_{k_0} & \dots & E_{k_0} & I_n \end{bmatrix},$$

$$(40)\quad \hat{Y}(\Delta) = \begin{bmatrix} 0 & 0 & 0 & \dots & 0 & 0 \\ 0 & L_{k_0,1} & 0 & \dots & 0 & 0 \\ 0 & L_{k_0,2} & L_{k_0,2} & \dots & 0 & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & L_{k_0,\omega-1} & L_{k_0,\omega-1} & \dots & L_{k_0,\omega-1} & 0 \end{bmatrix},$$

$$(41)\quad \hat{N}(\Delta) = \begin{bmatrix} I_{\omega(\overline{m}-n)} & 0 & 0 & 0 & \dots & 0 & 0 \\ 0 & I_n & -I_n & 0 & \dots & 0 & 0 \\ 0 & 0 & I_n & -I_n & \dots & 0 & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & 0 & 0 & \dots & I_n & -I_n \\ 0 & I_n & 0 & 0 & \dots & 0 & 0 \end{bmatrix},$$

$$(42)\quad \hat{X}(\Delta) = \begin{bmatrix} 0 & 0 & \dots & 0 & 0 \\ (E_{k_0})^{-1}J_{k_0,0} & 0 & \dots & 0 & 0 \\ 0 & (E_{k_0})^{-1}J_{k_0,1} & \dots & 0 & 0 \\ \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & (E_{k_0})^{-1}J_{k_0,\omega-2} & 0 \\ 0 & 0 & \dots & 0 & 0 \end{bmatrix},$$

where

$$(43)\qquad A_i := I_n,\ i = 0,\dots,\omega-2,\ A_{\omega-1} := E_{k_0},$$

$$(44)\qquad B_i := (E_{k_0})^{-1}J_{k_0,i},\ i = 0,\dots,\omega-2,\ B_{\omega-1} := J_{k_0,\omega-1},$$

$$C_i := L_{k_0,i},\ i = 0, \dots, \omega - 1, \tag{45}$$

$$D_i := P_{k_0,ii},\ i = 0, \dots, \omega - 1, \tag{46}$$

and the block columns (respectively, rows) of $\hat{M}(\Delta)$ and $\hat{Y}(\Delta)$ (respectively, $\hat{N}(\Delta)$ and $\hat{X}(\Delta)$) in the same position have the same number of scalar columns (respectively, rows).

Since (38) exhibits a structure of the same form as (24) with $\nu = \overline{m} - n$, a suitable transformation of strict system equivalence on $\hat{S}^M_{k_0}(\Delta)$, followed by an operation of type (b) with $\nu = \overline{m} - n$, allow us to obtain the matrix $S^S_{k_0}(\Delta)$ defined by (20) with

$$A(k_0 + i + h\omega) := A_i, i = 0, \dots, \omega - 1 \quad \forall h \in \mathbf{Z}, \tag{47}$$

$$B(k_0 + i + h\omega) := B_i, i = 0, \dots, \omega - 1 \quad \forall h \in \mathbf{Z}, \tag{48}$$

$$C(k_0 + i + h\omega) := C_i, i = 0, \dots, \omega - 1 \quad \forall h \in \mathbf{Z}, \tag{49}$$

$$D(k_0 + i + h\omega) := D_i, i = 0, \dots, \omega - 1 \quad \forall h \in \mathbf{Z}. \tag{50}$$

Thus system (16), (17), with $A(k), B(k), C(k)$, and $D(k)$ defined by (47)–(50), is system equivalent at time k_0 to the original model (5), (6) by construction. □

Remark 4.5. The constructive procedure provided by the sufficiency proof of Theorem 4.4 gives rise, under Assumption 4.3 and conditions (i) and (ii) of Proposition 2.5, to an ω-periodic system (16), (17) which is time reversible, since $A(k)$ is nonsingular for all $k \in \mathbf{Z}$ because of the assumption $\det \mathcal{T}_{k_0}(z)|_{z=0} \neq 0$ by Proposition 3.3. It is stressed that many properties and features of the ω-periodic system (16), (17) thus found are determined by the original model (5), (6) (see Remark 3.4). In addition, by Assumption 4.3, model (5), (6) and, hence, system (16), (17) have no null characteristic multiplier; thus, they have no null input decoupling zero and output decoupling zero [13], and, therefore, system (16), (17) is reachable (respectively, observable) at all times, or is not, according to the properties of model (5), (6) and irrespective of the time k_0 used in the procedure.

Remark 4.6. If a larger equivalence relation is introduced between two ω-periodic models, in which it is allowed to add to the original model of the form (5), (6) some further pseudostate components in order to store the input values, then for this new equivalence relation a theorem wholly similar to Theorem 4.4 can be obtained by merely substituting Assumption 4.3 with the weaker Assumption 2.1 [17]. In this way, under no assumption about time reversibility, an ω-periodic system can be found such that its ω-stacked transfer matrix at any initial time, its output solutions at any initial time, its nonzero characteristic multipliers at any time, and its nonzero input (output) decoupling zeros at any time are still the same as those of the original model, as well as the corresponding ordered sets of structural indices (apart from some null structural indices), although their orders do not coincide, in general. Therefore, the asymptotic stability and the rate of convergence of the free motions, the controllability, the reconstructibility, the stabilizability, the detectability, and even the number and the dimensions of the Jordan blocks corresponding to each nonzero characteristic multiplier in the Jordan form of the monodromy matrix of such a system at any time k_0 are still determined by the properties of the original model, as well as the existence of a solution of the robust tracking and regulation problem [17].

5. Conclusions.

In this paper a polynomial time-invariant description of a discrete-time linear periodic process has been used in order to obtain a representation of it in state-space form.

A solution of this problem has been proposed within the class of models which are system equivalent at time k_0 to the given one, under an assumption on the latter implying the time reversibility of the corresponding state-space representation. It has been shown that the order, the characteristic multipliers, and the stacked transfer matrix at any initial time of the periodic system thus obtained coincide with those of the original periodic model, and that the asymptotic stability, the reachability, the observability, the controllability, the reconstructibility, the stabilizability, the detectability, and even the Jordan form of the monodromy matrix of such a system at any initial time are determined by the original periodic model, as well as the existence of a solution of the robust tracking and regulation problem.

A. Appendix.

Proof of Proposition 2.5. Item (i) trivially follows by contradiction from causality, taking into account that the ω-stacked model (10), (11) is just a compact description of the original periodic model (5), (6). For item (ii), call $\bar{h}$ the maximum degree in Δ among all the elements of $\mathcal{U}_{k_0}(\Delta)$ and $\mathcal{W}_{k_0}(\Delta)$, assume $u_{k_0}(h) = 0$ for all $h \neq \bar{h}$, and assume all the initial conditions of $\xi_{k_0}(h)$ to be zero, so that item (i) and the application of the z-transform to (10), (11) show that $y_{k_0}(\bar{h}) = Q_{k_0} u_{k_0}(\bar{h})$. Then, item (ii) follows by contradiction, since otherwise $y(k_0 + \bar{h}\omega + i)$ should depend on $u(k_0 + \bar{h}\omega + j)$ for some $i, j \in [0, \omega - 1], i < j$, thus preventing the model to be causal.

Now, let (i) and (ii) hold for $k_0 = \bar{k}_0$, and $G^M_{\bar{k}_0}(z)$ be partitioned as follows:

$$G^M_{\bar{k}_0}(z) := \begin{bmatrix} G_{11}(z) & G_{12}(z) \\ G_{21}(z) & G_{22}(z) \end{bmatrix}, \tag{51}$$

where $G_{11}(z)$ has dimensions $q \times p$ and is proper, $G_{12}(z)$ has dimensions $q \times (\omega - 1)p$ and is strictly proper, $G_{21}(z)$ has dimensions $(\omega - 1)q \times p$ and is proper, $G_{22}(z)$ has dimensions $(\omega - 1)q \times (\omega - 1)p$ and is proper, and item (ii), rewritten for $G_{22}(z)$ instead of $G^M_{\bar{k}_0}(z)$, holds. Then relation (22) yields

$$G^M_{\bar{k}_0+1}(z) := \begin{bmatrix} G_{22}(z) & z^{-1}G_{21}(z) \\ zG_{12}(z) & G_{11}(z) \end{bmatrix}, \tag{52}$$

thus proving (i) and (ii) for $k_0 = \bar{k}_0 + 1$. The proof is completed recursively by virtue of the ω-periodicity. □

Proof of Proposition 3.1. The reflexivity and symmetry properties are obvious. For transitivity, given three ω-periodic models $\mathcal{M}_i, i = 1, 2, 3$, having inputs and outputs of the same dimensions p and q, respectively, and having ω-stacked models $\mathcal{M}^i_{k_0}$ and ω-stacked pseudo-states $\xi^i_{k_0}(h)$ at time k_0, $i = 1, 2, 3$, assume that the pairs $\mathcal{M}_1$ and $\mathcal{M}_2$ and, respectively, $\mathcal{M}_2$ and $\mathcal{M}_3$, are system equivalent at time k_0. That is, let $\overline{\mathcal{M}}^1_{k_0}$ and $\overline{\mathcal{M}}^2_{k_0}$ (respectively, $\tilde{\mathcal{M}}^2_{k_0}$ and $\tilde{\mathcal{M}}^3_{k_0}$) be the ω-stacked models at time k_0 obtained after operations of type (a) or (b) are accomplished on $\mathcal{M}^1_{k_0}$ and $\mathcal{M}^2_{k_0}$ (respectively, $\mathcal{M}^2_{k_0}$ and $\mathcal{M}^3_{k_0}$), such that their two corresponding ω-stacked system matrices at time k_0 are strictly system equivalent. Let $\bar{\xi}^i_{k_0}(h), i = 1, 2,$ (respectively, $\tilde{\xi}^i_{k_0}(h), i = 2, 3$), be their ω-stacked pseudo-states, and let $\bar{\nu}_i \gtrless 0, i = 1, 2,$ (respectively, $\tilde{\nu}_i \gtrless 0, i = 2, 3$), be the number of zero scalar components added to or removed from each of the ω vector components of $\xi^i_{k_0}(h)$ for obtaining $\bar{\xi}^i_{k_0}(h)$ (respectively, $\tilde{\xi}^i_{k_0}(h)$), where $\bar{\nu}_i$ (respectively, $\tilde{\nu}_i$) is defined to be < 0 if $|\bar{\nu}_i|$ (respectively, $|\tilde{\nu}_i|$) scalar components are removed. Without loss of generality assume that $\overline{\nu}_2 \geq \tilde{\nu}_2$ (otherwise the ordering of the $\mathcal{M}_i$ could be reversed). Make two operations of type (a) on both $\tilde{\mathcal{M}}^2_{k_0}$ and $\tilde{\mathcal{M}}^3_{k_0}$ by adding the same number $\overline{\nu}_2 - \tilde{\nu}_2$ of scalar components to each of the ω vector components of $\tilde{\xi}^2_{k_0}(h)$ and $\tilde{\xi}^3_{k_0}(h)$, and denote by $\hat{\mathcal{M}}^2_{k_0}$ and $\hat{\mathcal{M}}^3_{k_0}$ the ω-stacked models at time k_0 thus obtained from $\tilde{\mathcal{M}}^2_{k_0}$ and $\tilde{\mathcal{M}}^3_{k_0}$, respectively. The proposition follows from the transitivity property of strict system equivalence by noting that $\hat{\mathcal{M}}^2_{k_0} = \overline{\mathcal{M}}^2_{k_0}$ and that the ω-stacked system matrices at time k_0 associated with $\hat{\mathcal{M}}^2_{k_0}$ and $\hat{\mathcal{M}}^3_{k_0}$ are strictly system equivalent, as well as those of $\overline{\mathcal{M}}^1_{k_0}$ and $\overline{\mathcal{M}}^2_{k_0}$. □

REFERENCES

[1] M. Araki and K. Yamamoto, *Multivariable multirate sampled-data systems: State space description, transfer characteristics, and Nyquist criterion*, IEEE Trans. Automat. Control, AC-31 (1986), pp. 145–154.

[2] S. Bittanti, *Deterministic and stochastic linear periodic systems*, in Time Series and Linear Systems, S. Bittanti, ed., Springer-Verlag, Berlin, 1986, pp. 141–182.

[3] S. Bittanti and P. Bolzern, *Discrete-time linear periodic systems: Gramian and modal criteria for reachability and controllability*, Internat. J. Control, 41 (1985), pp. 909–928.

[4] F. M. Callier and C. A. Desoer, *Multivariable Feedback Systems*, Springer-Verlag, New York, 1982.

[5] C. T. Chen, *Linear System Theory and Design*, Holt, Rinehart and Winston, New York, NY, 1984.

[6] P. Colaneri, *Zero-error regulation of discrete-time linear periodic systems*, Systems Control Lett., 15 (1990), pp. 161–167.

[7] D. S. Evans, *Finite-dimensional realizations of discrete-time weighting patterns*, SIAM J. Appl. Math., 22 (1972), pp. 45–67.

[8] P. A. Fuhrmann, *On strict system equivalence and similarity*, Internat. J. Control, 25 (1977), pp. 5–10.

[9] O. M. Grasselli, *A canonical decomposition of linear periodic discrete-time systems*, Internat. J. Control, 40 (1984), pp. 201–214.

[10] O. M. Grasselli and F. Lampariello, *Dead-beat control of linear periodic discrete-time systems*, Internat. J. Control, 33 (1981), pp. 1091–1106.

[11] O. M. Grasselli and S. Longhi, *Block decoupling with stability of linear periodic systems*, J. Math. Systems Estim. Control, 3 (1993), pp. 427–458.

[12] ———, *Disturbance localization by measurement feedback for linear periodic discrete-time systems*, Automatica J. IFAC, 24 (1988), pp. 375–385.

[13] ———, *Finite zero structure of linear periodic discrete-time systems*, Internat. J. Systems Sci., 22 (1991), pp. 1785–1806.

[14] ———, *The geometric approach for linear periodic discrete-time systems*, Linear Algebra Appl., 158 (1991), pp. 27–60.

[15] ———, *Pole placement for non-reachable periodic discrete-time systems*, Math. Control Signals Systems, 4 (1991), pp. 439–455.

[16] ———, *Robust tracking and regulation of linear periodic discrete-time systems*, Internat. J. Control, 54 (1991), pp. 613–633.

[17] O. M. Grasselli, S. Longhi, and A. Tornambè, *System equivalence for periodic models and systems*, Tech. report R-92-03, Dipartimento di Ingegneria Elettronica, II Università degli Studi di Roma "Tor Vergata", Roma, Italy, 1992.

[18] O. M. Grasselli and A. Tornambè, *On obtaining a realization of a polynomial matrix description of a system*, IEEE Trans. Automat. Control, 37 (1992), pp. 852–856.

[19] T. Kailath, *Linear Systems*, Prentice–Hall, Englewood Cliffs, NJ, 1980.

[20] P. P. Khargonekar, K. Poolla, and A. Tannenbaum, *Robust control of linear time-invariant plants using periodic compensation*, IEEE Trans. Automat. Control, AC-30 (1985), pp. 1088–1096.

[21] M. Kono, *Eigenvalue assignment in linear periodic discrete-time systems*, Internat. J. Control, 32 (1980), pp. 149–158.

[22] S. Longhi, A. M. Perdon, and G. Conte, *Geometric and algebraic structure at infinity of discrete-time linear periodic systems*, Linear Algebra Appl., 122–124 (1989), pp. 245–271.

[23] R. A. MEYER AND C. S. BURRUS, *A unified analysis of multirate and periodically time-varying digital filters*, IEEE Trans. Circuits Systems, 22 (1975), pp. 162–168.

[24] M. MORF, *Extended system and transfer function matrices and system equivalence*, in Proc. 1975 IEEE Conf. on Decision and Control, Houston, TX, 1975, pp. 199–206.

[25] B. PARK AND E. I. VERRIEST, *Canonical forms on discrete linear periodically time-varying systems and a control application*, in Proc. 28th IEEE Conf. Decision Control, Tampa, FL, 1989, pp. 1220–1225.

[26] L. A. PERNEBO, *Notes on strict system equivalence*, Internat. J. Control, 25 (1977), pp. 21–38.

[27] J. A. RICHARDS, *Analysis of Periodically Time-Varying Systems*, Springer-Verlag, Berlin, 1983.

[28] H. H. ROSENBROCK, *State-Space and Multivariable Theory*, Nelson, London, 1970.

[29] E. I. VERRIEST, *The operational transfer function and parametrization of N-periodic systems*, in Proc. 27th IEEE Conf. Decision Control, Austin, TX, 1988, pp. 1994–1999.

[30] W. A. WOLOVICH AND R. GUIDORZI, *A general algorithm for determining state-space representations*, Automatica J. IFAC, 13 (1977), pp. 295–299.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 33, No. 2, pp. 469–497, March 1995

© 1995 Society for Industrial and Applied Mathematics

008

SUPERVISORY CONTROL OF NONDETERMINISTIC SYSTEMS WITH DRIVEN EVENTS VIA PRIORITIZED SYNCHRONIZATION AND TRAJECTORY MODELS*

MARK A. SHAYMAN[†] AND RATNESH KUMAR[‡]

Abstract. The supervisory control of nondeterministic discrete event dynamical systems (DEDSs) with driven events in the setting of *prioritized synchronization* and *trajectory models* introduced by Heymann are studied. Prioritized synchronization captures the notions of controllable, uncontrollable, and driven events in a natural way, and the authors use it for constructing supervisory controllers. The trajectory model is used for characterizing the behavior of nondeterministic DEDSs since it is a sufficiently detailed model (in contrast to the less detailed language or failures models), and serves as a *language congruence* with respect to the operation of prioritized synchronization. Results concerning controllability and observability in this general setting are obtained.

Key words. discrete event systems, supervisory control, nondeterministic automata, driven events, prioritized synchronization, trajectory models

AMS subject classifications. 68Q75, 93B25, 93C83

1. Introduction.

Supervisory control of discrete event dynamical systems (DEDSs) was introduced by Ramadge and Wonham [23]. In this approach, the behavior of a DEDS, called the plant, is described by its language, that is the collection of all possible sequences of events (traces) that it can generate. The task is to design a controller, called a supervisor, which, based on the observation of the sequence of events, disables some of the *controllable* events so that the language generated by the controlled plant either equals a prespecified desired language, called a target language, or remains confined to a prespecified range of languages. Various extensions of this basic problem such as control under partial observation, decentralized and modular control, hierarchical control, and optimal control have also been studied. Refer to [24] and references therein for an overview of research in this area (up to 1989).

Most of the research on supervisory control of DEDSs assumes that the plant can be modeled as a *deterministic* system [10]. In other words, given a state of the system and an event that occurs in that state, the state reached after the occurrence of the event is uniquely known. Such an assumption is not satisfied whenever unmodeled dynamics, partial observation, or inherent nondeterminism are present. Hence the assumption of a deterministic plant is quite strong. In this paper, we relax this assumption and consider the control of a *nondeterministic* plant [10], [17], [18], [9], [11], [7].

A *modeling framework* m over a finite event set Σ is an equivalence relation on all DEDSs representable as state machines with arbitrary state space (finite or denumerable) having ϵ-transitions and event set Σ. We identify m with the projection π_m, which maps each state machine $\mathcal{P}$ to its equivalence class or *model* $\pi_m(\mathcal{P})$. If

* Received by the editors November 2, 1992; accepted for publication (in revised form) November 9, 1993. This research was supported in part by the Center for Robotics and Manufacturing, University of Kentucky, the National Science Foundation under the Engineering Research Centers Program Grant CDR-8803012, the Minta Martin Fund for Aeronautical Research, and the General Research Board at the University of Maryland.

[†] Department of Electrical Engineering and Institute for Systems Research, University of Maryland, College Park, Maryland 20742 (shayman@eng.umd.edu).

[‡] Department of Electrical Engineering, University of Kentucky, Lexington, Kentucky 40506-0046 (kumar@engr.uky.edu).

the equivalence class of $\mathcal{P}$ is uniquely characterized by an attribute, is common to its members, we will freely identify $\pi_m(\mathcal{P})$ with this attribute.

We say that a modeling framework π_m is *more detailed* than another modeling framework π_n if the equivalence relation π_m *refines* the equivalence relation π_n. Obviously, it is desirable to use the least detailed modeling framework which is sufficient for the design task at hand. A complex system is generally synthesized by combining simpler systems using various types of interconnections. Since specifications for the logical behavior of a DEDS are typically given in terms of the *language* of the system, a basic requirement is that the modeling framework should contain sufficient detail so that if the models for each subsystem are known, then the language of the interconnected system is uniquely determined. A modeling framework with such a property for a given class of admissible interconnections is referred to as a *language congruence* [7].

The language modeling framework associates with a system its language, the collection of all possible finite traces which are executable. Thus, the language model of a system is a subset of Σ^*, the set of all finite sequences of events in Σ including ϵ, the zero-length sequence. For deterministic systems and deterministic operators such as *strict synchronous composition* (SSC), the language modeling framework is a language congruence. If operators which introduce nondeterminism (e.g., internal choice, event internalization) are admissible, then the language modeling framework is no longer a language congruence and a more detailed modeling framework such as the *failures model* introduced by Hoare [9] must be used to have a language congruence. The failures model consists of the set of all *failures* of the system—pairs (s, Σ') where s is a trace and $\Sigma' \subseteq \Sigma$ is a refusal set with the property that if the environment restricts the possible events to Σ', the system can deadlock following execution of s. Thus, a failures model is a subset of $\Sigma^* \times 2^{\Sigma}$.[1]

In the work of Kumar, Garg, and Marcus [14], control design is accomplished by constructing a supervisor which operates in strict synchronization with the plant. In the work of Balemi et al. [3], the set of events Σ is partitioned into two disjoint subsets: *commands*, which are generated by the supervisor and sent to the plant, and *responses*, which are generated by the plant and sent to the supervisor. It is required that the plant and supervisor be *mutually receptive*, which means that the plant executes every command generated by the supervisor and the supervisor executes every response generated by the plant. Thus, this design also requires that every event be executed synchronously.

There are several reasons for considering control designs which do not require complete synchronization between the plant and supervisor. Uncontrollable events are generated spontaneously by the plant, and the supervisor is not permitted to interfere with their execution. Consequently, there is no a priori reason to assume that the supervisor needs to "track" every such event by undergoing a transition synchronously with the plant. Also, certain uncontrollable events in the plant may not be sensed and hence are invisible to the supervisor. It is unrealistic to require the supervisor to execute such events synchronously.

In many applications, it is not realistic to expect (or require) the plant to respond synchronously to every event generated by the supervisor. (Such events are referred to as *forcible* [6], *driven* [7], or *command* [3] events in the literature.) By permitting the supervisor to place commands which are not executed by the plant, nondeterminism in the plant may be resolved and performance improved. For example, not every piece

[1] For simplicity, we ignore the possibility of divergence.

of equipment in a factory will trigger an alarm upon breakdown. Breakdown may only be discovered when an action is requested by the supervisor and not executed by the plant. Thus, the unsensed state of the plant is determined by a synchronization failure.

Another motivation for relaxing the requirement of strict synchronization comes from systems in which a single supervisor controls more than one plant. For example, in a walking machine, there could be separate modules (viewed here as plants), which perform motion control and vision control, respectively. At a higher level, there could be a single supervisor which controls and coordinates the two modules. Some of the commands issued by the supervisor may apply to both the modules, while others may be relevant to only one of them and should be ignored by the other.

Heymann [7] has proposed a type of interconnection, called *prioritized synchronous composition* (PSC), which relaxes the synchronization requirements on the plant and supervisor. Each process in a PSC interconnection is assigned a *priority set* of events. For an event to be enabled in the interconnected system, it must be enabled in all processes whose priority sets contain that event. Also, when an enabled event occurs, it occurs in each subsystem in which the event is enabled. In the context of supervisory control, the priority set of the plant contains the controllable and uncontrollable events, while the priority set of the supervisor contains the controllable and driven events. Thus, controllable events require the participation of both plant and supervisor; uncontrollable events require the participation of the plant and will occur synchronously in the supervisor whenever possible; driven events require the participation of the supervisor and will occur synchronously in the plant whenever possible.

It is important to distinguish between PSC and other types of parallel composition in the literature. For example, Hoare [9] defines a concurrent composition operator in which each process has its own event set and the processes synchronize on the events in the intersection of their event sets. This is generalized to trace-dependent event sets, called event-control sets, by Inan and Varaiya [11]. The key difference between concurrent composition and PSC is that in PSC, although a process cannot block events which are outside its priority set, it may be able to execute these events and, whenever possible, will execute these events synchronously when they occur in the other process.[2]

It is shown in [7, Ex. 7] that two systems with the same failures model may yield different languages when composed in prioritized synchronization with a fixed system. Thus, if PSC is included as an admissible interconnection operator, a more detailed modeling framework than the failures model is required to serve as a language congruence. One such modeling framework, called the *trajectory model*, is proposed by Heymann [7] and Heymann and Meyer [8].[3] The trajectory model of a system consists of the set of all *trajectories* or *refusal-traces*—finite sequences of the type $\Sigma_0(\sigma_1, \Sigma_1)\cdots(\sigma_k, \Sigma_k)$, where $\sigma_1 \cdots \sigma_k$ is the trace executed by the system, while

[2] If applied to so-called *improper* processes, the parallel operator defined by Inan [12] can be viewed as a generalized form of PSC, but only in the deterministic setting. However, when supervisory control is considered in this reference, the assumption is made that the plant is proper and has a constant event control set. This assumption excludes driven events.

[3] The trajectory model is similar to the *failure-trace model* (also called the refusal-testing model) of Phillips [20], but differs from this model in its treatment of silent transitions (transitions labeled with ϵ). The trajectory model treats silent transitions in a way that is consistent with the failures model. While more detailed than the failures model, the failure-trace model is less detailed than the ready-trace model [21], [1], and hence less detailed than the bisimulation model [19], [18]. Comparison of various semantics for nondeterministic systems can be found in [27], [2].

$\Sigma_j \subseteq \Sigma$ $(j = 0, \dots, k)$ is a refusal set, a set of events which can result in deadlock if presented to the system by the environment at the indicated point in the refusal-trace. Thus, a trajectory model is a subset of $2^{\Sigma} \times (\Sigma \times 2^{\Sigma})^*$ and refines the failures model by including the intermediate refusal sets.

Although we use the trajectory model for describing the behavior of a nondeterministic plant, it is assumed that the desired specification is given only in terms of a language model (as in [23]), and not in terms of a trajectory model. This is a reasonable assumption, for in most applications, we are only interested in the sequences of events that a system can execute, and not in the events that the system may "refuse" to execute after execution of a certain event in a certain event sequence. Hence we address the following supervisory control problem:

> Given (i) a partition $\Sigma = \Sigma_c \cup \Sigma_u \cup \Sigma_d$ of the event set into subsets of controllable, uncontrollable, and driven events, (ii) a nondeterministic plant with trajectory model $P \subseteq 2^{\Sigma} \times (\Sigma \times 2^{\Sigma})^*$ whose priority set is $A = \Sigma_c \cup \Sigma_u$, and (iii) a (prefix-closed) target language $K \subseteq \Sigma^*$, design a supervisor—another trajectory model, denoted $S \subseteq 2^{\Sigma} \times (\Sigma \times 2^{\Sigma})^*$—whose priority set is $B = \Sigma_c \cup \Sigma_d$ such that the language of the PSC of P and S equals K.

The interconnection of the plant and the supervisor by PSC results in the disabling of some of the controllable events and the forcing of some of the driven events, while never preventing any of the uncontrollable events from occurring in the plant. Thus we investigate the supervisory control of DEDSs in the general setting of trajectory models and PSC, as opposed to language models and SSC studied by Kumar, Garg, and Marcus [14].

We obtain a necessary and sufficient condition for the existence of a supervisor for the general problem with driven events, and also provide a technique for synthesizing a supervisor. For ease of implementation, we design supervisors which are deterministic. We also address the control problem when some of the uncontrollable events are not observed by the supervisor. While the primary goal of this paper is to obtain necessary and sufficient conditions for the control of nondeterministic systems with driven events, a secondary goal is to provide a rigorous mathematical foundation for the theory of trajectory models and PSC.

The organization of this paper is as follows. In §2, an example is presented that motivates the design techniques to be developed in the remainder of the paper. In §3, the trajectory model of a nondeterministic state machine (NSM) with ϵ-moves is defined and its properties derived from those of NSMs. An algorithm for constructing a canonical NSM from a given trajectory model is presented and its correctness is proven. In §4, the PSC of NSMs is defined, and it is shown that this induces a PSC operation on trajectory models. It is also proven that the trajectory modeling framework is a language congruence relative to PSC. Properties of the PSC of trajectory models are described in §5, and the technique of *augmentation* is introduced. In §6, the supervisory control problem with driven events under both complete and partial observation is solved, and the results are applied to obtain a control design for the example system from §2.

An abbreviated version of this paper appeared in [25]. Extensions of many of the results to include nonclosed specifications and marking can be found in [15].

2. Motivating example.

In this section, we describe an example that motivates the results described in this paper. Figure 1(a) gives a deterministic model for a plant that processes a single type of part. Event α represents inputting a part. Event

β_1 represents successful completion and outputting of the part. Event β_2 represents completion and outputting of the part, but accompanied by an undetectable misalignment of an internal mechanism. If this has occurred, another part may be input, but this event α can be followed by an event λ that represents jamming of the machine. When this occurs, further processing is impossible. The event μ represents realignment of the misaligned internal mechanism. Since the misalignment of the internal mechanism is undetectable, the observation mask $M(\cdot)$ identifies the events β_1 and β_2, i.e., $M(\beta_1) = M(\beta_2) := \beta$. It is assumed that α is controllable and that $\beta_1, \beta_2, \lambda$ are uncontrollable. A natural performance specification is that λ should never occur and that the closed-loop generated language should include $(\alpha(\beta_1 + \beta_2\mu))^*$, i.e., cyclic operation should be possible.

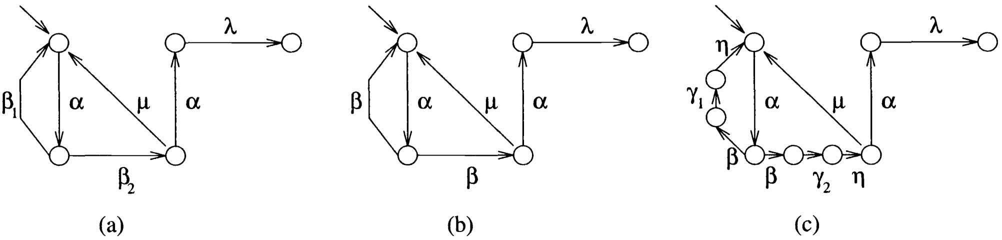


FIG. 1. *Diagram illustrating example of* §2.

Let us regard μ as a controllable event and consider whether the specifications can be met by a supervisor S of the Ramadge–Wonham type that is consistent with the observation mask. Since λ is uncontrollable, such a supervisor would need to disable α following any occurrence of β_2. Since the mask identifies β_1 and β_2, the supervisor must also disable α following β_1. Consequently, the generated closed-loop language imposed by any such supervisor is contained in $pr((\alpha\beta_2\mu)^*\alpha\beta_1)$ and thus fails to meet the lower-bound specification. ($pr(\cdot)$ denotes the prefix-closure operation.)

The design problem is also unsolvable using a forcing supervisor of the type considered by Golaszewski and Ramadge [6]. If such a supervisor forces μ following β_2, it must also force μ following β_1. Since the plant cannot execute μ after β_1, the controlled system would deadlock after the first occurrence of β_1. Thus, the lower-bound specification is not satisfied.

We could transform the partially observed deterministic system by identifying the events β_1, β_2 in the plant model and representing both of these events by their common mask value β. This yields the completely observed nondeterministic model depicted in Fig. 1(b). However, this results in a loss of information. In the first model, β_1, β_2 are indistinguishable only from the viewpoint of observation, while in the second model, they are indistinguishable for specification and control, as well as for observation. Since β_1, β_2 are uncontrollable, it is irrelevant whether they are distinguishable for the purpose of control. However, to be able to translate the original lower-bound specification into a corresponding specification on the transformed system, it is important that the events remain distinguishable from the viewpoint of specification. This can be accomplished by replacing β_1, β_2 by the three-event sequences $\beta\gamma_1\eta$, $\beta\gamma_2\eta$, respectively, where γ_1, γ_2 are completely unobservable, i.e., have mask value ϵ. If $M'(\cdot)$ denotes the mask for the transformed system, then $M'(\beta\gamma_1\eta) = M'(\beta)M'(\eta) = M'(\beta\gamma_2\eta)$. Thus, the substitution preserves the modeling assumption that the events β_1, β_2 have the same mask value. On the other hand, since γ_1, γ_2 are distinct event labels, the distinguishability of β_1, β_2 for the purpose of specification is also preserved. To model the uncontrollability of β_1, β_2, we designate β, γ_1, γ_2, η to be uncontrollable events. The "sandwiching" of the unobservable event γ_i between the observable events β, η reflects the fact that in the original model, the occurrence of β_i is known to the supervisor even though the supervisor cannot determine which of β_1, β_2 has occurred. With the new model, the supervisor knows that neither of the pair γ_1, γ_2 has occurred if β has not been observed. Similarly, it knows that one of the pair has occurred if η has been observed.

The substituted events can also be given a physical interpretation. β represents commenced processing of a part with or without internal undetectable misalignment. γ_1 and γ_2 represent the registering of faultless processing and of faulty processing, respectively. These are internal events that are modeled but whose occurrence is unobservable to an external process such as a supervisor. η represents the completion of processing and outputting of the part.

The transformed system is shown in Fig. 1(c). It is a partially observed nondeterministic system in which the mask is a natural projection and the unobservable events are uncontrollable. We will treat μ as a driven event rather than a controllable event as would be done in the Ramadge–Wonham theory. In the context of PSC-based control design, this allows for the possibility that the plant may refuse a request from the supervisor to execute this event. Using the results of §6, we will construct a PSC-based supervisor that meets the control specifications. (See Example 5.) The flexibility obtained by permitting the plant to refuse a supervisor-initiated event is an essential feature of the successful control design.

3. Trajectory model.

A plant, or a DEDS to be controlled, is modeled as an NSM with ϵ-moves. Letting $\mathcal{P}$ denote an NSM, it is defined to be the four tuple [10]: $\mathcal{P} := (X_{\mathcal{P}}, \Sigma, \delta_{\mathcal{P}}, x^0_{\mathcal{P}})$, where $X_{\mathcal{P}}$ denotes the state space of $\mathcal{P}$, Σ denotes the event set of $\mathcal{P}$, $\delta_{\mathcal{P}} : X_P \times \Sigma \cup \{\epsilon\} \to 2^{X_{\mathcal{P}}}$ denotes the nondeterministic[4] transition function of $\mathcal{P}$, and $x^0_{\mathcal{P}} \in X_{\mathcal{P}}$ denotes the initial state of $\mathcal{P}$. A triple $(x_1, \sigma, x_2) \in X_{\mathcal{P}} \times (\Sigma \cup \{\epsilon\}) \times X_{\mathcal{P}}$ is called a transition in $\mathcal{P}$ if $x_2 \in \delta_{\mathcal{P}}(x_1, \sigma)$. A transition (x_1, ϵ, x_2) is referred to as a *silent* transition. We assume that the plant NSM is finitely branching and cannot undergo an unbounded sequence of silent transitions.

3.1. Language model of a nondeterministic state machine.

As mentioned in §1, although trajectory models are used for describing the behaviors of nondeterministic systems, language or trace models are used for describing the desired or target specifications. Hence in this section we define the language model of an NSM $\mathcal{P}$. We first define the ϵ-closure of a state, which is the set of states reached by executing a finite sequence of silent transitions.

DEFINITION 1. *The* ϵ-closure *map,* $\epsilon^*_{\mathcal{P}} : X_{\mathcal{P}} \to 2^{X_{\mathcal{P}}}$, *is recursively defined to be*:

- $\forall x \in X_{\mathcal{P}} : \begin{cases} x \in \epsilon^*_{\mathcal{P}}(x), \\ x' \in \epsilon^*_{\mathcal{P}}(x) \Rightarrow \delta_{\mathcal{P}}(x', \epsilon) \subseteq \epsilon^*_{\mathcal{P}}(x). \end{cases}$

Using the definition of ϵ-closure, we extend the definition of the transition function from events to traces as follows.

DEFINITION 2. *The extension of the transition function to traces, denoted* $\delta^*_{\mathcal{P}} : X_{\mathcal{P}} \times \Sigma^* \to 2^{X_{\mathcal{P}}}$, *is defined inductively on the length of the traces as*:

- $\forall x \in X_{\mathcal{P}} : \begin{cases} \delta^*_{\mathcal{P}}(x, \epsilon) := \epsilon^*_{\mathcal{P}}(x), \\ \forall s \in \Sigma^*, \sigma \in \Sigma : \delta^*_{\mathcal{P}}(x, s\sigma) := \epsilon^*_{\mathcal{P}}(\delta_{\mathcal{P}}(\delta^*_{\mathcal{P}}(x, s), \sigma)), \end{cases}$

[4] The transition function $\delta_{\mathcal{P}}(\cdot, \cdot)$ is deterministic if and only if it is of the type $\delta_{\mathcal{P}} : X_{\mathcal{P}} \times \Sigma \to X_{\mathcal{P}}$, in which (i) there are no transitions labeled ϵ, and (ii) given a state and an event, either a unique state is reached upon execution of that event in that state, or that event is undefined in that state.

where in the last equality, the transition map is extended to $\delta_{\mathcal{P}} : 2^{X_{\mathcal{P}}} \times \Sigma \cup \{\epsilon\} \rightarrow 2^{X_{\mathcal{P}}}$, and the ϵ-closure map is extended to $\epsilon^*_{\mathcal{P}} : 2^{X_{\mathcal{P}}} \rightarrow 2^{X_{\mathcal{P}}}$ in the natural way. The set of states reached by executing a trace $s \in \Sigma^*$ from a state $x \in X_{\mathcal{P}}$ is given by the set $\delta^*_{\mathcal{P}}(x, s)$. It is clear that if $\mathcal{P}$ is deterministic, then the extension of the transition function to traces is also a deterministic partial map $\delta^*_{\mathcal{P}} : X_{\mathcal{P}} \times \Sigma^* \rightarrow X_{\mathcal{P}}$. (It is a partial map since it is generally defined only on a subset of $X_{\mathcal{P}} \times \Sigma^*$.)

The preceding definition can be used to obtain the language or trace model for the plant $\mathcal{P}$, denoted $L(\mathcal{P}) \subseteq \Sigma^*$, as follows:

- $L(\mathcal{P}) := \{s \in \Sigma^* \mid \delta^*_{\mathcal{P}}(x^0_{\mathcal{P}}, s) \neq \emptyset\}$.

3.2. Trajectory model of a nondeterministic state machine. As discussed in §1, language models are not adequate for characterizing the behavior of nondeterministic systems. Hence, we next define the trajectory model for an NSM $\mathcal{P}$. We first need to define the refusal map, and extend the transition function from events to refusal-traces.

DEFINITION 3. *The* refusal *map, $\Re_{\mathcal{P}} : X_{\mathcal{P}} \rightarrow 2^{\Sigma}$, is defined as*:

- $\forall x \in X_{\mathcal{P}} : \Re_{\mathcal{P}}(x) := \{\sigma \in \Sigma \mid \delta_{\mathcal{P}}(x', \sigma) = \emptyset, \forall x' \in \epsilon^*_{\mathcal{P}}(x)\}$.

Thus the refusal map defines, at each state, a set of events such that the system "refuses" to execute any of the events belonging to that set at that state. An event $\sigma \in \Sigma$ belongs to the refusal set of a state $x \in X_{\mathcal{P}}$ if and only if it is undefined at each state belonging to the ϵ-closure of x. Figure 2 depicts several NSMs defined over the event set $\{a, b\}$; each of the states is labeled with its refusal event set.

DEFINITION 4. *The extension of the transition function to refusal-traces, denoted $\delta^T_{\mathcal{P}} : X_{\mathcal{P}} \times (2^{\Sigma} \times (\Sigma \times 2^{\Sigma})^*) \rightarrow 2^{X_{\mathcal{P}}}$, is defined inductively on the length of the refusal-traces as*:

$$\bullet\ \forall x \in X_{\mathcal{P}} : \begin{cases} \forall \Sigma' \subseteq \Sigma : \delta^T_{\mathcal{P}}(x, \Sigma') := \{x' \in \epsilon^*_{\mathcal{P}}(x) \mid \Sigma' \subseteq \Re_{\mathcal{P}}(x')\}, \\ \forall e \in 2^{\Sigma} \times (\Sigma \times 2^{\Sigma})^*, \sigma \in \Sigma, \Sigma' \subseteq \Sigma : \\ \delta^T_{\mathcal{P}}(x, e(\sigma, \Sigma')) := \{x' \in \epsilon^*_{\mathcal{P}}(\delta_{\mathcal{P}}(\delta^T_{\mathcal{P}}(x, e), \sigma)) \mid \Sigma' \subseteq \Re_{\mathcal{P}}(x')\}. \end{cases}$$

A state $x' \in X_{\mathcal{P}}$ is reached by executing a "zero-length" refusal-trace $\Sigma' \subseteq \Sigma$ from a state $x \in X_{\mathcal{P}}$ if (i) x' belongs to the ϵ-closure of x, and (ii) the refusal set of x' contains Σ'. A state $x' \in X_{\mathcal{P}}$ is reached by executing a refusal-trace $e(\sigma, \Sigma') \in 2^{\Sigma} \times (\Sigma \times 2^{\Sigma})^*$ from a state $x \in X_{\mathcal{P}}$ if (i) x' belongs to the ϵ-closure of a state reached by executing the event σ from a state reached after executing the refusal-trace e from x, and (ii) the refusal set of x' contains Σ'. It is clear that if $\mathcal{P}$ is deterministic, then the extension of the transition function to refusal-traces is also a deterministic partial map $\delta^T_{\mathcal{P}} : X_{\mathcal{P}} \times (2^{\Sigma} \times (\Sigma \times 2^{\Sigma})^*) \rightarrow X_{\mathcal{P}}$. Based on the above extension of the transition function from events to refusal-traces, we define the trajectory model of the plant $\mathcal{P}$, which we denote as $T(\mathcal{P})$:

- $T(\mathcal{P}) := \{e \in 2^{\Sigma} \times (\Sigma \times 2^{\Sigma})^* \mid \delta^T_{\mathcal{P}}(x^0_{\mathcal{P}}, e) \neq \emptyset\}$.

We refer to the elements of this set as the *refusal-traces* or *trajectories* of $\mathcal{P}$.

Remark 1. There is a subtle but important difference in the meaning of the refusal sets in a trajectory model as opposed to those in an NSM. In the NSM, $\Re_{\mathcal{P}}(x)$ represents events that *must* be refused at the state x if offered by the environment. In contrast, the refusal set $\Sigma_i(e)$ in the refusal-trace e represents a set of events which *can* be refused if offered by the environment following execution of the previous fragment of the refusal-trace. The reason for this is that the *refusal-trace fragment* does not uniquely determine the state of the NSM due to nondeterminism. (Refer to Algorithm 1.)

Let $e = \Sigma_0(e)(\sigma_1(e), \Sigma_1(e)) \cdots (\sigma_n(e), \Sigma_n(e)) \in 2^{\Sigma} \times (\Sigma \times 2^{\Sigma})^*$ be a refusal-trace, where $n \in \mathcal{N}$, $\Sigma_i(e) \subseteq \Sigma$ and $\sigma_i(e) \in \Sigma$ for each $i \leq n$. We call n the length of e, and denote it as $|e| = n$. $\Sigma_i(e)$ is called the ith refusal set of e, and $\sigma_i(e)$ the ith event of e. For each $i \leq |e|$, we use e^i to denote the prefix of length i of e, i.e., $e^i := \Sigma_0(e) \cdots (\sigma_i(e), \Sigma_i(e))$. The notation $pr(e) \subseteq 2^\Sigma \times (\Sigma \times 2^\Sigma)^*$ is used to denote the set of all prefixes of e. Let $e, f \in 2^\Sigma \times (\Sigma \times 2^\Sigma)^*$ be refusal-traces. If f is a prefix of e, we indicate this by the notation $f \leq e$. We say that f is *dominated* by e, denoted $f \sqsubseteq e$, if $|f| = |e| := n$, $\sigma_i(f) = \sigma_i(e)$, and $\Sigma_i(f) \subseteq \Sigma_i(e)$ for each $i \leq n$. The notation $\text{dom}(e) \subseteq 2^\Sigma \times (\Sigma \times 2^\Sigma)^*$ is used to denote the set of all refusal-traces dominated by e.

Example 1. Consider a system $\mathcal{P}$ that deadlocks, i.e., cannot execute any transition, at its initial state. Then $T(\mathcal{P}) = \{\Sigma' \mid \Sigma' \subseteq \Sigma\}$, i.e., the trajectory model of $\mathcal{P}$ consists of all zero-length refusal-traces. We use $\Delta_\Sigma := \{\Sigma' \mid \Sigma' \subseteq \Sigma\}$ to denote the trajectory model of the *deadlock* system. Given $\sigma \in \Sigma$ and a trajectory model $P \subseteq 2^\Sigma \times (\Sigma \times 2^\Sigma)^*$, we use $\sigma \rightarrow P$ to denote the system that first executes the event σ and then follows with a refusal-trace in P. In other words, $\sigma \rightarrow P := pr\{\Sigma'(\sigma, e) \mid \Sigma' \subseteq \Sigma - \{\sigma\}, e \in P\}$. $\sigma \rightarrow P$ is called the *σ-prefix* operation on the trajectory model P. Given trajectory models $P_1, P_2 \subseteq 2^\Sigma \times (\Sigma \times 2^\Sigma)^*$, and $\sigma_1, \sigma_2 \in \Sigma$ with $\sigma_1 \neq \sigma_2$, the *external choice* between the trajectory models $\sigma_1 \rightarrow P_1$ and $\sigma_2 \rightarrow P_2$, denoted $(\sigma_1 \rightarrow P_1) + (\sigma_2 \rightarrow P_2)$, is defined to be the trajectory model:

$$(\sigma_1 \rightarrow P_1) + (\sigma_2 \rightarrow P_2) := \{e \in (\sigma_1 \rightarrow P_1) \cup (\sigma_2 \rightarrow P_2) \mid e^0 \in (\sigma_1 \rightarrow P_1) \cap (\sigma_2 \rightarrow P_2)\}.$$

This is a system which initially makes a deterministic choice between σ_1 and σ_2. If σ_i is executed, then the remainder of the refusal-trace is in P_i. The notation $P_1 \oplus P_2$ denotes the system that nondeterministically chooses to execute refusal-traces either in P_1 or in P_2. $P_1 \oplus P_2$ is called the *internal choice* between P_1 and P_2, and $P_1 \oplus P_2 := P_1 \cup P_2$.

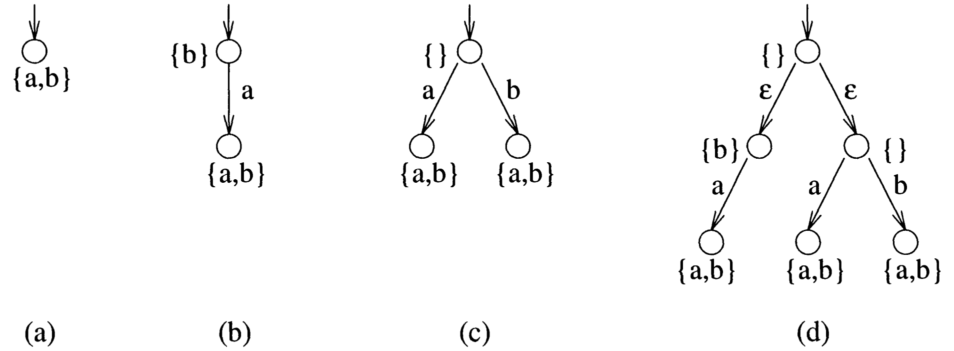


FIG. 2. *Diagram illustrating Example* 1.

Figures 2(a)–(d) depict NSMs defined over the event set $\{a, b\}$; each of the states is labeled with the set of events that are refused at that state. Figure 2(a) depicts an NSM that deadlocks. Hence its trajectory model is $\Delta_{\{a,b\}}$. Figure 2(b) depicts an NSM that initially executes the event a and then deadlocks. Hence its trajectory model is $a \rightarrow \Delta_{\{a,b\}}$. Figure 2(c) depicts an NSM that initially makes a deterministic choice between the events a and b and deadlocks after executing either of the events. Hence its trajectory model is given by $(a \rightarrow \Delta_{\{a,b\}}) + (b \rightarrow \Delta_{\{a,b\}})$. Figure 2(d) depicts an NSM that initially makes a nondeterministic choice between the systems of Figs. 2(b) and (c). Hence its trajectory model is $(a \rightarrow \Delta_{\{a,b\}}) \oplus [(a \rightarrow \Delta_{\{a,b\}}) + (b \rightarrow \Delta_{\{a,b\}})]$.

It follows from the definition of the trajectory model $T(\mathcal{P})$ that it satisfies the following five properties, denoted (T1), (T2), (T3), (T4), and (T5).

PROPOSITION 1. *The trajectory model $T(\mathcal{P})$ of an NSM $\mathcal{P}$ satisfies the following properties:*

(T1) (*nonemptiness*): $\emptyset \in T(\mathcal{P}) \Rightarrow T(\mathcal{P}) \neq \emptyset$;

(T2) (*prefix closure*): $\forall e \in T(\mathcal{P}), f \in 2^\Sigma \times (\Sigma \times 2^\Sigma)^* : f < e \Rightarrow f \in T(\mathcal{P})$;

(T3) (*dominance closure*): $\forall e \in T(\mathcal{P}), f \in 2^\Sigma \times (\Sigma \times 2^\Sigma)^* : f \sqsubset e \Rightarrow f \in T(\mathcal{P})$;

(T4) (*refusal of infeasible*): $\forall e \in T(\mathcal{P}), i \leq |e|, \sigma \in \Sigma : e^i(\sigma, \emptyset) \notin T(\mathcal{P}) \Rightarrow e^{i-1}(\sigma_i(e), \Sigma_i(e) \cup \{\sigma\}) \dots (\sigma_{|e|}, \Sigma_{|e|}) \in T(\mathcal{P})$;

(T5) (*persistence of refused*): $\forall e \in T(\mathcal{P}), i \leq |e|, \sigma \in \Sigma : \sigma \in \Sigma_i(e) \Rightarrow \sigma_{i+1}(e) \neq \sigma$.

Proof. (T1), (T2), and (T5) follow immediately from the definition of the trajectory model. To prove (T3), we note that a straightforward induction on length of refusal-traces shows that if $f \sqsubset e$, then $\delta^T_{\mathcal{P}}(x^0_{\mathcal{P}}, e) \subseteq \delta^T_{\mathcal{P}}(x^0_{\mathcal{P}}, f)$, which immediately yields (T3). It remains to prove (T4). Fix i, and suppose that $e^i(\sigma, \emptyset) \notin T(\mathcal{P})$. Then $\delta_{\mathcal{P}}(\delta^T_{\mathcal{P}}(x^0_{\mathcal{P}}, e^i), \sigma) = \emptyset$. Since $\epsilon^*_{\mathcal{P}}(\delta^T_{\mathcal{P}}(x^0_{\mathcal{P}}, e^i)) = \delta^T_{\mathcal{P}}(x^0_{\mathcal{P}}, e^i)$, this implies that $\sigma \in \Re_{\mathcal{P}}(x)$ for all $x \in \delta^T_{\mathcal{P}}(x^0_{\mathcal{P}}, e^i)$. It follows immediately that if $\bar{e}$ is obtained from e by replacing $\Sigma_i(e)$ with $\Sigma_i(e) \cup \{\sigma\}$, then $\delta^T_{\mathcal{P}}(x^0_{\mathcal{P}}, \bar{e}) = \delta^T_{\mathcal{P}}(x^0_{\mathcal{P}}, e)$, which implies that $\bar{e} \in T(\mathcal{P})$. □

Remark 2. In contrast to [8] where the properties of the trajectory model are defined axiomatically, we regard the NSM as the fundamental object and *derive* the properties of the trajectory model from the properties of NSMs.

3.3. Construction of canonical nondeterministic state machine. In this section we develop an algorithm for constructing a canonical nondeterministic state machine for any given set of refusal-traces satisfying (T1)–(T5).

DEFINITION 5. *Let $P \subseteq 2^\Sigma \times (\Sigma \times 2^\Sigma)^*$ be a refusal-trace set satisfying* (T1)–(T5). *If $e \in P$ is any refusal-trace which has the property that for each $i \leq |e|$, $\sigma \in \Sigma_i(e)$ whenever $e^i(\sigma, \emptyset) \notin P$, then we say that e is* saturated. *The* saturated trajectory model, *denoted P_{sat}, of P is defined to be*: $P_{\text{sat}} := \{e \in P \mid e$ *is saturated*$\}$.

It is easy to see that a prefix of a saturated refusal-trace is also saturated, and each refusal-trace of P is dominated by a saturated refusal-trace of P, so $\text{dom}(P_{\text{sat}}) = P$. Thus P_{sat} is equivalent in detail of description to P. So, we use the set of saturated refusal-traces for the construction of the canonical nondeterministic state machine. Given a finite number of event sets $\Sigma_1, \dots, \Sigma_n \subseteq \Sigma$ for some $n \in \mathcal{N}$, we use the notation $\min(\Sigma_1, \dots, \Sigma_n)$ to denote the collection of minimal sets from among the given n sets, i.e.,

- $\min(\Sigma_1, \dots, \Sigma_n) := \{\Sigma_i, 1 \leq i \leq n \mid \nexists j$ such that $1 \leq j \leq n; j \neq i; \Sigma_j \subset \Sigma_i\}$.

LEMMA 1. *Let $P \subseteq 2^\Sigma \times (\Sigma \times 2^\Sigma)^*$ satisfy* (T1)–(T5).

1. *P_{sat} contains a unique minimal zero-length refusal-trace $\Sigma^0_{\min} := \{\sigma' \in \Sigma \mid \sigma' \notin L(P)\}$.*
2. *If $e \in P_{\text{sat}}$ and $e(\sigma, \emptyset) \in P$, then the family $\{\Sigma' \subseteq \Sigma \mid e(\sigma, \Sigma') \in P_{\text{sat}}\}$ has a unique minimal element given by $\Sigma^{(e,\sigma)}_{\min} := \{\sigma' \in \Sigma \mid e(\sigma, \emptyset)(\sigma', \emptyset) \notin P\}$.*

Proof. The proof of the first part is similar to that of the second part, so we include only the latter. Since $e(\sigma, \emptyset) \in P$, repeated application of (T4) yields $e(\sigma, \Sigma^{(e,\sigma)}_{\min}) \in P$. Since e is saturated, in order to show that $e(\sigma, \Sigma^{(e,\sigma)}_{\min})$ is saturated, it suffices to show that if $e(\sigma, \Sigma^{(e,\sigma)}_{\min})(\sigma', \emptyset) \notin P$, then $\sigma' \in \Sigma^{(e,\sigma)}_{\min}$. Suppose $\sigma' \notin \Sigma^{(e,\sigma)}_{\min}$. Then $e(\sigma, \emptyset)(\sigma', \emptyset) \in P$. By repeated application of T4, it follows that $e(\sigma, \Sigma^{(e,\sigma)}_{\min})(\sigma', \emptyset) \in P$, which is a contradiction. Thus, $e(\sigma, \Sigma^{(e,\sigma)}_{\min}) \in P_{\text{sat}}$.

Finally, suppose $e(\sigma, \Sigma') \in P_{\rm sat}$ and $\sigma' \in \Sigma_{\min}^{(e,\sigma)}$, i.e., $e(\sigma, \emptyset)(\sigma', \emptyset) \notin P$. By (T3), it follows that $e(\sigma, \Sigma')(\sigma', \emptyset) \notin P$. Since $e(\sigma, \Sigma')$ is saturated, $\sigma' \in \Sigma'$, so $\Sigma_{\min}^{(e,\sigma)} \subseteq \Sigma'$. □

ALGORITHM 1. *Given* $P \subseteq 2^\Sigma \times (\Sigma \times 2^\Sigma)^*$ *satisfying* (T1)–(T5), *construct a nondeterministic state machine (with* ϵ*-moves)* $\mathcal{P} := (X_\mathcal{P}, \Sigma, \delta_\mathcal{P}, x_\mathcal{P}^0)$, *where*

- $X_\mathcal{P} := P_{\rm sat}$ *is the state space of* $\mathcal{P}$,
- $x_\mathcal{P}^0 := \Sigma_{\min}^0$ *is the initial state of* $\mathcal{P}$,
- $\delta_\mathcal{P} : X_\mathcal{P} \times \Sigma \cup \{\epsilon\} \to 2^{X_\mathcal{P}}$ *is the transition function of* $\mathcal{P}$ *defined as:*

(1) $\forall e \in P_{\rm sat}$, $\sigma \in \Sigma$:

$$\delta_\mathcal{P}(e, \sigma) := \begin{cases} e(\sigma, \Sigma_{\min}^{(e,\sigma)}) & \text{if } e(\sigma, \emptyset) \in P \\ \emptyset & \text{otherwise,} \end{cases}$$

(2(a)) $\forall \Sigma' \subseteq \Sigma$ *such that* $\Sigma' \in P_{\rm sat}$:

$$\delta_\mathcal{P}(\Sigma', \epsilon) := \min(\{\Sigma'' \subseteq \Sigma | \; \Sigma'' \in P_{\rm sat}, \; \Sigma' \subset \Sigma''\}),$$

(2(b)) $\forall e \in 2^\Sigma \times (\Sigma \times 2^\Sigma)^*, \sigma \in \Sigma, \Sigma' \subseteq \Sigma$ *such that* $e(\sigma, \Sigma') \in P_{\rm sat}$:

$$\delta_\mathcal{P}(e(\sigma, \Sigma'), \epsilon) := \{e(\sigma, \Sigma'') \,|\, \Sigma'' \in \min(\{\hat\Sigma \subseteq \Sigma \,|\, e(\sigma, \hat\Sigma) \in P_{\rm sat}, \Sigma' \subset \hat\Sigma\})\}.$$

Algorithm 1 provides a procedure for constructing a canonical NSM $\mathcal{P}$ for a given set of refusal-traces P satisfying (T1)–(T5). The state space of $\mathcal{P}$ equals $P_{\rm sat}$, the set of saturated refusal-traces of P, and the initial state of $\mathcal{P}$ is the minimal zero-length saturated refusal-trace $\Sigma_{\min}^0$ of P. The state reached by executing a nonepsilon event $\sigma \in \Sigma$ from a state $e \in P_{\rm sat}$ equals the minimal saturated refusal-trace of the type $e(\sigma, \Sigma')$ dominating $e(\sigma, \emptyset)$. The set of states reached by executing an epsilon transition from a zero-length refusal-trace $\Sigma' \in P_{\rm sat} = X_\mathcal{P}$ equals the minimal elements of the set of zero-length saturated refusal-traces dominating Σ'. Also, the set of states reached by executing an epsilon transition from a refusal-trace $e(\sigma, \Sigma') \in P_{\rm sat} = X_\mathcal{P}$ equals the set of saturated refusal-traces of the type $e(\sigma, \Sigma'')$ dominating $e(\sigma, \Sigma')$ with Σ'' minimal. Note that the canonical NSM constructed using Algorithm 1 has as many states as the number of saturated refusal-traces.

Remark 3. A construction which bears some similarity to Algorithm 1 was informally described in [8, Alg. 12.1]. However, a proof showing that the trajectory model of the canonical NSM equals P was omitted in that reference. There is also an important difference between the two algorithms. The construction in [8, Alg. 12.1] is based on prefixes of dominant refusal-traces, i.e., refusal-traces which are maximal with respect to $\sqsubseteq$ partial order. In contrast, Algorithm 1 is based on saturated refusal-traces. The use of saturated refusal-traces for the states has the advantage of avoiding the need to introduce certain "auxiliary states," which is the case when prefixes of dominant refusal-traces are used. This advantage arises because the saturated refusal-traces satisfy the properties described in Lemma 1.

Example 2. The trajectory model of the NSM shown in Fig. 2(d) is $P := pr(\mathrm{dom}(P'))$, where $P' := \{\{b\}(a, \{a, b\}), \; \emptyset(b, \{a, b\})\}$. Since P is obtained as the trajectory model of an NSM, it follows from Proposition 1 that P satisfies (T1)–(T5). Consequently, Algorithm 1 may be applied to obtain the canonical NSM $\mathcal{P}$ with trajectory model P. The state space of $\mathcal{P}$ is the set of saturated refusal-traces

$$P_{\rm sat} = pr(\{\emptyset(a, \{a, b\}), \; \{b\}(a, \{a, b\}), \; \emptyset(b, \{a, b\})\}),$$

which contains five elements. The canonical NSM $\mathcal{P}$ with these five states is depicted in Fig. 3(a). Each node is labeled with the name of the state—a saturated refusal-trace—that it represents. The NSM depicted in Fig. 3(b) with four states also has the same trajectory model.

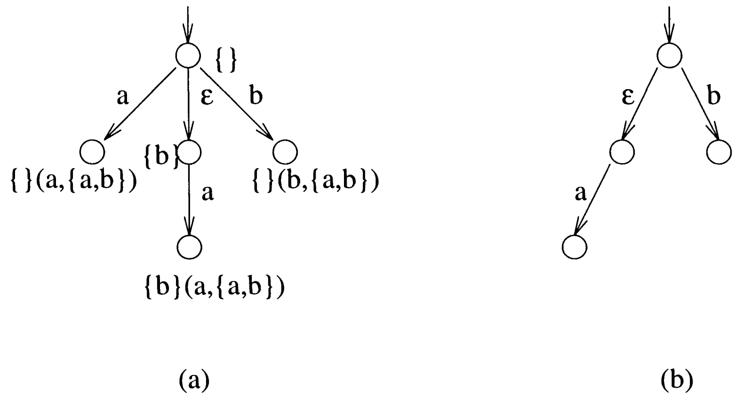


FIG. 3. *Diagram illustrating Example* 2.

The equality of the *failures* models of the two NSMs shown in Figs. 3(a) and (b) and the NSM shown in Fig. 2(d) is an instance of a basic property of the failures model concerning the operations of external choice and event concealment [9, Law L10, p. 113]. Since the three NSMs have identical trajectory models, the additional detail present in the trajectory model still does not distinguish among the three systems. However, the failure-trace model of Phillips does distinguish between the systems analogous to those in Figs. 2(d) and 3(b) with τ in place of ϵ [20, Ex. 3, p. 250].

We now prove the correctness of Algorithm 1, i.e., that the trajectory model of the canonical NSM $\mathcal{P}$ equals P.

PROPOSITION 2. *Let* $P \subseteq 2^\Sigma \times (\Sigma \times 2^\Sigma)^*$ *satisfy* (T1)–(T5). *Then* $T(\mathcal{P}) = P$, *where* $\mathcal{P}$ *is as constructed in Algorithm* 1.

Proof. We begin by showing that

$$(1) \qquad \forall\, e = \bar{e}(\sigma, \Sigma') \in P_{\text{sat}}, \quad \Re_{\mathcal{P}}(e) = \Sigma'.$$

It follows from the definition of $\delta_{\mathcal{P}}$ that $\sigma' \in \Re_{\mathcal{P}}(e)$ if and only if for each $f = \bar{e}(\sigma, \Sigma'') \in P_{\text{sat}}$ such that $\Sigma' \subseteq \Sigma''$, $f(\sigma', \emptyset) \notin P$. If $\sigma' \in \Sigma'$, then $\sigma' \in \Sigma''$ for all such Σ''. By T5, $f(\sigma', \emptyset) \notin P$, so $\sigma' \in \Re_{\mathcal{P}}(e)$. Thus, $\Sigma' \subseteq \Re_{\mathcal{P}}(e)$. On the other hand, if $\sigma' \notin \Sigma'$, then since $e \in P_{\text{sat}}$, it follows that $e(\sigma', \emptyset) \in P$, so $\sigma' \notin \Re_{\mathcal{P}}(e)$. Thus, $\Re_{\mathcal{P}}(e) \subseteq \Sigma'$, proving (1).

Next, we claim that

$$(2) \qquad \delta^T_{\mathcal{P}}(x^0_{\mathcal{P}}, e) = \{f \in P_{\text{sat}} \mid e \sqsubseteq f\} \quad \forall\, e \in 2^\Sigma \times (\Sigma \times 2^\Sigma)^*.$$

We prove (2) by induction on $|e|$. Let $e = \Sigma' \subseteq \Sigma$, a zero-length refusal-trace. Using the definition of $\delta_{\mathcal{P}}$ and (1) gives $\delta^T_{\mathcal{P}}(x^0_{\mathcal{P}}, \Sigma') = \{\Sigma'' \in \epsilon^*_{\mathcal{P}}(x^0_{\mathcal{P}}) \mid \Sigma' \subseteq \Re_{\mathcal{P}}(\Sigma'')\} = \{\Sigma'' \in P_{\text{sat}} \mid \Sigma' \subseteq \Sigma''\}$. This establishes (2) in the zero-length case.

For the induction step, let $e = \bar{e}(\sigma, \Sigma') \in 2^\Sigma \times (\Sigma \times 2^\Sigma)^*$. Using the induction hypothesis on $\bar{e}$, (1), and the fact that P_{sat} is prefix-closed gives

$$\delta^T_{\mathcal{P}}(x^0_{\mathcal{P}}, e) = \{f \in \epsilon^*_{\mathcal{P}}(\delta_{\mathcal{P}}(\delta^T_{\mathcal{P}}(x^0_{\mathcal{P}}, \bar{e}), \sigma)) \mid \Sigma' \subseteq \Re_{\mathcal{P}}(f)\}$$

$$\begin{aligned}
&= \{f \in \epsilon^*_{\mathcal{P}}(\delta_{\mathcal{P}}(\{\bar{f} \in P_{\text{sat}} |\, \bar{e} \sqsubseteq \bar{f}\}, \sigma)) |\, \Sigma' \subseteq \Re_{\mathcal{P}}(f)\} \\
&= \{\bar{f}(\sigma, \Sigma'') \in P_{\text{sat}} |\, \bar{e} \sqsubseteq \bar{f},\ \Sigma' \subseteq \Sigma''\} \\
&= \{f \in P_{\text{sat}} |\, e \sqsubseteq f\}.
\end{aligned}$$

This completes the induction step and establishes (2).

If $e \in P_{\text{sat}}$, (2) implies that $e \in \delta^T_{\mathcal{P}}(x^0_{\mathcal{P}}, e)$. Thus, $\delta^T_{\mathcal{P}}(x^0_{\mathcal{P}}, e)$ is nonempty, so $e \in T(\mathcal{P})$. Hence $P_{\text{sat}} \subseteq T(\mathcal{P})$. Since every refusal-trace in P is dominated by a saturated refusal-trace and $T(\mathcal{P})$ satisfies (T3), this implies that $P \subseteq T(\mathcal{P})$.

On the other hand, if $e \in T(\mathcal{P})$, then $\delta^T_{\mathcal{P}}(x^0_{\mathcal{P}}, e)$ is nonempty, so there exists $f \in P_{\text{sat}}$ which dominates e. Since P satisfies (T3), this implies that $e \in P$, so $T(\mathcal{P}) \subseteq P$, which completes the proof. □

The following result is an immediate consequence of the proof of Proposition 2.

COROLLARY 1. *If P is a trajectory model with canonical* NSM *$\mathcal{P}$, then for each $e \in P$, $\delta^T_{\mathcal{P}}(x^0_{\mathcal{P}}, e) = \{f \in P_{sat} |\, e \sqsubseteq f\}$.*

The following result is an immediate consequence of Propositions 1 and 2.

THEOREM 1. *Let $P \subseteq 2^\Sigma \times (\Sigma \times 2^\Sigma)^*$. Then P is the trajectory model of a nondeterministic state machine (with ϵ-moves) if and only if P satisfies properties* (T1)–(T5).

3.4. Deterministic trajectory models. Recall that a state machine $\mathcal{P}$ is deterministic if and only if its transition function is a partial map $\delta_{\mathcal{P}} : X_{\mathcal{P}} \times \Sigma \to X_{\mathcal{P}}$, i.e., there are no ϵ-transitions and $\delta_{\mathcal{P}}(x, \sigma)$ is either empty or contains exactly one element.

DEFINITION 6. *$P \subseteq 2^\Sigma \times (\Sigma \times 2^\Sigma)^*$ is called a* deterministic trajectory model *if and only if there exists a deterministic state machine $\mathcal{P}$ such that $T(\mathcal{P}) = P$.*

For any NSM $\mathcal{P}$, the language model can be obtained from the trajectory model via the trace map defined below. In the special case when the system $\mathcal{P}$ is deterministic, the trajectory model can be recovered from the language model via the inverse operation of the trace map, called the trajectory map, also defined below. Consequently, for deterministic systems, the language model is equivalent in detail of description to the trajectory model.

DEFINITION 7. *The* trace *map from refusal-traces to traces, denoted $tr : 2^\Sigma \times (\Sigma \times 2^\Sigma)^* \to \Sigma^*$, is defined inductively on the length of the refusal-traces as*:

- $\forall \Sigma' \subseteq \Sigma : tr(\Sigma') := \epsilon$,
- $\forall e \in 2^\Sigma \times (\Sigma \times 2^\Sigma)^*, \sigma \in \Sigma, \Sigma' \subseteq \Sigma : tr(e(\sigma, \Sigma')) := tr(e)\sigma$.

It is clear that $tr(T(\mathcal{P})) = L(\mathcal{P})$, where the trace operator is extended to the set of trajectory models in the natural way. Given a trajectory model $P \subseteq 2^\Sigma \times (\Sigma \times 2^\Sigma)^*$, we use $L(P) := tr(P)$ to denote the language model associated with P.

DEFINITION 8. *Let $K \subseteq \Sigma^*$ be a nonempty prefix-closed language. The* trajectory *map from traces to refusal-traces for the language model K, denoted $trj_K : K \to 2^\Sigma \times (\Sigma \times 2^\Sigma)^*$, is defined inductively on the length of traces of K as*:

- $trj_K(\epsilon) := \{\sigma \in \Sigma \mid \sigma \notin K\}$,
- $\forall s \in K, \sigma \in \Sigma$ *s.t.* $s\sigma \in K : trj_K(s\sigma) := trj_K(s)(\sigma, \{\sigma' \in \Sigma \mid s\sigma\sigma' \notin K\})$.

LEMMA 2. *Let $\mathcal{P}$ be an* NSM *with language model $K := L(\mathcal{P})$. Then*

(1) *$trj_K(K) \subseteq T(\mathcal{P})$,*

(2) *If $\mathcal{P}$ is deterministic, then $trj_K(K) = (T(\mathcal{P}))_{\text{sat}}$.*

Proof. Let $s \in K$ be a trace of length r. If $r = 0$, then $s = \epsilon$; otherwise, let $s = \sigma_1\sigma_2, \ldots, \sigma_r$. Let s^i denote the length-i prefix of s, and define $\hat{\Sigma}_i = \{\sigma \in \Sigma |\, s^i\sigma \notin K\}$.

Set

$$e := trj_K(s) = \hat{\Sigma}_0(\sigma_1, \hat{\Sigma}_1) \cdots (\sigma_r, \hat{\Sigma}_r).$$

Since $s \in L(\mathcal{P}) = L(T(\mathcal{P}))$, it follows from (T3) that the refusal-trace $\emptyset(\sigma_1, \emptyset) \cdots (\sigma_r, \emptyset) \in T(\mathcal{P})$. By repeated application of (T4), this implies that $e \in T(\mathcal{P})$, proving the first part.

Now assume that $\mathcal{P}$ is deterministic. To prove the second part, it suffices to show that e is the unique refusal-trace in $(T(\mathcal{P}))_{\text{sat}}$ with $tr(e) = s$. Also, since there must exist a saturated refusal-trace with trace s, it suffices to show that if $f \in (T(\mathcal{P}))_{\text{sat}}$ with $tr(f) = s$, then $f = e$. We use induction on $r = |e|$ to prove this together with the assertion that

$$\delta^T_{\mathcal{P}}(x^0_{\mathcal{P}}, e) = \delta^*_{\mathcal{P}}(x^0_{\mathcal{P}}, tr(e)). \tag{3}$$

Note that since $\mathcal{P}$ is deterministic, $\epsilon^*_{\mathcal{P}}(x) = x$, and given any $s \in K$, there exists a unique $x_s \in \delta^*_{\mathcal{P}}(x^0_{\mathcal{P}}, s)$. Furthermore, $\Re_{\mathcal{P}}(x_s) = \{\sigma \in \Sigma \mid s\sigma \notin K\}$.

If $r = 0$, then $f = \Sigma_0$ with $\Sigma_0 \subseteq \Re_{\mathcal{P}}(x^0_{\mathcal{P}}) = \hat{\Sigma}_0$. Since f is saturated, $\hat{\Sigma}_0 \subseteq \Sigma_0$, so $f = e$. Also, $\delta^T_{\mathcal{P}}(x^0_{\mathcal{P}}, e) = x^0_{\mathcal{P}} = \delta^*_{\mathcal{P}}(x^0_{\mathcal{P}}, \epsilon) = \delta^*_{\mathcal{P}}(x^0_{\mathcal{P}}, tr(e))$ as required.

For the induction step, express e and f as $e = \bar{e}(\sigma_r, \hat{\Sigma}_r)$, $f = \bar{f}(\sigma_r, \Sigma_r)$. Since the prefix of a saturated refusal-trace is saturated, $\bar{f} \in (T(\mathcal{P}))_{\text{sat}}$. Therefore, by induction hypothesis, we may assume that $\bar{f} = \bar{e}$. Using (3) applied to $\bar{e}$, it follows that

$$\begin{align}
\delta^T_{\mathcal{P}}(x^0_{\mathcal{P}}, f) &= \{x \in \epsilon^*_{\mathcal{P}}(\delta_{\mathcal{P}}(\delta^T_{\mathcal{P}}(x^0_{\mathcal{P}}, \bar{e}), \sigma_r)) \mid \Sigma_r \subseteq \Re_{\mathcal{P}}(x)\} \tag{4}\\
&= \{x \in \delta_{\mathcal{P}}(\delta^*_{\mathcal{P}}(x^0_{\mathcal{P}}, tr(\bar{e})), \sigma_r) \mid \Sigma_r \subseteq \Re_{\mathcal{P}}(x)\} \tag{5}\\
&= \{x \in \delta^*_{\mathcal{P}}(x^0_{\mathcal{P}}, tr(e)) \mid \Sigma_r \subseteq \Re_{\mathcal{P}}(x)\} \tag{6}\\
&= \begin{cases} x_s & \text{if } \Sigma_r \subseteq \Re_{\mathcal{P}}(x_s), \\ \emptyset & \text{otherwise.} \end{cases} \tag{7}
\end{align}$$

Since f is a refusal-trace of $\mathcal{P}$, $\delta^T_{\mathcal{P}}(x^0_{\mathcal{P}}, f)$ is nonempty, so $\Sigma_r \subseteq \Re_{\mathcal{P}}(x_s) = \hat{\Sigma}_r$. Since f is saturated, $\hat{\Sigma}_r \subseteq \Sigma_r$, so $f = e$. Also, by replacing f by e and Σ_r by $\hat{\Sigma}_r$ in equalities (4)–(7), we get $\delta^T_{\mathcal{P}}(x^0_{\mathcal{P}}, e) = x_s = \delta^*_{\mathcal{P}}(x^0_{\mathcal{P}}, tr(e))$. This completes the induction step. □

PROPOSITION 3. *Let $K \subseteq \Sigma^*$ be a nonempty prefixed-closed language, and let* $\det(K) := \operatorname{dom}(trj_K(K))$. *Then*

(1) $\det(K)$ *is a deterministic trajectory model.*

(2) *If P is any trajectory model with $L(P) = K$, then* $\det(K) \subseteq P$, *with equality if and only if P is deterministic.*

Proof. By a standard result, there exists a deterministic state machine $\mathcal{Q}$ such that $L(\mathcal{Q}) = K$. Setting $Q = T(\mathcal{Q})$ gives $L(Q) = K$. Since $\mathcal{Q}$ is deterministic, it follows from Lemma 2 that $trj_K(K) = (T(\mathcal{Q}))_{\text{sat}}$, which implies that $\det(K) = \operatorname{dom}((T(\mathcal{Q}))_{\text{sat}}) = Q$. Thus, $\det(K)$ is a deterministic trajectory model.

Let P be any trajectory model with $L(P) = K$. Then there exists a state machine $\mathcal{P}$ such that $T(\mathcal{P}) = P$. By Lemma 2, $trj_K(K) \subseteq P$, so $\det(K) \subseteq P$. If P is deterministic, then we can take $\mathcal{P}$ to be deterministic, so Lemma 2 implies that $trj_K(K) = (T(\mathcal{P}))_{\text{sat}}$, and hence $\det(K) = P$. On the other hand, if $\det(K) = P$, then P is deterministic by the first part. □

Remark 4. It follows from Proposition 3 that, given a nonempty prefix-closed language K, there is a unique deterministic trajectory model with language K. Furthermore, this trajectory model $\det(K)$ can be constructed from K by applying the map $trj_K(\cdot)$ and taking dominance closure. This trajectory model is the unique minimal element (with respect to inclusion) of the family of trajectory models having language K.

4. Prioritized synchronous composition. In this section, we define the PSC of two NSMs (with ϵ-moves), which induces a PSC operation on trajectory models. We also prove that the trajectory modeling framework is a language congruence with respect to PSC. Our definition of the PSC of NSMs is more general than the one in [7], since the silent transitions, i.e., transitions labeled ϵ, were not included. As discussed in §1, a priority set is associated with a system. This means that for an event which belongs to the priority set of a system to occur in the PSC with another system, the former system must participate.

DEFINITION 9. *Let $\mathcal{P} = (X_\mathcal{P}, \Sigma, \delta_\mathcal{P}, x^0_\mathcal{P})$ and $\mathcal{Q} = (X_\mathcal{Q}, \Sigma, \delta_\mathcal{Q}, x^0_\mathcal{Q})$ be two* NSM*s (with ϵ-moves). Let $A, B \subseteq \Sigma$ be the priority sets of $\mathcal{P}, \mathcal{Q}$, respectively. Then the* PSC *of $\mathcal{P}$ and $\mathcal{Q}$, denoted $\mathcal{P}\ {}_A\|_B\ \mathcal{Q}$, is another* NSM *defined as: $\mathcal{P}\ {}_A\|_B\ \mathcal{Q} := \mathcal{R} := (X_\mathcal{R}, \Sigma, \delta_\mathcal{R}, x^0_\mathcal{R})$, where $X_\mathcal{R} := X_\mathcal{P} \times X_\mathcal{Q}, x^0_\mathcal{R} := (x^0_\mathcal{P}, x^0_\mathcal{Q})$, and the transition function $\delta_\mathcal{R} : X_\mathcal{R} \times \Sigma \cup \{\epsilon\} \to 2^{X_\mathcal{R}}$ is defined as:*

- $\forall x_r = (x_p, x_q) \in X_\mathcal{R}$,

$$\forall \sigma \in \Sigma : \delta_\mathcal{R}(x_r, \sigma) := \begin{cases} \delta_\mathcal{P}(x_p, \sigma) \times \delta_\mathcal{Q}(x_q, \sigma) & \textit{if } \delta_\mathcal{P}(x_p, \sigma) \neq \emptyset, \delta_\mathcal{Q}(x_q, \sigma) \neq \emptyset, \\ \delta_\mathcal{P}(x_p, \sigma) \times \{x_q\} & \textit{if } \delta_\mathcal{P}(x_p, \sigma) \neq \emptyset, \sigma \in \Re_\mathcal{Q}(x_q), \sigma \notin B, \\ \{x_p\} \times \delta_\mathcal{Q}(x_q, \sigma) & \textit{if } \delta_\mathcal{Q}(x_q, \sigma) \neq \emptyset, \sigma \in \Re_\mathcal{P}(x_p), \sigma \notin A, \\ \emptyset & \textit{otherwise,} \end{cases}$$

$$\delta_\mathcal{R}(x_r, \epsilon) := [\delta_\mathcal{P}(x_p, \epsilon) \cup \{x_p\}] \times [\delta_\mathcal{Q}(x_q, \epsilon) \cup \{x_q\}] - \{(x_p, x_q)\}.$$

Thus, if an event is executable in the current states of both $\mathcal{P}$ and $\mathcal{Q}$, then it can be executed in $\mathcal{R}$, in which case both $\mathcal{P}$ and $\mathcal{Q}$ change their states synchronously according to their respective transitions. An event can be executed asynchronously by one of the systems if it is executable by that system and is not in the priority set of and cannot be executed in any state in the ϵ-closure of the current state of the other system. In this case, a state transition occurs in one system while no state change occurs in the other system. The silent transitions, i.e., those labeled by ϵ, can occur either synchronously or asynchronously. It is clear that an event in $A \cap B$ occurs only synchronously. Such synchronous execution is not required for events that do not belong to $A \cap B$. However, if an event that does not belong to $A \cap B$ is defined at states $x_p \in X_\mathcal{P}$ and $x_q \in X_\mathcal{Q}$, then it occurs synchronously at state $x_r = (x_p, x_q) \in X_\mathcal{R}$. Synchronous execution of such events is called *broadcast synchronization.*

Remark 5. If $A = B = \Sigma$, then an event is executable in the composed system if and only if it is executable in both systems. Thus this case corresponds to SSC. In contrast, if $A = B = \emptyset$, then an event is executable in the composed system if and only if it is executable in either of the systems.[5] This corresponds to an interleaving composition of the systems modified by the requirement that events which are executable by both systems are executed synchronously.

If $\mathcal{P}$ represents an uncontrolled plant, $\mathcal{Q}$ a supervisor, and $\mathcal{P}\ {}_A\|_B\ \mathcal{Q}$ the controlled plant or the closed-loop system, then (i) $A \cap B$ is the set of strict synchronization events and can be used to represent the set of *controllable* events; (ii) $A - B$ is the

[5] If an event is executable in the current state of one system but not in the current state of the other system, yet is executable in the second system following a silent transition, the event cannot occur in the composed system until the silent transition has occurred.

set of priority events only of $\mathcal{P}$ and can be used to represent the set of *uncontrollable* events; (iii) $B - A$ is the set of priority events only of $\mathcal{Q}$ and can be used to represent the set of *driven* events; (iv) $\Sigma - (A \cup B)$ is assumed to be empty, because events in $\Sigma - A \cup B$ belong neither to the priority set of the plant nor to that of the supervisor.

To simplify future notation, we define for any sets $\Sigma', \Sigma_1, \Sigma_2, \Sigma'' \subseteq \Sigma$:

- $\Sigma' \,_{\Sigma_1}\!\bigotimes_{\Sigma_2} \Sigma'' := (\Sigma' \cap \Sigma'') \cup (\Sigma' \cap \Sigma_1) \cup (\Sigma'' \cap \Sigma_2)$.

The following lemma gives two useful properties of the PSC of NSMs. It is a straightforward consequence of the definition of PSC.

LEMMA 3. *If* $\mathcal{R} = \mathcal{P} \,_A\|_B\, \mathcal{Q}$ *and* $x_r = (x_p, x_q) \in X_{\mathcal{R}}$*, then*

(1) $\epsilon^*_{\mathcal{R}}(x_r) = \epsilon^*_{\mathcal{P}}(x_p) \times \epsilon^*_{\mathcal{Q}}(x_q)$,

(2) $\Re_{\mathcal{R}}(x_r) = \Re_{\mathcal{P}}(x_p) \,_A\!\bigotimes_B \Re_{\mathcal{Q}}(x_q)$.

In other words, a state $x'_r = (x'_p, x'_q) \in X_{\mathcal{R}}$ belongs to the ϵ-closure of $x_r = (x_p, x_q)$ if and only if x'_p (respectively, x'_q) belongs to ϵ-closure of x_p (respectively, x_q). Also, an event is refused in $\mathcal{P} \,_A\|_B\, \mathcal{Q}$ if and only if either it is refused in both $\mathcal{P}$ and $\mathcal{Q}$, or it belongs to the priority set of $\mathcal{P}$ and is refused in $\mathcal{P}$, or it belongs to the priority set of $\mathcal{Q}$ and is refused in $\mathcal{Q}$.

We next consider the trajectory model of the PSC of two systems, and obtain its relationship to the trajectory models of the component systems. Using the definition of $\mathcal{P} \,_A\|_B\, \mathcal{Q}$ and that of its refusal map $\Re_{\mathcal{P} \,_A\|_B\, \mathcal{Q}}$, the trajectory model $T(\mathcal{P} \,_A\|_B\, \mathcal{Q})$ is easily obtained from its definition developed in the previous section. To obtain the relationship between $T(\mathcal{P})$, $T(\mathcal{Q})$, and $T(\mathcal{P} \,_A\|_B\, \mathcal{Q})$, we first define the PSC of a pair of refusal-traces.

DEFINITION 10. *Let* $e_p \in T(\mathcal{P})$ *and* $e_q \in T(\mathcal{Q})$*. Then the* PSC *of* e_p *and* e_q *(with respect to* $T(\mathcal{P})$ *and* $T(\mathcal{Q})$*), denoted* $e_p \,_A\|_B\, e_q$*, is defined inductively on* $|e_p| + |e_q|$ *as follows:*

- $\forall \Sigma_p, \Sigma_q \subseteq \Sigma$ *s.t.* $\Sigma_p \in T(\mathcal{P}), \Sigma_q \in T(\mathcal{Q})$:

$$\Sigma_p \,_A\|_B\, \Sigma_q := \{\Sigma' \subseteq \Sigma_p \,_A\!\bigotimes_B \Sigma_q\},$$

- $\forall e_p \in T(\mathcal{P}); e_q \in T(\mathcal{Q}); \sigma_p, \sigma_q \in \Sigma; \Sigma_p, \Sigma_q \subseteq \Sigma$ *s.t.* $e_p(\sigma_p, \Sigma_p) \in T(\mathcal{P}), e_q(\sigma_q, \Sigma_q) \in T(\mathcal{Q})$:

$$e_p(\sigma_p, \Sigma_p) \,_A\|_B\, e_q(\sigma_q, \Sigma_q) := T_1 \cup T_2 \cup T_3,$$

where

$$T_1 := \begin{cases} \{e(\sigma_p, \Sigma') \mid e \in e_p \,_A\|_B\, e_q(\sigma_q, \Sigma_q); \Sigma' \subseteq \Sigma_p \,_A\!\bigotimes_B \Sigma_q\} & \textit{if } \sigma_p \notin B \textit{ and } \\ & \quad e_q(\sigma_q, \Sigma_q)(\sigma_p, \emptyset) \notin T(\mathcal{Q}), \\ \emptyset & \textit{otherwise}; \end{cases}$$

$$T_2 := \begin{cases} \{e(\sigma_q, \Sigma') \mid e \in e_p(\sigma_p, \Sigma_p) \,_A\|_B\, e_q; \Sigma' \subseteq \Sigma_p \,_A\!\bigotimes_B \Sigma_q\} & \textit{if } \sigma_q \notin A \textit{ and } \\ & \quad e_p(\sigma_p, \Sigma_p)(\sigma_q, \emptyset) \notin T(\mathcal{P}), \\ \emptyset & \textit{otherwise}; \end{cases}$$

$$T_3 := \begin{cases} \{e(\sigma, \Sigma') \mid e \in e_p \,_A\|_B\, e_q; \Sigma' \subseteq \Sigma_p \,_A\!\bigotimes_B \Sigma_q\} & \textit{if } \sigma_p = \sigma_q := \sigma, \\ \emptyset & \textit{otherwise}. \end{cases}$$

It should be noted that $e_p \,_A\|_B\, e_q$ is a set of refusal-traces that depends on $T(\mathcal{P}), T(\mathcal{Q})$ as well as on the particular refusal-traces e_p, e_q. The dependence on $T(\mathcal{P}), T(\mathcal{Q})$ is not explicitly indicated in the notation.

The PSC of two zero-length refusal-traces $\Sigma_p \in T(\mathcal{P})$ and $\Sigma_q \in T(\mathcal{Q})$, which correspond to initial refusal sets of $T(\mathcal{P})$ and $T(\mathcal{Q})$, respectively, is obtained by computing $\Sigma_p \,{}_A\!\bigotimes_B\, \Sigma_q$, which corresponds to an initial refusal set of $T(\mathcal{P} \,{}_A\|_B\, \mathcal{Q})$. Next, the PSC of two refusal-traces $e_p(\sigma_p, \Sigma_p) \in T(\mathcal{P})$ and $e_q(\sigma_q, \Sigma_q) \in T(\mathcal{Q})$ is obtained by considering these three possible cases: (i) a refusal-trace belonging to $e_p \,{}_A\|_B\, e_q(\sigma_q, \Sigma_q)$ has already been executed in the composed system, and at this point σ_p is executable in $\mathcal{P}$ (indicated by $e_p(\sigma_p, \Sigma_p) \in T(\mathcal{P})$), the occurrence of σ_p cannot be blocked by $\mathcal{Q}$ (indicated by $\sigma_p \notin B$), and $\mathcal{Q}$ cannot participate in the occurrence of σ_p (indicated by $e_q(\sigma_q, \Sigma_q)(\sigma_p, \emptyset) \notin T(\mathcal{Q})$); (ii) a refusal-trace belonging to $e_p(\sigma_p, \Sigma_p) \,{}_A\|_B\, e_q$ has already been executed in the composed system, and at this point σ_q is executable in $\mathcal{Q}$, and $\mathcal{P}$ can neither block the occurrence of σ_q, nor can it participate in the occurrence of σ_q; (iii) $\sigma_p = \sigma_q := \sigma$; a refusal-trace belonging to $e_p \,{}_A\|_B\, e_q$ has already been executed in the composed system, and at this point σ is executable in both $\mathcal{P}$ and $\mathcal{Q}$.

Remark 6. It is clear from Definition 10 that if $A = B = \Sigma$, which corresponds to the case of SSC, then the sets $T_1 = T_2 = \emptyset$ since the conditions "$\sigma_p \notin B$" and "$\sigma_q \notin A$" both evaluate to "false." Hence the PSC of $e_p(\sigma_p, \Sigma_p) \in T(\mathcal{P})$ and $e_q(\sigma_q, \Sigma_q) \in T(\mathcal{Q})$ is nonempty if and only if the set T_3 is nonempty, which requires that $\sigma_p = \sigma_q$. Using induction, it can be easily concluded that the SSC of refusal-traces $e_p \in T(\mathcal{P})$ and $e_q \in T(\mathcal{Q})$ is a nonempty set if and only if $tr(e_p) = tr(e_q)$, in which case, $tr(e_p \,{}_\Sigma\|_\Sigma\, e_q) = tr(e_p) = tr(e_q)$, and for each $i \leq |e_p| = |e_q|$, the ith refusal set of any trace in $e_p \,{}_\Sigma\|_\Sigma\, e_q$ is any subset of the union of the ith refusal set of e_p and the ith refusal set of e_q, since $\Sigma_i(e_p) \,{}_\Sigma\!\bigotimes_\Sigma\, \Sigma_i(e_q) = \Sigma_i(e_p) \cup \Sigma_i(e_q)$.

We can extend the definition of the PSC of a pair of refusal-traces to the PSC of the trajectory models. With a slight abuse of notation, we use the same symbol ${}_A\|_B$ for the PSC of the NSMs $\mathcal{P}, \mathcal{Q}$ and for the PSC of their corresponding trajectory models $T(\mathcal{P}), T(\mathcal{Q})$.

DEFINITION 11. *The* PSC *of the trajectory models* $T(\mathcal{P}), T(\mathcal{Q})$ *is defined to be*

- $T(\mathcal{P}) \,{}_A\|_B\, T(\mathcal{Q}) := \bigcup_{e_p \in T(\mathcal{P}), e_q \in T(\mathcal{Q})} e_p \,{}_A\|_B\, e_q$.

The following result shows that the trajectory model of the PSC of NSMs is the PSC of their corresponding trajectory models. Equivalently, it states that the PSC operation on NSMs induces a PSC operation on trajectory models, and the induced operation is precisely the one described in Definition 11.

THEOREM 2. *For any* NSM*s* $\mathcal{P}, \mathcal{Q}$, $T(\mathcal{P} \,{}_A\|_B\, \mathcal{Q}) = T(\mathcal{P}) \,{}_A\|_B\, T(\mathcal{Q})$.

Proof. Refer to Appendix A. □

COROLLARY 2. *The trajectory model is a language congruence with respect to the operation of* PSC.

Proof. Let $\mathcal{P}_1, \mathcal{P}_2, \mathcal{Q}_1, \mathcal{Q}_2$ be NSMs with $T(\mathcal{P}_1) = T(\mathcal{P}_2)$ and $T(\mathcal{Q}_1) = T(\mathcal{Q}_2)$. Theorem 2 implies that $T(\mathcal{P}_1 \,{}_A\|_B\, \mathcal{Q}_1) = T(\mathcal{P}_2 \,{}_A\|_B\, \mathcal{Q}_2)$. Hence $L(\mathcal{P}_1 \,{}_A\|_B\, \mathcal{Q}_1) = L(T(\mathcal{P}_1 \,{}_A\|_B\, \mathcal{Q}_1)) = L(T(\mathcal{P}_2 \,{}_A\|_B\, \mathcal{Q}_2)) = L(\mathcal{P}_2 \,{}_A\|_B\, \mathcal{Q}_2)$. □

We will need the following result which shows that PSC of trajectory models preserves determinism.

COROLLARY 3. *If* P *and* Q *are deterministic trajectory models, then so is* $P \,{}_A\|_B\, Q$.

Proof. By definition, there exist deterministic state machines $\mathcal{P}, \mathcal{Q}$ such that $T(\mathcal{P}) = P$, $T(\mathcal{Q}) = Q$. From Definition 9, it is clear that $\mathcal{P} \,{}_A\|_B\, \mathcal{Q}$ is deterministic. Since Theorem 2 implies that $P \,{}_A\|_B\, Q = T(\mathcal{P} \,{}_A\|_B\, \mathcal{Q})$, we conclude that $P \,{}_A\|_B\, Q$ is deterministic. □

Remark 7. Theorem 2 shows that the trajectory model of $\mathcal{P} \,{}_A\|_B\, \mathcal{Q}$ can be described using only $T(\mathcal{P})$ and $T(\mathcal{Q})$, and not $\mathcal{P}, \mathcal{Q}$ directly. This is in contrast to the situation with the failures model. Theorem 2 and Corollary 2 both fail if the trajectory model is replaced with the failures model. The equality of failures models does not necessarily imply the equality of failures models, or even language models, under prioritized synchronous composition with a fixed system [7, Ex. 7]. The result in Corollary 2 was mentioned without proof in [7], [8]. However, its rigorous demonstration depends on the precise definitions given above for the PSC of NSMs (with ϵ-moves) as well as for the trajectory model of an NSM.

5. Properties of prioritized synchronous composition. In this section we describe some of the properties of the PSC of two or more trajectory models, which are used in §6 for the synthesis of supervisors which control the behavior of nondeterministic plants via PSC.

5.1. Associativity. We begin by providing a proof for the following result which is stated without proof as part of [8, Thm. 13.4].

THEOREM 3. *For any trajectory models P, Q, R and priority sets $A, B, C \subseteq \Sigma$*

$$(P\, {}_A\|_B\, Q)\, {}_{A\cup B}\|_C\, R = P\, {}_A\|_{B\cup C}\, (Q\, {}_B\|_C\, R).$$

Proof. Refer to Appendix A. □

This can be interpreted as an associative property as follows. Let P, Q denote trajectory models with event set Σ, and let A, B be subsets of Σ. We refer to the pairs (P, A), (Q, B) as *prioritized systems*, and define their *synchronous composition* to be the prioritized system

- $(P, A) \parallel (Q, B) := (P\, {}_A\|_B\, Q,\ A \cup B)$.

Then Theorem 3 asserts that $[(P, A) \parallel (Q, B)] \parallel (R, C) = (P, A) \parallel [(Q, B) \parallel (R, C)]$. Thus, the result is simply the associative property for the synchronous composition of prioritized systems.

5.2. Augmentation and prioritized synchronous composition. We define augmentation of both NSMs and trajectory models, and show that the prioritized synchronous composition of two trajectory models is identical to *strict* synchronous composition of their augmentations, provided the two priority sets exhaust the set of events.

Let $\mathcal{P}$ be an NSM with event set Σ, and let $D \subseteq \Sigma$. We denote by $\mathcal{D}$ the deterministic state machine with one state and self-loops labeled by every event in D. The *augmentation of $\mathcal{P}$ by D*, denoted $\mathcal{P}^D$, is defined to be the NSM $\mathcal{P}^D := \mathcal{P}\, {}_\emptyset\|_\emptyset\, \mathcal{D}$. The state space of $\mathcal{P}^D$ can be identified with the state space of $\mathcal{P}$, and $\mathcal{P}^D$ is then obtained from $\mathcal{P}$ by adding self-loops at each $x \in X_{\mathcal{P}}$ labeled by every event in $D \cap \Re_{\mathcal{P}}(x)$. It is clear that $\mathcal{P}^D$ is deterministic whenever $\mathcal{P}$ is deterministic.

If P is a trajectory model, the *augmentation of P by D*, denoted P^D, is defined to be the trajectory model $P^D := P\, {}_\emptyset\|_\emptyset\, \det(D^*)$. Note that since both priority sets are empty, P^D represents interleaving of P and $\det(D^*)$ except that the broadcast synchronization requirement means that events in D which can also occur in P occur synchronously in both P and $\det(D^*)$.

Remark 8. Since $\det(D^*)$ can always execute every event in D and can never execute any event in $\Sigma - D$, it follows that for any $A \subseteq \Sigma - D$ and any $B \subseteq D$,

$$P^D := P\, {}_\emptyset\|_\emptyset\, \det(D^*) = P\, {}_A\|_B\, \det(D^*).$$

It follows from Theorem 2 that given an NSM $\mathcal{P}$ and an event set $D \subseteq \Sigma$, $T(\mathcal{P}^D) = [T(\mathcal{P})]^D$, and it follows from Corollary 3 that if P is a deterministic trajectory model, then so is its augmentation P^D.

The following result shows that augmentation can be used to reduce prioritized synchronous composition to strict synchronization.

PROPOSITION 4. *If* $A \cup B = \Sigma$, *then* $P\ {}_A\|_B\ Q = P^{B-A}\ {}_\Sigma\|_B\ Q = P^{B-A}\ {}_\Sigma\|_\Sigma\ Q^{A-B}$.

Proof. It suffices to prove the first equality since the second equality follows from symmetry and a second application of the first equality. Using Remark 8 and Theorem 3 gives

$$\begin{aligned} P^{B-A}\ {}_\Sigma\|_B\ Q &= (P\ {}_A\|_{B-A}\ \det((B-A)^*))\ {}_\Sigma\|_B\ Q \\ &= \det((B-A)^*)\ {}_{B-A}\|_\Sigma\ (P\ {}_A\|_B\ Q) \\ &= P\ {}_A\|_B\ Q. \end{aligned}$$

The final equality is an easy consequence of two facts: the priority set of $P\ {}_A\|_B\ Q$ is Σ, so $\det((B-A)^*)$ cannot execute any events which do not occur in $P\ {}_A\|_B\ Q$; $\det((B-A)^*)$ can alway execute each event in its priority set, so it cannot block any events in $P\ {}_A\|_B\ Q$. □

6. Supervisory control with driven events.

In this section, we derive results concerning supervisory control by prioritized synchronous composition in the presence of driven events.

6.1. Control under complete observability.

We begin with a result which shows that in a prioritized synchronous composition, a deterministic system participates in every event of any refusal-trace whose trace belongs to its language.

LEMMA 4. *Let* P, Q *be trajectory models with* Q *deterministic. If* $e \in e_p\ {}_A\|_B\ e_q \subseteq P\ {}_A\|_B\ Q$ *with* $tr(e) \in L(Q)$, *then* $tr(e) = tr(e_q)$.

Proof. The result follows as a special case of Lemma 5 below. □

The following result gives necessary and sufficient conditions for a given (prefix-closed) language to be realizable as the closed-loop language for a plant supervised by prioritized synchronous composition. The basic assumption is that every event in Σ belongs to the priority set A of the plant P or the priority set B of the supervisor. The interpretation is that Σ is partitioned into disjoint subsets Σ_c, Σ_u and Σ_d consisting of the controllable, uncontrollable, and driven events, and $A = \Sigma_c \cup \Sigma_u$ while $B = \Sigma_c \cup \Sigma_d$.

THEOREM 4. *Let* P *be a trajectory model,* $A \cup B = \Sigma$, *and let* K *be a nonempty prefix-closed sublanguage of* $L(P^{B-A})$. *Then there exists a trajectory model* S *such that* $L(P\ {}_A\|_B\ S) = K$ *if and only if*

$$K(A-B) \cap L(P^{B-A}) \subseteq K, \tag{8}$$

in which case S *can be chosen to be the deterministic trajectory model* $\det(K)$.

Proof. We begin with sufficiency. Suppose that equation (8) holds. Since K is a nonempty prefix-closed sublanguage of $L(P^{B-A})$, there exists a trajectory model S such that $L(P^{B-A}) \cap L(S) = K$. Without loss of generality, we may assume that S is deterministic. (In particular, we can choose $S = \det(K)$.)

We claim that

$$L(P^{B-A}) \cap L(S^{A-B}) = K. \tag{9}$$

Obviously, $K = L(P^{B-A}) \cap L(S) \subseteq L(P^{B-A}) \cap L(S^{A-B})$. We establish the reverse inclusion by contradiction. Suppose $L(P^{B-A}) \cap L(S^{A-B})$ strictly contains K. Let $t = s\sigma$ be a minimal length trace in $L(P^{B-A}) \cap L(S^{A-B}) - K$. Then $s \in K = L(P^{B-A}) \cap L(S)$. Since $s\sigma \in L(S^{A-B})$, there exists $g = \bar{g}(\sigma, \emptyset) \in S^{A-B}$ such that $tr(\bar{g}) = s$. Hence, there exist $e = \bar{e}(\sigma', \Sigma') \in S$, $f = \bar{f}(\sigma'', \Sigma'') \in \det((A-B)^*)$

such that $g \in e\,{}_{\emptyset}\|_{\emptyset}\, f$. First suppose $\sigma \notin A - B$. Then $\sigma \neq \sigma''$, so $\sigma = \sigma'$ and $\bar{g} \in \bar{e}\,{}_{\emptyset}\|_{\emptyset}\, f$. Since S is deterministic and $tr(\bar{g}) = s \in L(S)$, it follows from Lemma 4 that $tr(\bar{g}) = tr(\bar{e})$. Thus, $s\sigma = tr(g) = tr(e) \in L(S)$, which implies that $t \in K$, a contradiction. On the other hand, if $\sigma \in A - B$, then it follows from (8) that $t \in K$, again a contradiction. Thus, (9) holds.

Using Proposition 4, we get $K = L(P^{B-A}) \cap L(S^{A-B}) = L(P^{B-A}\,{}_{\Sigma}\|_{\Sigma}\, S^{A-B}) = L(P\,{}_{A}\|_{B}\, S)$, showing that S solves the supervisory control problem.

Conversely, suppose there exists a trajectory model S such that $L(P\,{}_{A}\|_{B}\, S) = K$. Then (9) holds. Let $t = s\sigma \in K(A - B) \cap L(P^{B-A})$. Since $s \in K \subseteq L(S^{A-B})$ and $\sigma \in A - B$, it follows that $s\sigma \in L(S^{A-B})$. Thus, $t \in L(P^{B-A}) \cap L(S^{A-B}) = K$, so (8) holds. □

Remark 9. Theorem 4 states that K is realizable as the closed-loop language if and only if it is controllable (in the sense of Ramadge and Wonham [24]) *with respect to the language of the augmented plant,* $L(P^{B-A})$, which depends on the trajectory model P—not simply on $L(P)$. Knowledge of $L(P)$ is not sufficient to determine if the supervisory control problem is solvable for a given target language K. This is illustrated by the following example.

Example 3. We consider a very simple air traffic control problem. The plant represents the aircraft and pilot, while the supervisor represents the air traffic controller. Let $\Sigma = \{a, b\}$ where $a \in \Sigma_u$ represents a flight maneuver, while $b \in \Sigma_d$ represents a command from the tower not to execute the flight maneuver. The execution of b by the supervisor indicates that the command has been broadcasted, whereas the execution of b by the plant indicates that the command has been received.

We consider two alternative trajectory models for the plant:

$$P_1 = (a \rightarrow \Delta_\Sigma) + (b \rightarrow \Delta_\Sigma), \quad P_2 = P_1 \oplus (a \rightarrow \Delta_\Sigma).$$

NSMs with trajectory models P_1 and P_2 are depicted in Figs. 2(c) and (d), respectively. In P_1, the pilot can initially execute the maneuver or receive the command not to do so. However, in P_2 there is an initial nondeterministic choice between P_1 and the trajectory model $(a \rightarrow \Delta_\Sigma)$ in which the maneuver is possible but the command cannot be received. Thus, P_2 models the possibility of aircraft radio receiver failure. Note that $L(P_1) = L(P_2)$. However, it can be verified that $L(P_1^{B-A}) = (a + \epsilon)b^*$, while $L(P_2^{B-A}) = b^*(a + \epsilon)b^*$. Suppose that the target language K is not completely specified but is required to contain the trace b and not contain any trace in which the event a occurs after the event b has occurred. In other words, the tower should be initially able to broadcast the command b, and if the command has been broadcasted, the pilot must not be able to execute the maneuver a.

The supervisory control problem is clearly solvable for the plant model P_1. For example, if we choose $S = P_1$, then $P_1\,{}_{A}\|_{B}\, S = P_1$, so the closed-loop language is $L(P_1) = \{\epsilon, a, b\}$, which meets the specifications for K. On the other hand, the supervisory control problem is not solvable for the plant model P_2. For any target language K which satisfies the specifications, we have $ba \in K(A - B) \cap L(P_2^{B-A}) - K$. It follows from Theorem 4 that there is no supervisor S such that $L(P_2\,{}_{A}\|_{B}\, S) = K$.

It is worth noting that if P_2 is the correct plant model, i.e., receiver failure can occur, then the supervisory control problem can be made solvable by changing the protocol between the pilot and tower. If the pilot is required to obtain clearance from the tower in order to execute maneuver a, then a becomes a controllable event and it is then trivial to construct a supervisor that meets the specifications.

When there are no driven events, then $A = \Sigma_c \cup \Sigma_u = \Sigma$ and $B = \Sigma_c$. In this case Theorem 4 specializes to give the following corollary.

COROLLARY 4. *Let K be a nonempty prefix-closed sublanguage of $L(P)$. Then the following are equivalent:*

(1) *There exists a trajectory model S such that $L(P\,_{\Sigma}\|_{\Sigma_c}\,S) = K$.*

(2) $L(P\,_{\Sigma}\|_{\Sigma_c}\,\det(K)) = K$.

(3) $K\Sigma_u \cap L(P) \subseteq K$.

Remark 10. Corollary 4 shows that when there are no driven events, the necessary and sufficient conditions for supervisory control by prioritized synchronous composition are the same as those in the Ramadge–Wonham framework [23]. The equivalence of the first and second conditions of Corollary 4 was stated without proof in [7, Thm. 1] and [8, Thm. 14.2]. The equivalence of the first and third conditions of Corollary 4 was stated in [8, Thm. 14.1] accompanied by an incomplete proof.

Remark 11. The proof of Theorem 4 shows that if K satisfies condition (8) and N is any prefix-closed sublanguage of Σ^* with $L(P^{B-A}) \cap N = K$, then the deterministic supervisor $S := \det(N)$ results in K as the closed-loop language $L(P\,_A\|_B\,S)$. Since $K \subseteq N$, it follows from Lemma 4 that every event executed by the closed-loop system occurs in S. In particular, every uncontrollable event is executed by the supervisor even though such events do not belong to its priority set. This behavior is induced by the broadcast synchronization requirement in prioritized synchronous composition.

It is interesting to specialize this observation to the case where there are no driven events. Since $A = \Sigma$, the plant also participates in every event. Thus, the plant and supervisor function as though they are connected by *strict synchronization* rather than by prioritized synchronous composition. In particular, this is the case when the supervisor is chosen to be $\det(K)$. The determinism of S is essential here. If S is a nondeterministic trajectory model with $L(S) = N$, there is no guarantee that the closed-loop language will be K. This is demonstrated by the next example.

Example 4. Let $\Sigma = \{a,b\}$, $\Sigma_c = \{a\}$, $\Sigma_u = \{b\}$, $P = (a \to \Delta_\Sigma) + (b \to (a \to \Delta_\Sigma))$, $S = (a \to \Delta_\Sigma) \oplus (b \to \Delta_\Sigma)$. Then $L(P) = \{\epsilon, a, b, ba\}$, $L(S) = \{\epsilon, a, b\}$. Let $K = L(S)$. Then K satisfies the controllability condition (the third condition of Corollary 4) as well as $L(P) \cap L(S) = K$. A straightforward calculation shows that $P\,_{\Sigma}\|_{\Sigma_c}\,S = ((a \to \Delta_\Sigma) + (b \to (a \to \Delta_\Sigma))) \oplus (b \to \Delta_\Sigma)$. Thus, $L(P\,_{\Sigma}\|_{\Sigma_c}\,S) = \{\epsilon, a, b, ba\} = L(P) \neq K$. What happens is that since S is nondeterministic, the event b can be executed as the initial event solely in P even though $b \in L(S)$. (This cannot happen for deterministic S by Lemma 4.) Thus, strict synchronization is lost. This permits a trace of $P\,_{\Sigma}\|_{\Sigma_c}\,S$, which is not a trace of S.

6.2. Control under restricted unobservability. We continue to assume that $A \cup B = \Sigma$, where $A = \Sigma_c \cup \Sigma_u$ and $B = \Sigma_c \cup \Sigma_d$. In the closed-loop system $P\,_A\|_B\,S$, the events in $A - B$, i.e., the uncontrollable events, are generated by the plant P and are broadcast to the supervisor S, where they are synchronously executed whenever enabled. It may happen that information about the occurrence of certain uncontrollable events is unavailable for broadcasting due to lack of sensors, or it may be desirable to implement a simplified supervisor which ignores such information. This suggests a generalization of prioritized synchronous composition in which the broadcast synchronization requirement is disregarded for a specified subset $\Gamma \subseteq A - B$ of uncontrollable events. Since events in $A - B$ cannot occur spontaneously in S, this effectively prevents S from ever executing the events in Γ. Thus, instead of modifying the definition of prioritized synchronous composition, it is equivalent to restrict the admissible supervisors to those which do not execute events in Γ.

Let $\Pi_\Gamma : \Sigma \to \Sigma$ denote the natural projection defined by

- $\forall \sigma \in \Sigma : \Pi_\Gamma(\sigma) := \begin{cases} \epsilon & \text{for } \sigma \in \Gamma, \\ \sigma & \text{for } \sigma \in \Sigma - \Gamma. \end{cases}$

$\Pi_\Gamma(\cdot)$ extends to a map on Σ^* in the obvious way. We define the *restricted supervisory control problem* (RSCP) to be as follows: Given a prefix-closed sublanguage K of $L(P^{B-A})$ and $\Gamma \subseteq A - B$, determine if there exists a supervisor S such that

- $L(P\,{}_A\|_B\, S) = K$ and $\Pi_\Gamma(L(S)) = L(S)$.

Remark 12. There are two different ways to model an uncontrollable event in the plant which is unobservable to the supervisor. It can be completely suppressed and treated as an ϵ-event in P. Alternatively, it can be treated as a labeled event $\sigma \in \Sigma$ in the plant which does not label any transitions in the supervisor. The advantage of the second approach (which is the one taken in the RSCP) is that such an event can be included in the performance specifications, i.e., in the target language K. Hence, even though it is unobservable to the supervisor, its occurrence in the closed-loop system can be controlled—albeit subject to the conditions that must be satisfied by K for the solvability of the RSCP.

The next result generalizes Lemma 4 to the case where certain events in $A - B$ are not present in the second system Q.

LEMMA 5. *Let* $\Gamma \subseteq A-B$, *and* P, Q *be trajectory models with* Q *deterministic and satisfying* $\Pi_\Gamma(L(Q)) = L(Q)$. *If* $e \in e_p\,{}_A\|_B\, e_q \subseteq P\,{}_A\|_B\, Q$ *with* $\Pi_\Gamma(tr(e)) \in L(Q)$, *then* $\Pi_\Gamma(tr(e)) = tr(e_q)$.

Proof. The proof is by induction on $|e|$. The assertion holds trivially when $|e| = 0$. For the induction step, write $e = \bar{e}(\sigma, \Sigma')$ and let $\bar{e}_p$, $\bar{e}_q$ denote the prefixes of e_p, e_q obtained by deleting the final event and refusal set from each refusal-trace.

If σ occurs synchronously in both P and Q, then $\bar{e} \in \bar{e}_p\,{}_A\|_B\, \bar{e}_q$. Then $\sigma \notin \Gamma$, so $\Pi_\Gamma(tr(e)) = \Pi_\Gamma(tr(\bar{e}))\sigma$. Since $L(Q)$ is prefix-closed, $\Pi_\Gamma(tr(\bar{e})) \in L(Q)$. Applying the induction hypothesis gives $\Pi_\Gamma(tr(\bar{e})) = tr(\bar{e}_q)$. Thus, $\Pi_\Gamma(tr(e)) = \Pi_\Gamma(tr(\bar{e}))\sigma = tr(\bar{e}_q)\sigma = tr(e_q)$. The same argument applies in the case where $\bar{e} \in e_p\,{}_A\|_B\, \bar{e}_q$, i.e., when σ occurs only in Q.

Suppose $\bar{e} \in \bar{e}_p\,{}_A\|_B\, e_q$, i.e., σ occurs only in P. If $\sigma \in \Gamma$, then $\Pi_\Gamma(tr(e)) = \Pi_\Gamma(tr(\bar{e})) = tr(e_q)$, where the second equality follows from the induction hypothesis. Now suppose that $\sigma \notin \Gamma$. Since σ occurs only in P, it follows that $e_q(\sigma, \emptyset) \notin Q$. Since Q is deterministic, Proposition 3 implies that $tr(e_q)\sigma \notin L(Q)$. Since $L(Q)$ is prefix-closed, $\Pi_\Gamma(tr(\bar{e})) \in L(Q)$. Using the induction hypothesis, we have $tr(e_q)\sigma = \Pi_\Gamma(tr(\bar{e}))\sigma = \Pi_\Gamma(tr(e)) \in L(Q)$, a contradiction. Thus, this final case cannot occur. □

For the standard supervisory control problem with partial observations (and no driven events), a target language K is obtainable as the language of the closed-loop system if and only if K is controllable and observable relative to the language of the plant [16], [5]. The following result shows that in the presence of driven events, the RSCP is solvable if and only if K is controllable and observable *relative to the language of the augmented plant.*

THEOREM 5. *Let* $A \cup B = \Sigma$, $\Gamma \subseteq A - B$, *and let* K *be a nonempty prefix-closed sublanguage of* $L(P^{B-A})$. *Then there exists a trajectory model* S *such that*

$$(10) \qquad L(P\,{}_A\|_B\, S) = K, \qquad \Pi_\Gamma(L(S)) = L(S)$$

if and only if the following two conditions are satisfied:

$$(11) \qquad K(A-B) \cap L(P^{B-A}) \subseteq K,$$

$$(12) \qquad \forall \bar{s}, \bar{t} \in K, \sigma \in \Sigma : \Pi_\Gamma(\bar{s}) = \Pi_\Gamma(\bar{t}), \bar{s}\sigma \in K,\ \bar{t}\sigma \in L(P^{B-A}) \Rightarrow \bar{t}\sigma \in K.$$

In this case, S can be chosen to be the deterministic trajectory model $\det(\Pi_\Gamma(K))$.

Proof. We first show the necessity of the controllability condition (11) and observability condition (12). Suppose there exists a trajectory model S such that (10) is satisfied. Then (11) follows from Theorem 4. Let $\bar{s}$, $\bar{t} \in K$ with $\Pi_\Gamma(\bar{s}) = \Pi_\Gamma(\bar{t})$, and suppose that $\bar{s}\sigma \in K$, $\bar{t}\sigma \in L(P^{B-A})$. We need to show that $\bar{t}\sigma \in K$. Since $L(P^{B-A}) \cap L(S^{A-B}) = K$, it suffices to show that $\bar{t}\sigma \in L(S^{A-B})$. Since K is controllable, it suffices to consider $\sigma \in B$. Note that $S^{A-B} = (S^{A-B-\Gamma})^\Gamma$. Also, since $\Pi_\Gamma(L(S)) = L(S)$, events in Γ are never executed in S. Hence $\Pi_\Gamma(L(S^{A-B-\Gamma})) = L(S^{A-B-\Gamma})$. In other words, events in Γ are never executed in $S^{A-B-\Gamma}$. Hence the language $L(S^{A-B}) = L((S^{A-B-\Gamma})^\Gamma)$ is obtained by pure interleaving of the languages $L(S^{A-B-\Gamma})$ and $L(\det(\Gamma^*)) = \Gamma^*$. Since the trace $\bar{s}\sigma \in L(S^{A-B})$, we have $\Pi_\Gamma(\bar{s}\sigma) \in L(S^{A-B-\Gamma})$. Also, since $\Pi_\Gamma(\bar{s}\sigma) = \Pi_\Gamma(\bar{s})\sigma = \Pi_\Gamma(\bar{t})\sigma = \Pi_\Gamma(\bar{t}\sigma)$, we have $\Pi_\Gamma(\bar{t}\sigma) \in L(S^{A-B-\Gamma})$. Since $\bar{t}\sigma$ is a pure interleaving of $\Pi_\Gamma(\bar{t}\sigma) \in L(S^{A-B-\Gamma})$ and a trace in Γ^*, $\bar{t}\sigma \in L(S^{A-B})$. This establishes (12) and completes the proof of necessity.

To prove sufficiency, suppose that (11) and (12) both hold. Let $S = \det(\Pi_\Gamma(K))$. By Proposition 4, it is equivalent to prove $L(P^{B-A}\,{}_\Sigma\|_B\, S) = K$. Given any $t \in K$, there exists $e \in P^{B-A}$ with $tr(e) = t$ and there exists $f \in S$ with $tr(f) = \Pi_\Gamma(t)$. Since $\Gamma \cap B = \emptyset$ and S can never execute an event in Γ, it follows that $e\,{}_\Sigma\|_B\, f$ is nonempty and every refusal-trace which it contains has trace t. Thus, $K \subseteq L(P^{B-A}\,{}_\Sigma\|_B\, S)$.

It remains to prove

$$L(P^{B-A}\,{}_\Sigma\|_B\, S) \subseteq K. \tag{13}$$

We establish (13) by contradiction. Let $g = \bar{g}(\sigma, \Sigma') \in P^{B-A}\,{}_\Sigma\|_B\, S$ and suppose g has minimal length among the refusal-traces of $P^{B-A}\,{}_\Sigma\|_B\, S$, whose traces are not in K. Let $\bar{t}$ and $t = \bar{t}\sigma$ denote the traces of $\bar{g}$ and g, respectively. Then $t \notin K$ and

$$\bar{t} \in K, \quad t \in L(P^{B-A}), \tag{14}$$

where the final membership follows from the fact that the priority set of P^{B-A} is Σ.

If $\sigma \in A - B$, then it follows from (11) that $t \in K$, contrary to assumption. Thus, without loss of generality, we may assume that $\sigma \in B$. Since $\bar{r} := \Pi_\Gamma(\bar{t}) \in \Pi_\Gamma(K) = L(S)$, it follows from Lemma 5 that S executes every event in $\bar{r}$ while $P^{B-A}\,{}_\Sigma\|_B\, S$ executes the refusal-trace $\bar{g}$. Since $\sigma \in B$, the final event in g must occur synchronously in P^{B-A} and S. This implies that $\bar{r}\sigma \in L(S) = \Pi_\Gamma(K)$. Thus, there exists $s \in K$ such that $\Pi_\Gamma(s) = \bar{r}\sigma$. Since the last observable event in s is σ, by replacing s with a prefix if necessary, we may assume that $s = \bar{s}\sigma$. Then

$$\bar{s} \in K, \quad \Pi_\Gamma(\bar{s}) = \bar{r} = \Pi_\Gamma(\bar{t}). \tag{15}$$

From (14), (15), and the observability assumption it follows that $t \in K$, contrary to assumption. This establishes (13) and completes the proof of sufficiency. □

Remark 13. It follows from Lemma 5 that if $S := \det(\Pi_\Gamma(K))$ is used to solve the RSCP, then every event in $\Sigma - \Gamma$ which occurs in the closed-loop system is executed by the supervisor. The events in Γ are not observed by the supervisor and are executed only by the plant.

With Theorem 5 in hand, we can obtain a successful control design for the example presented in §2. Recall that for the given partially observed plant, there is no supervisor of the Ramadge–Wonham type that is consistent with the observation mask and satisfies the upper-bound specification of preventing jamming and the lower-bound specification of permitting cyclic operation.

Example 5. Let $\mathcal{P}$ denote the open-loop NSM described in §2 and depicted in Fig. 1(c). $\mathcal{P}$ is partially observed with a natural projection mask corresponding to $\Gamma = \{\gamma_1, \gamma_2\}$. The partition of the event set Σ into subsets of controllable, uncontrollable, and driven events is given by $\Sigma_c = \{\alpha\}$, $\Sigma_u = \{\beta, \gamma_1, \gamma_2, \eta, \lambda\}$, $\Sigma_d = \{\mu\}$. Thus, in a PSC-based control design, the priority sets of the plant and supervisor are $A = \{\alpha, \beta, \gamma_1, \gamma_2, \eta, \lambda\}$ and $B = \{\alpha, \mu\}$, respectively.

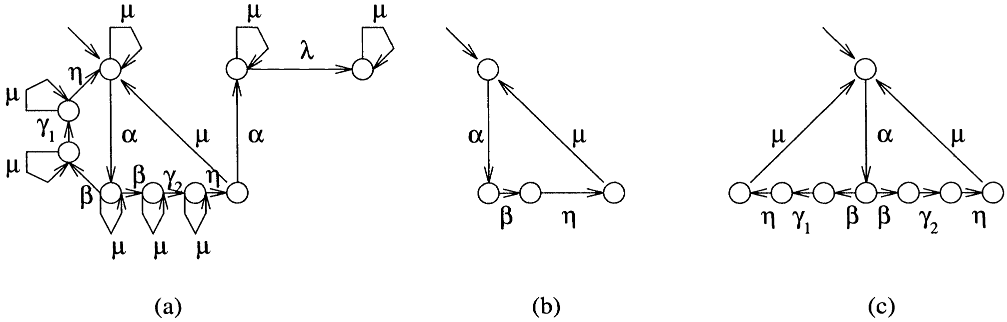


FIG. 4. *Diagram illustrating Example* 5.

The augmented plant $\mathcal{P}^{B-A}$ is shown in Fig. 4(a). The requirement that the jamming event λ never occur in the closed-loop system is represented by the specification $L(P\,_A\|_B\, S) \subseteq K$, where P is the trajectory model of $\mathcal{P}$, S is the trajectory model of a supervisor, and $K := \{s \in L(P^{B-A}) \mid \text{no event in } s \text{ is } \lambda\}$. Since $s := \alpha\beta\gamma_2\eta\alpha \in K$, $s\lambda \in L(P^{B-A}) - K$, and $\lambda \in A - B = \Sigma_u$, K is not controllable with respect to $A - B$ and $L(P^{B-A})$, and a straightforward calculation yields

$$K^{\uparrow} = pr((\mu^*\alpha\mu^*\beta\mu^*(\gamma_1\mu^*\eta + \gamma_2\mu^*\eta\mu))^*)$$

as the supremal controllable sublanguage [22]. However, $K^{\uparrow}$ is not observable relative to the augmented plant language $L(P^{B-A})$ and mask $\Pi_\Gamma(\cdot)$. In particular, if $\bar{t} \in (\mu^*\alpha\mu^*\beta\mu^*(\gamma_1\mu^*\eta + \gamma_2\mu^*\eta\mu))^*\mu^*\alpha\mu^*\beta\mu^*\gamma_2\mu^*\eta \subset K^{\uparrow}$, then $\bar{t}\alpha \in L(P^{B-A}) - K^{\uparrow}$, and $\exists \bar{s} \in (\mu^*\alpha\mu^*\beta\mu^*(\gamma_1\mu^*\eta + \gamma_2\mu^*\eta\mu))^*\mu^*\alpha\mu^*\beta\mu^*\gamma_1\mu^*\eta \subset K^{\uparrow}$ such that $\Pi_\Gamma(\bar{s}) = \Pi_\Gamma(\bar{t})$ and $\bar{s}\alpha \in K^{\uparrow}$.

It is easy to see that the supremal normal sublanguage [16], [4] of $K^{\uparrow}$ is given by

$$\hat{K} = pr((\mu^*\alpha\mu^*\beta\mu^*(\gamma_1 + \gamma_2)\mu^*\eta\mu)^*).$$

Since $\hat{K}$ is obtained from $K^{\uparrow}$ by disabling certain occurrences of the controllable event α, it is controllable. The fact that $\hat{K}$ is controllable also follows from the fact that it equals the closed and observable sublanguage of $K^{\uparrow}$, computed using the formula given in [13, Eq. 10], and the fact that controllability is preserved under such a computation [13, Thm. 5]. Since normality implies observability [16], it follows from Thm. 5 that $\hat{K}$ can be obtained as the closed-loop language by using an appropriate PSC-based supervisor, one choice being $S := \det(\Pi_\Gamma(\hat{K}))$.

While the supervisor $S = \det(\Pi_\Gamma(\hat{K}))$ is minimally restrictive, it possesses an undesirable trait: S can execute arbitrarily long sequences of the driven event μ. Since the plant can never execute more than one μ in succession, all but at most one of a sequence of μ's requested by the supervisor will be refused by the plant. We can remove this redundancy by replacing $\hat{K}$ by the sublanguage $\hat{K}' := pr((\alpha\beta(\gamma_1 + \gamma_2)\eta\mu)^*)$. $\hat{K}'$ is obtained by removing the self-loops on μ from $\hat{K}$, and is also both controllable and observable. By Theorem 5, the PSC-based supervisor $S' := \det(\Pi_\Gamma(\hat{K}'))$ will impose $\hat{K}'$ as the closed-loop language. A minimal state machine realization $\S'$ for the deterministic trajectory model S' is shown in Fig. 4(b), and the resulting closed-loop NSM $\mathcal{P}\ {}_A\|_B\ \S'$ is depicted in Fig. 4(c). The closed-loop language $\hat{K}'$ does not contain λ yet permits γ_1 and γ_2 to be executed arbitrarily many times. Hence, the dual objectives of preventing jamming and permitting cyclic operation are met.

The supervisor $\S'$ implements the following control strategy: $\S'$ tracks the inputting, commencement of processing, and completion/outputting of each part by executing α, β and η synchronously with $\mathcal{P}$. Between the synchronous executions of β and η, the plant executes either γ_1 or γ_2 without the participation or knowledge of the supervisor. Following the synchronous execution of η, the supervisor requests execution of the realignment event μ. If the mechanism is misaligned, i.e., γ_2 has preceded η, then the plant executes μ synchronously with the supervisor. This corrects the alignment and returns the plant to its initial state. On the other hand, if misalignment has not occurred, i.e., γ_1 has preceded η, then the plant refuses μ and this event occurs solely in the supervisor. The possibility that the plant can refuse an event offered by the supervisor is an essential feature of this control design.

7. Conclusion. In this paper we have studied the supervisory control of nondeterministic plants in the presence of driven events under complete as well as partial observation. We have shown that prioritized synchronous composition is an adequate control mechanism for this purpose. The trajectory model, used for describing the behavior of nondeterministic systems, is shown to be a language congruence with respect to prioritized synchronous composition. Hence it is quite useful for describing the behaviors of nondeterministic systems which may be controlled via PSC. It is shown that the supervisory control problem with driven events is solvable if and only if the target language is controllable and observable with respect to the language of the plant augmented by the set of driven events. In case the languages involved are regular, one way to perform the test for controllability/observability is to construct a deterministic system and language equivalent to the augmented plant, and apply a known test for controllability/observability [23], [14], [26]. However, it can be shown that by modifying the algorithms in these references, the tests can be performed without having to do such a nondeterministic to deterministic conversion. Hence it is possible to obtain algorithms of polynomial complexity (polynomial in the product of the number of states in the given plant NSM and that in the deterministic generator of the desired language) for testing controllability/observability. Due to the augmentation, the solvability depends on the trajectory model of the plant—not simply on its language. We have also described the associativity and augmentation properties of PSC, which are useful in the analysis of supervisory control.

Appendix A. Proof of Theorems 2 and 3.

Proof of Theorem 2. Let $\mathcal{R} = \mathcal{P}\ {}_A\|_B\ \mathcal{Q}$. First, we show that

$$T(\mathcal{R}) \subseteq T(\mathcal{P})\ {}_A\|_B\ T(\mathcal{Q}). \tag{16}$$

We prove by induction on length of refusal-trace that if $e \in T(\mathcal{R})$ and $x_r = (x_p, x_q) \in \delta^T_{\mathcal{R}}(x^0_{\mathcal{R}}, e)$, then there exist $e_p \in T(\mathcal{P})$, $e_q \in T(\mathcal{Q})$ such that

(i) the final refusal sets of e_p, e_q are $\Re_{\mathcal{P}}(x_p)$, $\Re_{\mathcal{Q}}(x_q)$, respectively;

(ii) $e \in e_p\ {}_A\|_B\ e_q$;

(iii) $x_r \in \delta^T_{\mathcal{P}}(x^0_{\mathcal{P}}, e_p) \times \delta^T_{\mathcal{Q}}(x^0_{\mathcal{Q}}, e_q)$.

Consider a zero-length refusal-trace $e = \Sigma' \in T(\mathcal{R})$. Then there exists $x_r = (x_p, x_q) \in \epsilon^*_{\mathcal{R}}(x^0_{\mathcal{R}})$ such that $\Sigma' \subseteq \Re_{\mathcal{R}}(x_r)$. Lemma 3 implies that $x_p \in \epsilon^*_{\mathcal{P}}(x^0_{\mathcal{P}})$, $x_q \in$

$\epsilon^*_{\mathcal{Q}}(x^0_{\mathcal{Q}})$, $\Sigma' \subseteq \Re_{\mathcal{P}}(x_p) \, {}_A\!\bigotimes_B \Re_{\mathcal{Q}}(x_q)$. Setting $e_p = \Re_{\mathcal{P}}(x_p)$, $e_q = \Re_{\mathcal{Q}}(x_q)$, it follows that (i), (ii), (iii) are satisfied.

For the induction step, consider a refusal-trace $e = \bar{e}(\sigma, \Sigma') \in T(\mathcal{R})$. Then there exist $\bar{x}_r = (\bar{x}_p, \bar{x}_q) \in \delta^T_{\mathcal{R}}(x^0_{\mathcal{R}}, \bar{e})$, $x'_r = (x'_p, x'_q) \in \delta_{\mathcal{R}}(\bar{x}_r, \sigma)$, $x_r = (x_p, x_q) \in \epsilon^*_{\mathcal{R}}(x'_r)$ such that $\Sigma' \subseteq \Re_{\mathcal{R}}(x_r)$. By induction hypothesis, there exist $\bar{e}_p \in T(\mathcal{P})$, $\bar{e}_q \in T(\mathcal{Q})$ with final refusal sets $\Re_{\mathcal{P}}(\bar{x}_p)$, $\Re_{\mathcal{Q}}(\bar{x}_q)$, respectively, such that $\bar{e} \in \bar{e}_p \, {}_A\|_B \, \bar{e}_q, \bar{x}_p \in \delta^T_{\mathcal{P}}(x^0_{\mathcal{P}}, \bar{e}_p), \bar{x}_q \in \delta^T_{\mathcal{Q}}(x^0_{\mathcal{Q}}, \bar{e}_q)$. Since σ is executable in $\bar{x}_r$, it follows from Definition 9 that there are three cases:

(a) $\delta_{\mathcal{P}}(\bar{x}_p, \sigma) \neq \emptyset$, $\delta_{\mathcal{Q}}(\bar{x}_q, \sigma) \neq \emptyset$;
(b) $\delta_{\mathcal{P}}(\bar{x}_p, \sigma) \neq \emptyset$, $\sigma \in \Re_{\mathcal{Q}}(\bar{x}_q)$, $\sigma \notin B$;
(c) $\delta_{Q}(\bar{x}_q, \sigma) \neq \emptyset$, $\sigma \in \Re_{\mathcal{P}}(\bar{x}_p)$, $\sigma \notin A$.

By symmetry, it suffices to consider cases (a) and (b).

In case (a), $x'_p \in \delta_{\mathcal{P}}(\bar{x}_p, \sigma)$, $x'_q \in \delta_{\mathcal{Q}}(\bar{x}_q, \sigma)$. Setting $e_p = \bar{e}_p(\sigma, \Re_{\mathcal{P}}(x_p))$, $e_q = \bar{e}_q(\sigma, \Re_{\mathcal{Q}}(x_q))$, and using the fact that

$$(17) \qquad \Sigma' \subseteq \Re_{\mathcal{R}}(x_r) = \Re_{\mathcal{P}}(x_p) \, {}_A\!\bigotimes_B \Re_{\mathcal{Q}}(x_q),$$

it follows easily that $e_p \in T(\mathcal{P})$, $e_q \in T(\mathcal{Q})$, and conditions (i), (ii), (iii) are satisfied.

In case (b), $x'_p \in \delta_{\mathcal{P}}(\bar{x}_p, \sigma)$, $x'_q = \bar{x}_q$. Set $e_p = \bar{e}_p(\sigma, \Re_{\mathcal{P}}(x_p))$ and let e_q be the refusal-trace obtained from $\bar{e}_q$ by replacing its final refusal set $\Re_{\mathcal{Q}}(\bar{x}_q)$ with the set $\Re_{\mathcal{Q}}(x_q)$. (Since $x_q \in \epsilon^*_{\mathcal{Q}}(\bar{x}_q)$, the new final refusal set will contain the old final refusal set.) Then $e_p \in T(\mathcal{P})$, $e_q \in T(\mathcal{Q})$ and conditions (i), (iii) are clearly satisfied. It follows from Definition 10 that $\bar{e} \in \bar{e}_p \, {}_A\|_B \, \bar{e}_q \subseteq \bar{e}_p \, {}_A\|_B \, e_q$. Since $\sigma \in \Re_{\mathcal{Q}}(\bar{x}_q) \subseteq \Re_{\mathcal{Q}}(x_q)$, it follows from property (T5) that $e_q(\sigma, \emptyset) \notin T(\mathcal{Q})$. Since $\sigma \notin B$ and (17) holds, it follows from Definition 10 that condition (ii) is satisfied. This completes the induction step and establishes (16).

It remains to show that

$$(18) \qquad T(\mathcal{P}) \, {}_A\|_B \, T(\mathcal{Q}) \subseteq T(\mathcal{R}).$$

We prove by induction on $|e_p| + |e_q|$ that if $e \in e_p \, {}_A\|_B \, e_q$ with $e_p \in T(\mathcal{P})$, $e_q \in T(\mathcal{Q})$, then

$$(19) \qquad \delta^T_{\mathcal{P}}(x^0_{\mathcal{P}}, e_p) \times \delta^T_{\mathcal{Q}}(x^0_{\mathcal{Q}}, e_q) \subseteq \delta^T_{\mathcal{R}}(x^0_{\mathcal{R}}, e).$$

Since the set on the left side is nonempty by assumption, this implies that the set on the right side is nonempty, i.e., that $e \in T(\mathcal{R})$.

Let $e_p = \Sigma_p$, $e_q = \Sigma_q$ be zero-length refusal-traces of $\mathcal{P}$, $\mathcal{Q}$, respectively, and let $x_p \in \delta^T_{\mathcal{P}}(x^0_{\mathcal{P}}, \Sigma_p)$, $x_q \in \delta^T_{\mathcal{Q}}(x^0_{\mathcal{Q}}, \Sigma_q)$. Then $x_p \in \epsilon^*_{\mathcal{P}}(x^0_{\mathcal{P}}), x_q \in \epsilon^*_{\mathcal{Q}}(x^0_{\mathcal{Q}}), \Sigma_p \subseteq \Re_{\mathcal{P}}(x_p), \Sigma_q \subseteq \Re_{\mathcal{Q}}(x_q)$. Let $x_r = (x_p, x_q)$. Then $x_r \in \epsilon^*_{\mathcal{R}}(x^0_{\mathcal{R}})$. It follows from Definition 10 and Lemma 3 that $e = \Sigma'$ with $\Sigma' \subseteq \Sigma_p \, {}_A\!\bigotimes_B \Sigma_q \subseteq \Re_{\mathcal{P}}(x_p) \, {}_A\!\bigotimes_B \Re_{\mathcal{Q}}(x_q) = \Re_{\mathcal{R}}(x_r)$. This shows that $x_r \in \delta^T_{\mathcal{R}}(x^0_{\mathcal{R}}, \Sigma')$, so (19) holds in the zero-length case.

For the induction step, write $e_p = \bar{e}_p(\sigma_p, \Sigma_p)$, $e_q = \bar{e}_q(\sigma_q, \Sigma_q)$, and suppose $e = \bar{e}(\sigma, \Sigma') \in e_p \, {}_A\|_B \, e_q$. It follows from Definition 10 that there are three cases to consider:

(d) $\sigma = \sigma_p = \sigma_q$, $\bar{e} \in \bar{e}_p \, {}_A\|_B \, \bar{e}_q$, $\Sigma' \subseteq \Sigma_p \, {}_A\!\bigotimes_B \Sigma_q$;
(e) $\sigma = \sigma_p$, $\sigma \notin B$, $e_q(\sigma, \emptyset) \notin T(\mathcal{Q})$, $\bar{e} \in \bar{e}_p \, {}_A\|_B \, e_q$, $\Sigma' \subseteq \Sigma_p \, {}_A\!\bigotimes_B \Sigma_q$;
(f) $\sigma = \sigma_q$, $\sigma \notin A$, $e_p(\sigma, \emptyset) \notin T(\mathcal{P})$, $\bar{e} \in e_p \, {}_A\|_B \, \bar{e}_q$, $\Sigma' \subseteq \Sigma_p \, {}_A\!\bigotimes_B \Sigma_q$.

By symmetry, it suffices to consider cases (d), (e).

For case (d), let $x_r = (x_p, x_q) \in \delta^T_{\mathcal{P}}(x^0_{\mathcal{P}}, e_p) \times \delta^T_{\mathcal{Q}}(x^0_{\mathcal{Q}}, e_q)$. Then $\Sigma_p \subseteq \Re_{\mathcal{P}}(x_p)$, $\Sigma_q \subseteq \Re_{\mathcal{Q}}(x_q)$, so

$$(20) \qquad \Sigma' \subseteq \Sigma_p \, {}_A\!\bigotimes_B \Sigma_q \subseteq \Re_{\mathcal{P}}(x_p) \, {}_A\!\bigotimes_B \Re_{\mathcal{Q}}(x_q)) = \Re_{\mathcal{R}}(x_r).$$

Since $x_p \in \delta^T_{\mathcal{P}}(x^0_{\mathcal{P}}, e_p)$, $x_q \in \delta^T_{\mathcal{Q}}(x^0_{\mathcal{Q}}, e_q)$, there exist $\bar{x}_p \in \delta^T_{\mathcal{P}}(x^0_{\mathcal{P}}, \bar{e}_p)$, $\bar{x}_q \in \delta^T_{\mathcal{Q}}(x^0_{\mathcal{Q}}, \bar{e}_q)$, $x'_p \in \delta_{\mathcal{P}}(\bar{x}_{\mathcal{P}}, \sigma)$, $x'_q \in \delta_{\mathcal{Q}}(\bar{x}_{\mathcal{Q}}, \sigma)$ such that $x_p \in \epsilon^*_{\mathcal{P}}(x'_p)$, $x_q \in \epsilon^*_{\mathcal{Q}}(x'_q)$. Let $\bar{x}_r = (\bar{x}_p, \bar{x}_q)$. It follows from Definition 9 that $(x'_p, x'_q) \in \delta_R(\bar{x}_r, \sigma)$, while by induction hypothesis we may assume that $\bar{x}_r \in \delta^T_{\mathcal{R}}(x^0_{\mathcal{R}}, \bar{e})$. Then

$$x_r \in \epsilon^*_{\mathcal{R}}((x'_p, x'_q)) \subseteq \epsilon^*_{\mathcal{R}}(\delta_{\mathcal{R}}(\bar{x}_r, \sigma)) \subseteq \epsilon^*_{\mathcal{R}}(\delta_{\mathcal{R}}(\delta^T_{\mathcal{R}}(x^0_{\mathcal{R}}, \bar{e}), \sigma)). \tag{21}$$

We conclude from (20) and (21) that $x_r \in \delta^T_{\mathcal{R}}(x^0_{\mathcal{R}}, e)$ as required.

For case (e), let $x_r = (x_p, x_q) \in \delta^T_{\mathcal{P}}(x^0_{\mathcal{P}}, e_p) \times \delta^T_{\mathcal{Q}}(x^0_{\mathcal{Q}}, e_q)$. The inclusions given by (20) hold as in the previous case. Since $x_p \in \delta^T_{\mathcal{P}}(x^0_{\mathcal{P}}, e_p)$, there exist $\bar{x}_p \in \delta^T_{\mathcal{P}}(x^0_{\mathcal{P}}, \bar{e}_p)$ and $x'_p \in \delta_{\mathcal{P}}(\bar{x}_{\mathcal{P}}, \sigma)$ such that $x_p \in \epsilon^*_{\mathcal{P}}(x'_p)$. Let $\bar{x}_r = (\bar{x}_p, x_q)$. We have $\delta_{\mathcal{Q}}(\epsilon^*_{\mathcal{Q}}(x_q), \sigma) \subseteq \delta_{\mathcal{Q}}(\delta^T_{\mathcal{Q}}(x^0_{\mathcal{Q}}, e_q), \sigma) = \emptyset$, where the final equality follows from the assumption that $e_q(\sigma, \emptyset) \notin T(\mathcal{Q})$. This implies that $\sigma \in \mathfrak{R}_{\mathcal{Q}}(x_q)$. It then follows from Definition 9 that $(x'_p, x_q) \in \delta_{\mathcal{R}}(\bar{x}_r, \sigma)$, while by induction hypothesis, we may assume that $\bar{x}_r \in \delta^T_{\mathcal{R}}(x^0_{\mathcal{R}}, \bar{e})$. Then

$$x_r \in \epsilon^*_{\mathcal{R}}((x'_p, x_q)) \subseteq \epsilon^*_{\mathcal{R}}(\delta_{\mathcal{R}}(\bar{x}_r, \sigma)) \subseteq \epsilon^*_{\mathcal{R}}(\delta_{\mathcal{R}}(\delta^T_{\mathcal{R}}(x^0_{\mathcal{R}}, \bar{e}), \sigma)). \tag{22}$$

We conclude from (20) and (22) that $x_r \in \delta^T_{\mathcal{R}}(x^0_{\mathcal{R}}, e)$ as required. This completes the induction step and establishes (18). □

The following result was proved in the course of the proof of Theorem 2.

COROLLARY 5. *For* NSM*s* $\mathcal{P}, \mathcal{Q}$ *and* $A, B \subseteq \Sigma$, *let* $\mathcal{R} := \mathcal{P}\ {}_A\|_B\ \mathcal{Q}$. *Then*

(1) *for each* $e \in T(\mathcal{R})$ *and* $x_r = (x_p, x_q) \in \delta^T_{\mathcal{R}}(x^0_{\mathcal{R}}, e)$, *there exists* $e_p \in T(\mathcal{P})$ *and* $e_q \in T(\mathcal{Q})$ *such that* $\Sigma_{|e_p|}(e_p) = \mathfrak{R}_{\mathcal{P}}(x_p), \Sigma_{|e_q|}(e_q) = \mathfrak{R}_{\mathcal{Q}}(x_q)$, $e \in e_p\ {}_A\|_B\ e_q$, *and* $x_r \in \delta^T_{\mathcal{P}}(x^0_{\mathcal{P}}, e_p) \times \delta^T_{\mathcal{Q}}(x^0_{\mathcal{Q}}, e_q)$;

(2) *for each* $e_p \in T(\mathcal{P}), e_q \in T(\mathcal{Q})$ *and* $e \in e_p\ {}_A\|_B\ e_q$; $\delta^T_{\mathcal{P}}(x^0_{\mathcal{P}}, e_p) \times \delta^T_{\mathcal{Q}}(x^0_{\mathcal{Q}}, e_q) \subseteq \delta^T_{\mathcal{R}}(x^0_{\mathcal{R}}, e)$.

In order to prove Theorem 3 we will use the following result which gives a monotonicity property of the PSC of refusal-traces with respect to the dominance partial order.

LEMMA 6. *Let* P, Q *be trajectory models,* $A, B \subseteq \Sigma$, $f_p, e_p \in P$, $f_q, e_q \in Q$, *with* $f_p \sqsubseteq e_p$, $f_q \sqsubseteq e_q$. *Then* $f_p\ {}_A\|_B\ f_q \subseteq e_p\ {}_A\|_B\ e_q$.

Proof. The proof is by induction on the sum of refusal-trace lengths $|e_p| + |e_q|$. If $|e_p| + |e_q| = 0$, then $f_p = \hat{\Sigma}_p$, $e_p = \Sigma_p$, $f_q = \hat{\Sigma}_q$, $e_q = \Sigma_q$ with $\hat{\Sigma}_p \subseteq \Sigma_p$, $\hat{\Sigma}_q \subseteq \Sigma_q$. Let $f \in f_p\ {}_A\|_B\ f_q$. Then $f = \Sigma' \subseteq \hat{\Sigma}_p\ {}_A\bigotimes_B\ \hat{\Sigma}_q \subseteq \Sigma_p\ {}_A\bigotimes_B\ \Sigma_q$. Thus, $f \in e_p\ {}_A\|_B\ e_q$.

For the induction step, write $f_p = \bar{f}_p(\sigma_p, \hat{\Sigma}_p)$, $e_p = \bar{e}_p(\sigma_p, \Sigma_p)$, $f_q = \bar{f}_q(\sigma_q, \hat{\Sigma}_q)$, $e_q = \bar{e}_q(\sigma_q, \Sigma_q)$ with $\bar{f}_p \sqsubseteq \bar{e}_p$, $\hat{\Sigma}_p \subseteq \Sigma_p$, $\bar{f}_q \sqsubseteq \bar{e}_q$, $\hat{\Sigma}_q \subseteq \Sigma_q$. Let $f = \bar{f}(\sigma, \Sigma') \in f_p\ {}_A\|_B\ f_q$. There are three cases to consider. (1) Suppose $\bar{f} \in \bar{f}_p\ {}_A\|_B\ \bar{f}_q$, $\sigma = \sigma_p = \sigma_q$, $\Sigma' \subseteq \hat{\Sigma}_p\ {}_A\bigotimes_B\ \hat{\Sigma}_q$. By the induction hypothesis, $\bar{f} \in \bar{e}_p\ {}_A\|_B\ \bar{e}_q$. Since $\Sigma' \subseteq \hat{\Sigma}_p\ {}_A\bigotimes_B\ \hat{\Sigma}_q \subseteq \Sigma_p\ {}_A\bigotimes_B\ \Sigma_q$, this implies that $f \in e_p\ {}_A\|_B\ e_q$. (2) Suppose $\bar{f} \in \bar{f}_p\ {}_A\|_B\ f_q$, $\sigma = \sigma_p \notin B$, $f_q(\sigma, \emptyset) \notin Q$, $\Sigma' \subseteq \hat{\Sigma}_p\ {}_A\bigotimes_B\ \hat{\Sigma}_q$. By induction hypothesis, $\bar{f} \in \bar{e}_p\ {}_A\|_B\ e_q$. Also, $e_q(\sigma, \emptyset) \notin Q$, since otherwise (T3) would imply that $f_q(\sigma, \emptyset) \in Q$, a contradiction. Since $\Sigma' \subseteq \hat{\Sigma}_p\ {}_A\bigotimes_B\ \hat{\Sigma}_q \subseteq \Sigma_p\ {}_A\bigotimes_B\ \Sigma_q$, this implies that $f \in e_p\ {}_A\|_B\ e_q$. (3) Suppose $\bar{f} \in f_p\ {}_A\|_B\ \bar{f}_q$, $\sigma = \sigma_q \notin A$, $f_p(\sigma, \emptyset) \notin P$, $\Sigma' \subseteq \hat{\Sigma}_p\ {}_A\bigotimes_B\ \hat{\Sigma}_q$. This case is analogous to case (2). Hence, the induction step is complete. □

Proof of Theorem 3. By symmetry, it suffices to prove the inclusion

$$(P\ {}_A\|_B\ Q)\ {}_{A \cup B}\|_C\ R \subseteq P\ {}_A\|_{B \cup C}\ (Q\ {}_B\|_C\ R).$$

By Lemma 6, it suffices to show that if e_p, e_q, e_r are *saturated* refusal-traces of P, Q, R respectively, then $(e_p\;{}_A\|_B\;e_q)\;{}_{A\cup B}\|_C\;e_r \subseteq P\;{}_A\|_{B\cup C}\;(Q\;{}_B\|_C\;R)$. To prove this, we show by induction on the sum of the lengths $|e_p| + |e_q| + |e_r|$ that

$$(e_p\;{}_A\|_B\;e_q)\;{}_{A\cup B}\|_C\;e_r \subseteq e_p\;{}_A\|_{B\cup C}\;(e_q\;{}_B\|_C\;e_r). \tag{23}$$

For future reference, we note that the following identity holds:

$$\begin{aligned}(\Sigma_p\;{}_A\!\bigotimes_B\Sigma_q)\;{}_{A\cup B}\!\bigotimes_C\Sigma_r &= (\Sigma_p\cap\Sigma_q\cap\Sigma_r)\cup(\Sigma_p\cap A)\cup(\Sigma_q\cap B)\cup(\Sigma_r\cap C)\\ &= \Sigma_p\;{}_A\!\bigotimes_{B\cup C}(\Sigma_q\;{}_B\!\bigotimes_C\Sigma_r) := \hat\Sigma\end{aligned} \tag{24}$$

If $e_p = \Sigma_p$, $e_q = \Sigma_q$, $e_r = \Sigma_r$ each have zero-length, then $(e_p\;{}_A\|_B\;e_q)\;{}_{A\cup B}\|_C\;e_r$ consists of all subsets of $(\Sigma_p\;{}_A\!\bigotimes_B\Sigma_q)\;{}_{A\cup B}\!\bigotimes_C\Sigma_r$, while $e_p\;{}_A\|_{B\cup C}\;(e_q\;{}_B\|_C\;e_r)$ consists of all subsets of $\Sigma_p\;{}_A\!\bigotimes_{B\cup C}(\Sigma_q\;{}_B\!\bigotimes_C\Sigma_r)$. Thus, (23) follows from (24).

For the induction step, let $e_p = \bar e_p(\sigma_p, \Sigma_p)$, $e_q = \bar e_q(\sigma_q, \Sigma_q)$, $e_r = \bar e_r(\sigma_r, \Sigma_r)$. $\bar e_p$, $\bar e_q$, $\bar e_r$ are saturated since they are prefixes of saturated refusal-traces. Let $f \in e_p\;{}_A\|_B\;e_q$ and let $h \in f\;{}_{A\cup B}\|_C\;e_r \subseteq (e_p\;{}_A\|_B\;e_q)\;{}_{A\cup B}\|_C\;e_r$. Let $h = \bar h(\sigma, \hat\Sigma)$. (There is no loss of generality in taking the final refusal set of h to be the maximal set $\hat\Sigma$.) To establish the induction step, we consider several cases. (Some cases will not apply if at least one of the refusal-traces e_p, e_q, e_r has zero-length.)

(1) $\bar h \in f\;{}_{A\cup B}\|_C\;\bar e_r, \sigma \notin A\cup B$, $f(\sigma,\emptyset) \notin P\;{}_A\|_B\;Q$, $\sigma = \sigma_r$. (This is when the final event in h occurs in R but not in $P\;{}_A\|_B\;Q$.)

(2a) $\bar h \in \bar f\;{}_{A\cup B}\|_C\;e_r$, $\sigma \notin C$, $e_r(\sigma,\emptyset) \notin R$, $\sigma = \sigma_p$, $\bar f \in \bar e_p\;{}_A\|_B\;e_q$, $\sigma \notin B$, $e_q(\sigma,\emptyset) \notin Q$. (This is when the final event in h occurs in $P\;{}_A\|_B\;Q$ but not in R, and within $P\;{}_A\|_B\;Q$, it occurs in P but not in Q.)

(2b) $\bar h \in \bar f\;{}_{A\cup B}\|_C\;e_r$, $\sigma \notin C$, $e_r(\sigma,\emptyset) \notin R$, $\sigma = \sigma_q$, $\bar f \in e_p\;{}_A\|_B\;\bar e_q$, $\sigma \notin A$, $e_p(\sigma,\emptyset) \notin P$. (This is when the final event in h occurs in $P\;{}_A\|_B\;Q$ but not in R, and within $P\;{}_A\|_B\;Q$, it occurs in Q but not in P.)

(2c) $\bar h \in \bar f\;{}_{A\cup B}\|_C\;e_r$, $\sigma \notin C$, $e_r(\sigma,\emptyset) \notin R$, $\sigma = \sigma_p = \sigma_q$, $\bar f \in \bar e_p\;{}_A\|_B\;\bar e_q$. (This is when the final event in h occurs in $P\;{}_A\|_B\;Q$ but not in R, and within $P\;{}_A\|_B\;Q$, it occurs in both P and Q.)

(3a) $\bar h \in \bar f\;{}_{A\cup B}\|_C\;\bar e_r$, $\sigma = \sigma_p = \sigma_r$, $\bar f \in \bar e_p\;{}_A\|_B\;e_q$, $\sigma \notin B$, $e_q(\sigma,\emptyset) \notin Q$. (This is when the final event in h occurs in $P\;{}_A\|_B\;Q$ and in R, and within $P\;{}_A\|_B\;Q$, it occurs in P but not in Q.)

(3b) $\bar h \in \bar f\;{}_{A\cup B}\|_C\;\bar e_r$, $\sigma = \sigma_q = \sigma_r$, $\bar f \in e_p\;{}_A\|_B\;\bar e_q$, $\sigma \notin A$, $e_p(\sigma,\emptyset) \notin P$. (This is when the final event in h occurs in $P\;{}_A\|_B\;Q$ and in R, and within $P\;{}_A\|_B\;Q$, it occurs in Q but not in P.)

(3c) $\bar h \in \bar f\;{}_{A\cup B}\|_C\;\bar e_r$, $\sigma = \sigma_p = \sigma_q = \sigma_r$, $\bar f \in \bar e_p\;{}_A\|_B\;\bar e_q$. (This is when the final event in h occurs in $P\;{}_A\|_B\;Q$ and in R, and within $P\;{}_A\|_B\;Q$, it occurs in both P and Q.)

We include a detailed proof for case (2a). The other cases are proven in a similar manner and are left to the reader. Under the assumptions of (2a),

$$\bar h \in \bar f\;{}_{A\cup B}\|_C\;e_r \subseteq (\bar e_p\;{}_A\|_B\;e_q)\;{}_{A\cup B}\|_C\;e_r \subseteq \bar e_p\;{}_A\|_{B\cup C}\;(e_q\;{}_B\|_C\;e_r),$$

where the last inclusion is the induction hypothesis. Thus, there exists

$$g = \bar g(\sigma', \Sigma_q\;{}_B\!\bigotimes_C\Sigma_r) \in e_q\;{}_B\|_C\;e_r$$

with $\sigma' \in \{\sigma_q, \sigma_r\}$ such that $\bar h \in \bar e_p\;{}_A\|_{B\cup C}\;g$. (By Lemma 6, there is no loss of generality in taking the final refusal set of g to be the maximal set $\Sigma_q\;{}_B\!\bigotimes_C\Sigma_r$.)

Since $e_q(\sigma,\emptyset) \notin Q$ and $e_r(\sigma,\emptyset) \notin R$, it follows from the assumption that e_q and e_r are saturated such that $\sigma \in \Sigma_q$ and $\sigma \in \Sigma_r$. Thus, $\sigma \in \Sigma_q \cap \Sigma_r \subseteq \Sigma_q\, {}_B\!\bigotimes_C \Sigma_r$. By (T5), $g(\sigma,\emptyset) \notin Q\, {}_B\|_C\, R$. This together with $\sigma \notin B \cup C$, implies that

$$h = \bar{h}(\sigma,\hat{\Sigma}) = \bar{h}(\sigma, \Sigma_p\, {}_A\!\bigotimes_{B\cup C}(\Sigma_q\, {}_B\!\bigotimes_C \Sigma_r)) \in e_p\, {}_A\|_{B\cup C}\, g \subseteq e_p\, {}_A\|_{B\cup C}\, (e_q\, {}_B\|_C\, e_r),$$

completing the induction step. □

REFERENCES

[1] J. C. M. Baeten, J. A. Bergstra, and J. W. Klop, *Ready-trace semantics for concrete process algebra with the priority operator*, The Comput. J., 30 (1987), pp. 498–506.

[2] J. C. M. Baeten and W. P. Weijland, *Process Algebra*, Cambridge University Press, Cambridge, 1990.

[3] S. Balemi, G. J. Hoffmann, P. Gyugyi, H. Wong-Toi, and G. F. Franklin, *Supervisory control of a rapid thermal multiprocessor*, IEEE Trans. Automat. Control, 38 (1993), pp. 1040–1059.

[4] R. D. Brandt, V. K. Garg, R. Kumar, F. Lin, S. I. Marcus, and W. M. Wonham, *Formulas for calculating supremal controllable and normal sublanguages*, Systems Control Lett., 15 (1990), pp. 111–117.

[5] R. Cieslak, C. Desclaux, A. Fawaz, and P. Varaiya, *Supervisory control of discrete event processes with partial observation*, IEEE Trans. Automat. Control, 33 (1988), pp. 249–260.

[6] C. H. Golaszewski and P. J. Ramadge, *Control of discrete event processes with forced events*, in Proc. 26th IEEE Conference on Decision and Control, Los Angeles, CA, 1987, pp. 247–251.

[7] M. Heymann, *Concurrency and discrete event control*, IEEE Control Systems Magazine, 10 (1990), pp. 103–112.

[8] M. Heymann and G. Meyer, *Algebra of discrete event processes*, Tech. report NASA 102848, NASA Ames Research Center, Moffett Field, CA, June 1991.

[9] C. A. R. Hoare, *Communicating Sequential Processes*, Prentice–Hall, Englewood Cliffs, NJ, 1985.

[10] J. E. Hopcroft and J. D. Ullman, *Introduction to Automata Theory, Languages and Computation*, Addison–Wesley, Reading, MA, 1979.

[11] K. Inan and P. Varaiya, *Algebras of discrete event models*, Proc. IEEE, 77 (1989), pp. 24–38.

[12] K. M. Inan, *An algebraic approach to supervisory control*, Math. Control Signals Systems, 5 (1992), pp. 151–164.

[13] R. Kumar, *Formulas for observability of discrete event dynamical systems*, in Proc. of 1993 Conference on Information Sciences and Systems, Baltimore, MD, Johns Hopkins University, 1993, pp. 581–586.

[14] R. Kumar, V. K. Garg, and S. I. Marcus, *On controllability and normality of discrete event dynamical systems*, Systems Control Lett., 17 (1991), pp. 157–168.

[15] R. Kumar and M. A. Shayman, *Non-blocking supervisory control of nondeterministic systems via prioritized synchronization*, Tech. report TR 93-58, Institute for Systems Research, University of Maryland, College Park, Maryland, July 1993.

[16] F. Lin and W. M. Wonham, *On observability of discrete-event systems*, Inform. Sci., 44 (1988), pp. 173–198.

[17] R. Milner, *A Calculus of Communicating Systems*, Springer-Verlag, Berlin, 1980.

[18] ———, *Communication and Concurrency*, Prentice–Hall, New York, 1989.

[19] D. M. R. Park, *Concurrency and automata on infinite sequences*, in Proc. 5th GI Conference, Lecture Notes in Computer Science 104, P. Deussen, ed., Springer-Verlag, New York, 1981, pp. 167–183.

[20] I. Phillips, *Refusal testing*, Theoret. Comput. Sci., 50 (1987), pp. 241–284.

[21] A. Pnueli, *Linear and branching structures in the semantics and logics of reactive systems*, in 12th International Colloquium on Automata, Languages, and Programming, Lecture Notes in Computer Science 194, W. Brauer, ed., Springer-Verlag, New York, 1985, pp. 15–32.

[22] P. J. Ramadge and W. M. Wonham, *On the supremal controllable sublanguage of a given language*, SIAM J. Control Optim., 25 (1987), pp. 637–659.

[23] ———, *Supervisory control of a class of discrete event processes*, SIAM J. Control Optim., 25 (1987), pp. 206–230.

[24] ———, *The control of discrete event systems*, Proc. IEEE: Special Issue on Discrete Event Systems, 77 (1989), pp. 81–98.

[25] M. A. SHAYMAN AND R. KUMAR, *Supervisory control of nondeterministic systems*, in Proc. of 1993 Conference on Information Sciences and Systems, Baltimore, MD, Johns Hopkins University, 1993, pp. 587–592.

[26] J. N. TSITSIKLIS, *On the control of discrete event dynamical systems*, Math. Control Signals Systems, 2 (1989), pp. 95–107.

[27] R. J. VAN GLABBEEK, *Comparative concurrency semantics, with refinement of actions*, Ph.D. thesis, Free University of Amsterdam, Amsterdam, the Netherlands, 1990.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 33, No. 2, pp. 498–527, March 1995

© 1995 Society for Industrial and Applied Mathematics
009

RISK-SENSITIVE PRODUCTION PLANNING OF STOCHASTIC MANUFACTURING SYSTEMS: A SINGULAR PERTURBATION APPROACH*

QING ZHANG†

Abstract. This paper is concerned with robust production planning of a stochastic manufacturing system in which the rate of machine breakdown and repair is much larger than the rate of fluctuation in demand. It is shown that the risk-sensitive production planning problem can be approximated by a limiting problem in which the stochastic machine availability process is replaced by its mean availability. The value function of the limiting problem is shown to be the only viscosity solution to the Isaacs equation associated with a zero-sum, two-player differential game. Near-optimal production plans are constructed from near-optimal controls of the limiting problem. Finally, these results are extended to problems with state constraints.

Key words. stochastic manufacturing systems, risk-sensitive control, differential games, Isaacs equations, viscosity solutions, near-optimal production planning

AMS subject classifications. 93E20, 90B30

1. Introduction.

Most manufacturing systems are large complex systems. Because of the large size of these systems, exact optimal policies for running them are quite difficult to obtain, both theoretically and computationally. One way to cope with these complexities is to develop methods of hierarchical controls for these systems. The idea of hierarchical control is to reduce the overall complex problem into manageable approximate problems or subproblems, solve these problems, and construct a solution for the original problem from solutions of these simpler problems. Development of such approaches for large complex systems has been identified as a particularly fruitful research area by the Committee on the Next Decade in Operations Research [3] as well as by the Panel on Future Directions in Control Theory [7]. A great deal of research in hierarchical control has been conducted by researchers in the areas of operations research, operations management, system theory, and control theory. Related literature can be found in recent papers by Rogers et al. [16], Saksena, O'Reilly, and Kokotovic [17], Gershwin [11], Lehoczky et al. [15], Soner [24], and the book by Sethi and Zhang [19].

An interesting approach in hierarchical control is the singular perturbation method. To illustrate this approach, let us consider a manufacturing system which consists of machines that are subject to breakdown and repair and which faces an uncertain demand. The objective of the system is to obtain the rate of production over time in order to meet the demand at the minimum expected discounted cost of production and inventory/shortages over the infinite horizon. Because of the uncertainties in machine availability and product demand, the exact optimal solution of such a problem is very difficult to obtain. To reduce the complexity, we consider the case in which the rate at which the machine breakdown and repair events occur is much larger than the rate of fluctuation in demand. The idea of a singular perturbation approach is to derive a limiting control problem, which is easier to solve than the original problem. This limiting problem is obtained by replacing the stochastic machine availability

* Received by the editors January 27, 1993; accepted for publication (in revised form) November 9, 1993. This research was supported in part by Natural Sciences and Engineering Research Council of Canada grant A4169.

† Faculty of Management, University of Toronto, Toronto, Ontario M5S 1V4, Canada. Present address: Department of Mathematics , University of Georgia, Athens, Georgia 30602.

process by the average total capacity of machines and by appropriately modifying the objective function. From its optimal control, one constructs an approximate optimal control of the original, more complex, problem. Research along this line can be found in Lehoczky et al. [15], Soner [24], Sethi and Zhang [20], [19], and Sethi, Zhang, and Zhou [21].

In this paper, we consider robust production plans of stochastic manufacturing systems with a risk-sensitive cost function (cf. Whittle [25]). This consideration is motivated by the following observations. First, since most manufacturing systems are large complex systems, it is very difficult to establish accurate mathematical models to describe them. Modeling errors are inevitable. Second, in practice, an optimal policy for a subdivision of a big corporation is usually not an optimal policy for the whole corporation. Therefore, optimal solutions with actual cost criteria may not be desirable in many real problems. An alternative approach is to consider robust controls. The design of robust controls emphasizes system stability rather than optimality. In some manufacturing systems, it is more desirable to consider controls that are robust enough to attenuate uncertain disturbances, which include modeling errors, and therefore to achieve the system stability. Robust control design is particularly important in manufacturing systems with unfavorable disturbances. There are two kinds of system disturbances in the system under consideration: (1) unfavorable internal disturbances—usually associated with unfavorable machine capacity fluctuations; (2) unfavorable external disturbances—usually associated with unfavorable fluctuations in demand.

The basic idea of the risk-sensitive control is to consider a risk-sensitive cost function that penalizes heavily on costs associated with large state trajectories and controls. Related literature on risk-sensitive control can be found, for example, in Whittle [25], Fleming and McEneaney [8], [9], James [14], and Glover and Doyle [12].

In [8], [14], risk-sensitive control problems of controlled diffusions are considered. Using the associated dynamic programming equations, the authors show that as the system noise goes to zero, the value function of the risk-sensitive control problem converges to the value function of a differential game problem. Then, a near-optimal policy for the differential game problem can be shown to be a near-optimal control for the risk-sensitive control problem. In this paper, we consider the risk-sensitive control of the manufacturing systems with stochastic production capacity and stochastic product demand. As the rate of fluctuation of the production capacity process goes to infinity, we show that the risk-sensitive control problem can be approximated by a limiting problem in which the stochastic capacity process can be averaged out and replaced by its average. We also show that the value function of the limiting problem satisfies the Isaacs equation of a zero-sum, two-player differential game. Then, we use a near-optimal control of the limiting problem to construct a near-optimal control for the original risk-sensitive control problem. In this paper, the machine capacity process will be assumed to be a finite state (jump) Markov chain. Our formulation is similar to that of [8], [14]. Because of the presence of the jump Markov disturbance processes, the dynamic programming approach used in [8], [14] will not work in our problem. To obtain the desired results, we first prove an asymptotic property of the machine capacity process (Lemma 3.1). Then, we use a combination of a probabilistic approach and a weak convergence approach to derive the convergence of the associated value functions (Theorems 4.1, 5.1, and 6.1) to the value function of a limiting problem. We show that the value function of the limiting problem is the only viscosity solution to the associated Isaacs equation (Theorem 5.2).

The plan of the paper is as follows. In the next section, we consider a relatively simple model of manufacturing systems. We formulate the risk-sensitive control problem and make some assumptions on the cost function and stochastic processes of the system. In §3, we give asymptotic estimates on the machine capacity process. Then, in §4 and §5, we show that the value function of the risk-sensitive control problem converges to the value function of a limiting problem that has a differential game interpretation. In §6, we construct near-optimal production plans for the original problem. In §7, we extend the results to a class of manufacturing systems with internal buffers. We present only the results obtained by considering a two-machine flowshop for simplicity in exposition. Finally, we conclude the paper by making some remarks.

2. Problem formulation. We consider a manufacturing system that produces n_0 distinct part types using m identical machines. Let $u_t \in R^{n_0}$ denote the vector of production rates, $x_t \in R^{n_0}$ denote the vector of total inventories/backlogs, and $z_t \in R^{n_0}$ denote the vector of demand rates. They satisfy the following system equation:

$$\dot{x}_t = u_t - z_t,\ x_0 = a \in R^{n_0}\ (a \text{ is given}). \tag{2.1}$$

We consider the manufacturing system that consists of machines that are subject to breakdown and repair. Let $\mathcal{M} = \{0, 1, \ldots, m\}$ denote the set of machine capacity states and let a random process $\alpha(\varepsilon, t) \in \mathcal{M}$, defined on a standard probability space $(\Omega, \mathcal{F}, P)$, denote the total capacity process for the manufacturing system, where ε is a small parameter to be specified later in this section. Since only a finite amount of production capacity is available at any given time t, it would impose an upper bound on the production rate u_t. For example, in the one-dimensional case ($n_0 = 1$), the production constraint is $0 \le u_t \le \alpha(\varepsilon, t)$.

We consider the risk sensitive cost function $J^{\varepsilon,\sqrt{\varepsilon}}(u.)$ defined by

$$J^{\varepsilon,\sqrt{\varepsilon}}(u.) = \sqrt{\varepsilon}\log E\left[\exp\left\{\frac{1}{\sqrt{\varepsilon}}\int_0^\infty e^{-\rho t}[h(x_t) + c(u_t)]dt\right\}\right], \tag{2.2}$$

where h is the inventory/backlog cost, c is the production cost, and $\rho > 0$ is the discount rate. The problem is to find an admissible control $u.$ (to be precisely defined later in this section) that minimizes $J^{\varepsilon,\sqrt{\varepsilon}}(u.)$.

In the above problem, we choose the risk-sensitivity parameter $\sqrt{\varepsilon} > 0$. This type of risk-sensitive control problem is classified as a risk-averse problem (cf. [25]). The cost function in (2.2) emphasizes the stability of the system since it penalizes heavily on large trajectory x_t and large control u_t when ε is small.

Notation. We make use of the following notation in this paper:

χ_D	the indicator function of any set D;
$O(y)$	a function of y such that $\sup_y \lvert O(y)\rvert/\lvert y\rvert < \infty$;
A'	the transpose of any matrix (or vector) A;
$\lVert\xi\rVert_\infty$	ess sup$\lvert\xi\rvert$ of any random variable ξ;
$f_x(x)$	the derivative of f at x;
$f_{x^+}(x)$	the right-hand derivative of f at x;
$C, C_0, C_1, \ldots$	multiplicative constants;
$k_h, k_0, k_1, \ldots$	exponential constants.

We now specify the production (or control) constraints. For each $i \in \mathcal{M} = \{0, 1, 2, \ldots, m\}$, let

$$\mathcal{U}(i) = \{l = (l_1, \ldots, l_{n_0}) \geq 0 : p \cdot l \leq i\} \subset R^{n_0}, \tag{2.3}$$

where $p = (p_1, \ldots, p_{n_0}) \geq 0$ are given constants with p_i representing the amount of capacity needed to produce part type i at rate 1. With this definition, the production constraint at time t is $u_t \in \mathcal{U}(\alpha(\varepsilon, t))$.

We make the following assumptions on the functions h and c and the random processes $\alpha(\varepsilon, t)$ and z_t.

(A1) h and c are convex functions. There exist constants C_0 and $k_h > 0$ such that for all x, x', u, and u',

$$\begin{aligned} &0 \leq h(x) \leq C_0(1 + |x|^{k_h}), \\ &|h(x) - h(x')| \leq C_0(1 + |x|^{k_h} + |x'|^{k_h})|x - x'|, \\ \text{and} \quad &|c(u) - c(u')| \leq C_0|u - u'|. \end{aligned}$$

(A2) Let $\varepsilon > 0$ denote a small parameter. The machine capacity process $\alpha(\varepsilon, t) \in \mathcal{M}$ is a finite-state Markov process governed by a generator $Q = Q^{(1)} + \varepsilon^{-1}Q^{(2)}$, where $Q^{(l)}$ is an $(m+1) \times (m+1)$ matrix such that $Q^{(l)} = (q_{ij}^{(l)})$ with $q_{ij}^{(l)} \geq 0$ if $i \neq j$ and $q_{ii}^{(l)} = -\sum_{j \neq i} q_{ij}^{(l)}$ for $l = 1, 2$. Moreover, $Q^{(2)}$ is irreducible.

(A3) The demand rate z_t is a bounded process which is independent of $\alpha(\varepsilon, t)$.

Remark 2.1. Assumptions (A1) and (A2) are used in [15], [20], [21], [22], [23], [27]. Soner [24] considers a model in which Q depends on the control variables. However, due to the jump Markov property of the machine capacity process, the viscosity solution method used in [24] will not work for the risk-sensitive control problem under consideration.

DEFINITION 2.1. *We say that a control* $u. = \{u_t : t \geq 0\}$ *is* admissible *if* u_t *is a* $\sigma\{\alpha(\varepsilon, s), z_s : s \leq t\}$ *adapted measurable process and* $u_t \in \mathcal{U}(\alpha(\varepsilon, t))$ *for all* $t \geq 0$. *We use* $\mathcal{A}^\varepsilon$ *to denote the set of all admissible controls. Then our control problem can be written as follows*:

$$\mathcal{P}^{\varepsilon,\sqrt{\varepsilon}} : \begin{cases} \text{min.} & J^{\varepsilon,\sqrt{\varepsilon}}(u.) \\ & = \sqrt{\varepsilon} \log E\left[\exp\left\{\dfrac{1}{\sqrt{\varepsilon}} \displaystyle\int_0^\infty e^{-\rho t}[h(x_t) + c(u_t)]dt\right\}\right], \\ \text{s.t.} & \dot{x}_t = u_t - z_t, \ x_0 = a, \ u. \in \mathcal{A}^\varepsilon, \\ \text{value fn.} & v^{\varepsilon,\sqrt{\varepsilon}} = \inf\limits_{u. \in \mathcal{A}^\varepsilon} J^{\varepsilon,\sqrt{\varepsilon}}(u.). \end{cases} \tag{2.4}$$

Let $\nu = (\nu_0, \nu_1, \ldots, \nu_m)'$ denote the equilibrium distribution of $Q^{(2)}$. That is, ν is the only positive solution to

$$\nu' Q^{(2)} = 0 \text{ and } \sum_{i=0}^{m} \nu_i = 1.$$

Intuitively, the machine capacity process $\alpha(\varepsilon, t)$ "converges" weakly to its equilibrium distribution as ε tends to zero. This property suggests that the problem $\mathcal{P}^{\varepsilon,\sqrt{\varepsilon}}$

can be approximated by a problem in which the stochastic machine availability is replaced by its equilibrium distribution.

Let $\mathcal{Z}_t$ denote the σ-algebra generated by z_s, $s \le t$, i.e., $\mathcal{Z}_t = \sigma\{z_s : s \le t\}$. As in [21], we consider the following control space:

$$\mathcal{A}^0 = \{U. = (u^0_\cdot, u^1_\cdot, \dots, u^m_\cdot) : u^i_t \in \mathcal{U}(i), \text{ and } U_t \text{ is a } \mathcal{Z}_t \text{ adapted process}\}.$$

We also consider two control problems $\mathcal{P}^{0,\sqrt{\varepsilon}}$ and $\mathcal{P}^{0,0}$ defined as follows:

$$(2.5)\quad \mathcal{P}^{0,\sqrt{\varepsilon}}: \begin{cases} \text{min.} & J^{0,\sqrt{\varepsilon}}(U.) \\ & = \sqrt{\varepsilon}\log\left[E\exp\left\{\frac{1}{\sqrt{\varepsilon}}\int_0^\infty e^{-\rho t}\left[h(x_t)+\sum_{i=0}^m \nu_i c(u^i_t)\right]dt\right\}\right], \\ \text{s.t.} & \dot{x}_t = \sum_{i=0}^m \nu_i u^i_t - z_t,\ x_0 = a,\ U. = (u^0_\cdot, \dots, u^m_\cdot) \in \mathcal{A}^0, \\ \text{value fn.} & v^{0,\sqrt{\varepsilon}} = \inf_{U.\in\mathcal{A}^0} J^{0,\sqrt{\varepsilon}}(U.), \end{cases}$$

and

$$(2.6)\quad \mathcal{P}^{0,0}: \begin{cases} \text{min.} & J^{0,0}(U.) = \left\|\int_0^\infty e^{-\rho t}\left[h(x_t)+\sum_{i=0}^m \nu_i c(u^i_t)\right]dt\right\|_\infty, \\ \text{s.t.} & \dot{x}_t = \sum_{i=0}^m \nu_i u^i_t - z_t,\ x_0 = a,\ U. = (u^0_\cdot, \dots, u^m_\cdot) \in \mathcal{A}^0, \\ \text{value fn.} & v^{0,0} = \inf_{U.\in\mathcal{A}^0} J^{0,0}(U.). \end{cases}$$

We will show that, when ε is small, $\mathcal{P}^{\varepsilon,\sqrt{\varepsilon}}$ can be approximated by $\mathcal{P}^{0,\sqrt{\varepsilon}}$, and $\mathcal{P}^{0,\sqrt{\varepsilon}}$ can be approximated further by $\mathcal{P}^{0,0}$. Therefore, $\mathcal{P}^{\varepsilon,\sqrt{\varepsilon}}$ can be approximated by $\mathcal{P}^{0,0}$. Then, a near optimal control for $\mathcal{P}^{0,0}$ will be used to construct controls for $\mathcal{P}^{\varepsilon,\sqrt{\varepsilon}}$ that are nearly optimal.

3. Asymptotic properties of $\alpha(\varepsilon, t)$.

In this section, we consider an asymptotic property of the process $\alpha(\varepsilon, t)$, which plays a very important role in this paper.

LEMMA 3.1. *For each $N_0 > 0$, there exist constants $\varepsilon_0 > 0$ and C such that for $0 < \varepsilon \le \varepsilon_0$, $i \in \mathcal{M}$, $t \ge 0$, and for any $\mathcal{Z}_t$ adapted process $\beta(t)$, $|\beta(t)| \le N_0$ a.s.,*

$$(3.1)\qquad E\left[\exp\left\{\frac{1}{\sqrt{\varepsilon}(t+1)^{\frac{3}{2}}}\left|\int_0^t (\chi_{\{\alpha(\varepsilon,s)=i\}} - \nu_i)\beta(s)ds\right|\right\}\middle|\mathcal{Z}_t\right] \le C, \text{ a.s.}$$

Proof. In view of Lemma A.2, it suffices to show that for any deterministic $\beta(t)$, $|\beta(t)| \le N_0$,

$$(3.2)\qquad E\exp\left\{\frac{1}{\sqrt{\varepsilon}(t+1)^{\frac{3}{2}}}\left|\int_0^t (\chi_{\{\alpha(\varepsilon,s)=i\}} - \nu_i)\beta(s)ds\right|\right\} \le C.$$

We let

$$\lambda(t) = (\chi_{\{\alpha(\varepsilon,t)=0\}}, \dots, \chi_{\{\alpha(\varepsilon,t)=m\}})'$$
$$\text{and}\quad w(t) = \lambda(t) - \lambda(0) - \int_0^t Q'\lambda(s)ds,$$

where $Q = Q^{(1)} + \varepsilon^{-1}Q^{(2)}$ is the generator of the process $\alpha(\varepsilon, t)$. Then, it is well known (cf. [5]) that $w(t) = (w_0(t), \ldots, w_m(t))'$ is a $\sigma\{\alpha(\varepsilon, s) : s \leq t\}$ martingale and

$$\lambda(t) = e^{Q't}\lambda(0) + \int_0^t e^{Q'(t-s)}dw(s). \tag{3.3}$$

Let

$$\bar{P} = \begin{pmatrix} \nu_0 & \cdots & \nu_0 \\ \vdots & \cdots & \vdots \\ \nu_m & \cdots & \nu_m \end{pmatrix}$$

and let $\Phi(t) = e^{Q't} - \bar{P}$. Then, $\Phi(t)$ is deterministic and by Lemma A.1

$$\Phi(t) = O(\varepsilon + e^{-k_0 t/\varepsilon}). \tag{3.4}$$

Moreover, since $\bar{P}\lambda(t) = \bar{P}\lambda(0) = \nu$ and $\bar{P}Q' = 0$,

$$\bar{P}w(t) = \bar{P}\left[\lambda(t) - \lambda(0) - \int_0^t Q'\lambda(s)ds\right] = \nu - \nu - \int_0^t \bar{P}Q'\lambda(s)ds = 0.$$

We combine (3.3) and the definition of $\Phi(t)$ to obtain

$$\begin{aligned} \lambda(t) - \nu = &\ \Phi(t)\lambda(0) + \int_0^t (\Phi(t-s) + \bar{P})dw(s) \\ = &\ \Phi(t)\lambda(0) + \int_0^t \Phi(t-s)dw(s). \end{aligned}$$

By integrating $(\lambda(s) - \nu)\beta(s)$ over $[0, t]$ and exchanging orders of integration, we have

$$\int_0^t (\lambda(s) - \nu)\beta(s)ds = \lambda(0)\int_0^t \Phi(s)\beta(s)ds + \int_0^t \left\{\int_s^t \Phi(r-s)\beta(r)dr\right\} dw(s).$$

Recall that $\beta(t)$ is assumed to be deterministic at the beginning of the proof. Both $\int_0^t \Phi(s)\beta(s)ds$ and $\int_s^t \Phi(r-s)\beta(r)dr$ are deterministic. Moreover,

$$\frac{1}{t+1}\int_0^t \Phi(s)\beta(s)ds = \frac{1}{t+1}\int_0^t O(\varepsilon + e^{-k_0 s/\varepsilon})ds = O(\varepsilon),$$

$$\frac{1}{t+1}\int_s^t \Phi(r-s)\beta(r)dr = \frac{1}{t+1}\int_s^t O(\varepsilon + e^{-k_0(r-s)/\varepsilon})dr = O(\varepsilon),$$

where $O(1)$ is deterministic and is independent of s and t. Thus,

$$\frac{1}{t+1}\int_0^t (\lambda(s) - \nu)\beta(s)ds = \varepsilon O(1) + \varepsilon\int_0^t O(1)dw(s). \tag{3.5}$$

Therefore,

$$\begin{aligned} &E\exp\left\{\frac{1}{\sqrt{\varepsilon}(t+1)^{\frac{3}{2}}}\left|\int_0^t (\lambda(s) - \nu)\beta(s)ds\right|\right\} \\ &\quad = \exp\frac{\sqrt{\varepsilon}O(1)}{\sqrt{t+1}} E\exp\left\{\frac{\sqrt{\varepsilon}}{\sqrt{t+1}}\left|\int_0^t O(1)dw(s)\right|\right\}. \end{aligned} \tag{3.6}$$

Recall that $w(t) = (w_0(t), \ldots, w_m(t))'$. It suffices to show that, for ε_0 small enough and for each $i \in \mathcal{M}$,

$$E \exp\left\{ \frac{\sqrt{\varepsilon}}{\sqrt{t+1}} \left| \int_0^t b(s,t) dw_i(s) \right| \right\} \leq C \tag{3.7}$$

for all bounded measurable functions $|b(s,t)| \leq N_1$ (for some fixed N_1) and $t \geq 0$. Now, for each $t_0 \geq 0$, let $b_0(s) = b(s, t_0)$. Then, by Lemma A.3,

$$\begin{aligned} & E \exp\left\{ \frac{\sqrt{\varepsilon}}{\sqrt{t+1}} \left| \int_0^t b_0(s) dw_i(s) \right| \right\} \\ & \leq e + (e-1) \sum_{j=1}^{\infty} e^j P\left(\frac{\sqrt{\varepsilon}}{\sqrt{t+1}} \left| \int_0^t b_0(s) dw_i(s) \right| \geq j \right). \end{aligned} \tag{3.8}$$

Now, we estimate

$$P\left(\frac{\sqrt{\varepsilon}}{\sqrt{t+1}} \left| \int_0^t b_0(s) dw_i(s) \right| \geq j \right).$$

Let $p(t) = \int_0^t b_0(s) dw_i(s)$. Then $p(t)$ is a local martingale. Let $q(\cdot)$ denote the only solution to the equation (cf. [4])

$$q(t) = 1 + \zeta \int_0^t q(s^-) dp(s),$$

where $q(s^-)$ is the left-hand limit of q at s and ζ is a positive constant to be determined later. Then, it is shown in [22] that

(1) $Eq(t) \leq 1$ for all $t \geq 0$;

(2) $q(t) = e^{\zeta p(t)} \prod_{s \leq t} (1 + \zeta \Delta p(s)) e^{-\zeta \Delta p(s)}$, where $\Delta p(s) := p(s) - p(s^-)$, $|\Delta p(s)| \leq N_1$;

(3) $q(t) \geq \exp\{\zeta p(t) - k_1 \zeta^2 N(t)\}$ for $0 < \zeta \leq \zeta_0$, $t > 0$, where k_1 and ζ_0 are positive constants and $N(t)$ is the number of jumps of $p(s)$ in $s \in [0,t]$. Therefore,

$$\begin{aligned} & P\left(\frac{\sqrt{\varepsilon}}{\sqrt{t+1}} \left| \int_0^t b_0(s) dw_i(s) \right| \geq j \right) \\ = \; & P\left(|p(t)| \geq \frac{j\sqrt{t+1}}{\sqrt{\varepsilon}} \right) \\ \leq \; & P\left(p(t) \geq \frac{j\sqrt{t+1}}{\sqrt{\varepsilon}} \right) + P\left(p(t) \leq -\frac{j\sqrt{t+1}}{\sqrt{\varepsilon}} \right). \end{aligned}$$

We first consider $P\left(p(t) \geq j\zeta\sqrt{t+1}/\sqrt{\varepsilon}\right)$. Let $a_j = j(t+1)/(8k_1\varepsilon)$. Then,

$$\begin{aligned} & P\left(p(t) \geq \frac{j\sqrt{t+1}}{\sqrt{\varepsilon}} \right) \\ & \leq P\left(q(t) \geq \exp\left\{ \frac{j\zeta\sqrt{t+1}}{\sqrt{\varepsilon}} - k_1\zeta^2 N(t) \right\} \right) \\ & \leq P\left(q(t) \geq \exp\left\{ \frac{j\zeta\sqrt{t+1}}{\sqrt{\varepsilon}} - k_1\zeta^2 N(t) \right\}, N(t) \leq a_j \right) + P(N(t) \geq a_j) \\ & \leq P\left(q(t) \geq \exp\left(\frac{j\zeta\sqrt{t+1}}{\sqrt{\varepsilon}} - k_1\zeta^2 a_j \right) \right) + P(N(t) \geq a_j) \\ & \leq \exp\left(-\frac{j\zeta\sqrt{t+1}}{\sqrt{\varepsilon}} + k_1\zeta^2 a_j \right) + P(N(t) \geq a_j). \end{aligned}$$

Now if we take $\zeta = 4\sqrt{\varepsilon}/\sqrt{t+1}$, then

$$\exp\left(-\frac{j\zeta\sqrt{t+1}}{\sqrt{\varepsilon}} + k_1\zeta^2 a_j\right) = e^{-2j}.$$

As in [22], we can show that for ε small enough,

$$P(N(t) \geq a_j) \leq 2\gamma^{a_j-1} \quad \text{for } j \geq j_0,$$

where $\gamma = 8ek_1/j_0 \in (0,1)$ and $j_0 > \max\{1, 8ek_1\}$. Thus,

$$P\left(p(t) \geq \frac{j\sqrt{t+1}}{\sqrt{\varepsilon}}\right) \leq \begin{cases} e^{-2j} + 2\gamma^{a_j-1} & \text{if } j \geq j_0, \\ 1 + e^{-2} & \text{if } j < j_0. \end{cases}$$

Repeating the same argument for the martingale $(-p(t))$ we get

$$P\left(p(t) \leq -\frac{j\sqrt{t+1}}{\sqrt{\varepsilon}}\right) \leq \begin{cases} e^{-2j} + 2\gamma^{a_j-1} & \text{if } j \geq j_0, \\ 1 + e^{-2} & \text{if } j < j_0. \end{cases}$$

Combining the above two inequalities we obtain

$$P\left(\frac{\sqrt{\varepsilon}}{\sqrt{t+1}}\left|\int_0^t b_0(s)dw_i(s)\right| \geq j\right) \leq \begin{cases} 2(e^{-2j} + 2\gamma^{a_j-1}) & \text{if } j \geq j_0, \\ 2(1 + e^{-2}) & \text{if } j < j_0. \end{cases}$$

Then, by (3.8),

$$E\exp\left\{\frac{\sqrt{\varepsilon}}{\sqrt{t+1}}\left|\int_0^t b(s,t)dw_i(s)\right|\right\} \leq C_1 + (e-1)\sum_{j=1}^{\infty} 2e^j(e^{-2j} + 2\gamma^{a_j-1}),$$

where $C_1 = e + 2j_0(e-1)(1+e^{-2})$. Now, we choose ε_0 small enough such that $e\gamma^{1/(8k_1\varepsilon_0)} \leq 1/2$. Then,

$$E\exp\left\{\frac{\sqrt{\varepsilon}}{\sqrt{t+1}}\left|\int_0^t b(s,t_0)dw_i(s)\right|\right\} \leq C_1 + 2 + 4(e-1)\gamma^{-1}.$$

Since t_0 is arbitrary, we may take $t_0 = t$ in the above inequality. Then,

$$E\exp\left\{\frac{\sqrt{\varepsilon}}{\sqrt{t+1}}\left|\int_0^t b(s,t)dw_i(s)\right|\right\} \leq C_1 + 2 + 4(e-1)\gamma^{-1}.$$

Combining this inequality with (3.6), we obtain

$$E\exp\left\{\frac{1}{\sqrt{\varepsilon}(t+1)^{\frac{3}{2}}}\left|\int_0^t (\lambda(s)-\nu)\beta(s)ds\right|\right\} \leq C \text{ for some constant } C.$$

This completes the proof of the lemma. □

COROLLARY 3.1. *In the above lemma, if* $Q^{(1)} = 0$, *i.e.*, $Q = \varepsilon^{-1}Q^{(2)}$, *then we have the following stronger estimate*:

$$E\left[\exp\left\{\frac{1}{\sqrt{\varepsilon}\sqrt{t+1}}\left|\int_0^t (\chi_{\{\alpha(\varepsilon,s)=i\}} - \nu_i)\beta(s)ds\right|\right\}\middle|\mathcal{Z}_t\right] \leq C, \ a.s. \tag{3.9}$$

Proof. If $Q^{(1)} = 0$, then Lemma A.1 becomes

$$|P(\alpha(\varepsilon, t) = i) - \nu_i| \leq Ce^{-k_0 t/\varepsilon};$$

see [20] for details. Thus, equation (3.5) can be replaced by

$$\int_0^t (\lambda(s) - \nu)\beta(s)ds = \varepsilon O(1) + \varepsilon \int_0^t O(1)dw(s).$$

The remaining proof of the corollary follows exactly as the rest of the proof of Lemma 3.1. □

COROLLARY 3.2. *For each $N_0 > 0$ and $0 < \delta < 1/2$, there exist constants $\varepsilon_0 > 0$ and C such that for $0 < \varepsilon \leq \varepsilon_0$, $i \in \mathcal{M}$, $t \geq 0$, and for any $\mathcal{Z}_t$ adapted process $\beta(t)$, $|\beta(t)| \leq N_0$,*

$$P\left(\left|\int_0^t (\chi_{\{\alpha(\varepsilon,s)=i\}} - \nu_i)\beta(s)ds\right| \geq \varepsilon^{\frac{1}{2}-\delta}\right) \leq C\exp\left\{-\frac{1}{\varepsilon^\delta (t+1)^{\frac{3}{2}}}\right\}. \tag{3.10}$$

Moreover, if $Q^{(1)} = 0$, then

$$P\left(\left|\int_0^t (\chi_{\{\alpha(\varepsilon,s)=i\}} - \nu_i)\beta(s)ds\right| \geq \varepsilon^{\frac{1}{2}-\delta}\right) \leq C\exp\left\{-\frac{1}{\varepsilon^\delta \sqrt{t+1}}\right\}. \tag{3.11}$$

Proof. Note in view of Lemma 3.1 that

$$\begin{aligned} & P\left(\left|\int_0^t (\chi_{\{\alpha(\varepsilon,s)=i\}} - \nu_i)\beta(s)ds\right| \geq \varepsilon^{\frac{1}{2}-\delta}\right) \\ = \; & P\left(\exp\left\{\frac{1}{\sqrt{\varepsilon}(t+1)^{\frac{3}{2}}}\left|\int_0^t (\chi_{\{\alpha(\varepsilon,s)=i\}} - \nu_i)\beta(s)ds\right|\right\} \geq \exp\left\{\frac{\varepsilon^{\frac{1}{2}-\delta}}{\sqrt{\varepsilon}(t+1)^{\frac{3}{2}}}\right\}\right) \\ \leq \; & C\exp\left\{-\frac{1}{\varepsilon^\delta (t+1)^{\frac{3}{2}}}\right\}. \end{aligned}$$

This proves (3.10). Similarly, (3.11) follows from Corollary 3.1. □

To conclude this section, we give a corollary that will be needed in §7.

COROLLARY 3.3. *Let $D(\varepsilon) = \{|\int_0^t (\chi_{\{\alpha(\varepsilon,s)=i\}} - \nu_i)\beta(s)ds| \geq \sqrt[4]{\varepsilon}\}$. Then, there exist $k > 0$ and C such that*

$$E\exp\left\{\frac{k}{\sqrt[4]{\varepsilon}}e^{-\rho t/2}\chi_{D(\varepsilon)}\right\} \leq C.$$

Proof. Note that

$$\exp\left\{\frac{k}{\sqrt[4]{\varepsilon}}e^{-\rho t/2}\chi_{D(\varepsilon)}\right\} \leq 1 + \left[\exp\left\{\frac{k}{\sqrt[4]{\varepsilon}}e^{-\rho t/2}\right\}\right]\chi_{D(\varepsilon)}. \tag{3.12}$$

By Corollary 3.2,

$$P(D(\varepsilon)) \leq C\exp\left\{-\frac{1}{\sqrt[4]{\varepsilon}(t+1)^{\frac{3}{2}}}\right\}. \tag{3.13}$$

Combine (3.12) and (3.13) to obtain

$$\begin{aligned} E\exp\left\{\frac{k}{\sqrt[4]{\varepsilon}}e^{-\rho t/2}\chi_{D(\varepsilon)}\right\} & \leq 1 + \exp\left\{\frac{k}{\sqrt[4]{\varepsilon}}e^{-\rho t/2}\right\}P(D(\varepsilon)) \\ & \leq 1 + C\exp\left\{\frac{k}{\sqrt[4]{\varepsilon}}e^{-\rho t/2}\right\}\exp\left\{-\frac{1}{\sqrt[4]{\varepsilon}(t+1)^{\frac{3}{2}}}\right\} \\ & = 1 + C\exp\left[\frac{k}{\sqrt[4]{\varepsilon}}e^{-\rho t/2} - \frac{1}{\sqrt[4]{\varepsilon}(t+1)^{\frac{3}{2}}}\right]. \end{aligned}$$

We now choose $k > 0$ and small enough such that

$$\frac{k}{\sqrt[4]{\varepsilon}} e^{-\rho t/2} \le \frac{1}{\sqrt[4]{\varepsilon}(t+1)^{\frac{3}{2}}}.$$

Then, for k small enough, we have

$$E \exp\left\{ \frac{k}{\sqrt[4]{\varepsilon}} e^{-\rho t/2} \chi_{D(\varepsilon)} \right\} \le 1 + C. \qquad \square$$

4. Asymptotic properties of the value functions. In this section, we study the asymptotic property of the value function $v^{\varepsilon,\sqrt{\varepsilon}}$. The next theorem shows that the value function $v^{\varepsilon,\sqrt{\varepsilon}}$ of $\mathcal{P}^{\varepsilon,\sqrt{\varepsilon}}$ can be approximated by the value function $v^{0,\sqrt{\varepsilon}}$ of $\mathcal{P}^{0,\sqrt{\varepsilon}}$.

THEOREM 4.1. *There exist constants $\varepsilon_0 > 0$ and C such that, for $0 < \varepsilon \le \varepsilon_0$,*

$$|v^{\varepsilon,\sqrt{\varepsilon}} - v^{0,\sqrt{\varepsilon}}| \le C\sqrt{\varepsilon}.$$

Proof. We first show that $v^{\varepsilon,\sqrt{\varepsilon}} \le v^{0,\sqrt{\varepsilon}} + C\sqrt{\varepsilon}$. For any given $\bar{U}_\cdot = (u^0_\cdot, \dots, u^m_\cdot) \in \mathcal{A}^0$, we construct a control $u^\varepsilon_t = \sum_{i=0}^m \chi_{\{\alpha(\varepsilon,t)=i\}} u^i_t$. Then, obviously, $u^\varepsilon_\cdot \in \mathcal{A}^\varepsilon$. Moreover, let $x_\cdot$ and $\bar{x}_\cdot$ denote, respectively, the corresponding states of the systems $\mathcal{P}^{\varepsilon,\sqrt{\varepsilon}}$ and $\mathcal{P}^{0,\sqrt{\varepsilon}}$ with the same initial value a, i.e.,

$$\dot{x}_t = u^\varepsilon_t - z_t, \; x_0 = a, \tag{4.1}$$

$$\dot{\bar{x}}_t = \sum_{i=0}^m \nu_i u^i_t - z_t, \; \bar{x}_0 = a. \tag{4.2}$$

Note that $c(u^\varepsilon_t) = \sum_{i=0}^m \chi_{\{\alpha(\varepsilon,t)=i\}} c(u^i_t)$. Then,

$$\begin{aligned}
&J^{\varepsilon,\sqrt{\varepsilon}}(u^\varepsilon_\cdot)\\
&= \sqrt{\varepsilon} \log E \left\{ \exp\left\{ \frac{1}{\sqrt{\varepsilon}} \int_0^\infty e^{-\rho t}[h(x_t) + c(u^\varepsilon_t)]dt \right\} \right\}\\
&= \sqrt{\varepsilon} \log E \left\{ \exp\left\{ \frac{1}{\sqrt{\varepsilon}} \int_0^\infty e^{-\rho t} \left[h(x_t) + \sum_{i=0}^m \chi_{\{\alpha(\varepsilon,t)=i\}} c(u^i_t) \right] dt \right\} \right\}\\
&= \sqrt{\varepsilon} \log E \left\{ \left(\exp\left\{ \frac{1}{\sqrt{\varepsilon}} \int_0^\infty e^{-\rho t} \left[h(\bar{x}_t) + \sum_{i=0}^m \nu_i c(u^i_t) \right] dt \right\} \right) \cdot I(\varepsilon) \right\},
\end{aligned}$$

where

$$I(\varepsilon) = \exp\left\{ \frac{1}{\sqrt{\varepsilon}} \int_0^\infty e^{-\rho t} \left[h(x_t) - h(\bar{x}_t) + \sum_{i=0}^m (\chi_{\{\alpha(\varepsilon,t)=i\}} - \nu_i) c(u^i_t) \right] dt \right\}.$$

Recall that $\bar{x}_t$ and $\bar{U}_t$ are $\mathcal{Z}_t$ adapted. Thus,

$$\begin{aligned}
&J^{\varepsilon,\sqrt{\varepsilon}}(u^\varepsilon_\cdot)\\
&= \sqrt{\varepsilon} \log E \left\{ E \left[\left(\exp\left\{ \frac{1}{\sqrt{\varepsilon}} \int_0^\infty e^{-\rho t} \left[h(\bar{x}_t) + \sum_{i=0}^m \nu_i c(u^i_t) \right] dt \right\} \right) \cdot I(\varepsilon) \middle| \mathcal{Z}_t \right] \right\}\\
&= \sqrt{\varepsilon} \log E \left\{ \left(\exp\left\{ \frac{1}{\sqrt{\varepsilon}} \int_0^\infty e^{-\rho t} \left[h(\bar{x}_t) + \sum_{i=0}^m \nu_i c(u^i_t) \right] dt \right\} \right) \cdot E[I(\varepsilon)|\mathcal{Z}_t] \right\}.
\end{aligned} \tag{4.3}$$

We now estimate $E[I(\varepsilon)|\mathcal{Z}_t]$. Note that by Assumption (A1) and the fact that $|x_t| + |\bar{x}_t| = O(1+t)$,

$$
\begin{aligned}
|h(x_t) - h(\bar{x}_t)| &\le C_0(1 + |x_t|^{k_h} + |\bar{x}_t|^{k_h})|x_t - \bar{x}_t| \\
&\le C_2(1 + t^{k_h})|x_t - \bar{x}_t| \\
&= C_2(1 + t^{k_h})\left|\int_0^t \sum_{i=0}^m (\chi_{\{\alpha(\varepsilon,s)=i\}} - \nu_i)u_s^i ds\right| \qquad (4.4) \\
&\le C_2(1 + t^{k_h})\sum_{i=0}^m \left|\int_0^t (\chi_{\{\alpha(\varepsilon,s)=i\}} - \nu_i)u_s^i ds\right|.
\end{aligned}
$$

Note also that, by integration by parts,

$$
\begin{aligned}
&\left|\int_0^\infty e^{-\rho t}\sum_{i=0}^m (\chi_{\{\alpha(\varepsilon,t)=i\}} - \nu_i)c(u_t^i)dt\right| \\
&= \left|\rho\int_0^\infty e^{-\rho t}\left(\sum_{i=0}^m \int_0^t (\chi_{\{\alpha(\varepsilon,s)=i\}} - \nu_i)c(u_s^i)ds\right)dt\right| \qquad (4.5) \\
&\le \rho\int_0^\infty e^{-\rho t}\sum_{i=0}^m \left|\int_0^t (\chi_{\{\alpha(\varepsilon,s)=i\}} - \nu_i)c(u_s^i)ds\right| dt.
\end{aligned}
$$

Combining (4.3), (4.4), and (4.5), we obtain

$$
\begin{aligned}
I(\varepsilon) \le \exp\left\{\frac{1}{\sqrt{\varepsilon}}\int_0^\infty e^{-\rho t}\right. & \left[C_2(1 + t^{k_h})\sum_{i=0}^m \left|\int_0^t (\chi_{\{\alpha(\varepsilon,s)=i\}} - \nu_i)u_s^i ds\right|\right. \\
&\left.\left. + \rho\sum_{i=0}^m \left|\int_0^t (\chi_{\{\alpha(\varepsilon,s)=i\}} - \nu_i)c(u_s^i)ds\right|\right]\right\}.
\end{aligned}
$$

By Jensen's inequality and the convexity of e^x,

$$
\exp\int_0^\infty e^{-\rho t}\gamma_t dt \le \frac{\rho}{2}\int_0^\infty e^{-\rho t/2}\exp\left[\frac{2}{\rho}e^{-\rho t/2}\gamma_t\right\} dt \qquad (4.6)
$$

for any function γ_t. Taking

$$
\begin{aligned}
\gamma_t = \frac{1}{\sqrt{\varepsilon}}&\left[C_2(1 + t^{k_h})\sum_{i=0}^m \left|\int_0^t (\chi_{\{\alpha(\varepsilon,s)=i\}} - \nu_i)u_s^i ds\right|\right. \\
&\left. + \rho\sum_{i=0}^m \left|\int_0^t (\chi_{\{\alpha(\varepsilon,s)=i\}} - \nu_i)c(u_s^i)ds\right|\right]
\end{aligned}
$$

in (4.6), we get

$$
\begin{aligned}
I(\varepsilon) \le \frac{\rho}{2}\int_0^\infty e^{-\rho t/2}\exp\left\{\frac{2}{\rho\sqrt{\varepsilon}}e^{-\rho t/2}\right. & \left[C_2(1 + t^{k_h})\sum_{i=0}^m \left|\int_0^t (\chi_{\{\alpha(\varepsilon,s)=i\}} - \nu_i)u_s^i ds\right|\right. \\
&\left.\left. + \rho\sum_{i=0}^m \left|\int_0^t (\chi_{\{\alpha(\varepsilon,s)=i\}} - \nu_i)c(u_s^i)ds\right|\right]\right\} dt.
\end{aligned}
$$

Note that there exists $k_2 > 0$ such that, for all $t \geq 0$,

$$\frac{2}{\rho}e^{-\rho t/2}C_2(1+t^{k_h})(t+1)^{\frac{3}{2}} \leq k_2$$

and

$$\frac{2}{\rho}e^{-\rho t/2}\rho(t+1)^{\frac{3}{2}} \leq k_2.$$

This implies that

(4.7)

$$\begin{aligned}I(\varepsilon) \leq \frac{\rho}{2}\int_0^\infty e^{-\rho t/2}\exp\Bigg\{\frac{k_2}{\sqrt{\varepsilon}(t+1)^{\frac{3}{2}}}\sum_{i=0}^m\Bigg[&\left|\int_0^t(\chi_{\{\alpha(\varepsilon,s)=i\}}-\nu_i)u_s^i ds\right| \\ &+\left|\int_0^t(\chi_{\{\alpha(\varepsilon,s)=i\}}-\nu_i)c(u_s^i)ds\right|\Bigg]\Bigg\}dt.\end{aligned}$$

Moreover, observe that

$$\begin{aligned}&E\Bigg[\exp\Bigg\{\frac{k_2}{\sqrt{\varepsilon}(t+1)^{\frac{3}{2}}}\sum_{i=0}^m\Bigg[\left|\int_0^t(\chi_{\{\alpha(\varepsilon,s)=i\}}-\nu_i)u_s^i ds\right| \\ &\qquad\qquad +\left|\int_0^t(\chi_{\{\alpha(\varepsilon,s)=i\}}-\nu_i)c(u_s^i)ds\right|\Bigg]\Bigg\}dt \mid \mathcal{Z}_t\Bigg] \\ \leq\ &\left[\prod_{i=0}^m E\left[\exp\left\{\frac{2(m+1)k_2}{\sqrt{\varepsilon}(t+1)^{\frac{3}{2}}}\left|\int_0^t(\chi_{\{\alpha(\varepsilon,s)=i\}}-\nu_i)u_s^i ds\right|\right\}\Bigg|\ \mathcal{Z}_t\right]\right]^{\frac{1}{2(m+1)}} \\ &\cdot\left[\prod_{i=0}^m E\left[\exp\left\{\frac{2(m+1)k_2}{\sqrt{\varepsilon}(t+1)^{\frac{3}{2}}}\left|\int_0^t(\chi_{\{\alpha(\varepsilon,s)=i\}}-\nu_i)c(u_s^i)ds\right|\right\}\Bigg|\ \mathcal{Z}_t\right]\right]^{\frac{1}{2(m+1)}} \\ \leq\ &[C_3^{2(m+1)}]^{\frac{1}{2(m+1)}} \quad \text{a.s.} \\ =\ &C_3.\end{aligned}$$

In view of (4.7), we have

$$E[I(\varepsilon)|\mathcal{Z}_t] \leq \frac{\rho}{2}\int_0^\infty e^{-\rho t/2}C_3 dt = C_3.$$

Using this inequality in (4.3) we obtain

$$\begin{aligned}&J^{\varepsilon,\sqrt{\varepsilon}}(u_\cdot^\varepsilon) \\ &\leq \sqrt{\varepsilon}\log E\left[C_3\exp\left\{\frac{1}{\sqrt{\varepsilon}}\int_0^\infty e^{-\rho t}\left[h(\bar{x}_t)+\sum_{i=0}^m \nu_i c(u_t^i)\right]dt\right\}\right] \\ &= \sqrt{\varepsilon}\log C_3 + \sqrt{\varepsilon}\log E\left[\exp\left\{\frac{1}{\sqrt{\varepsilon}}\int_0^\infty e^{-\rho t}\left[h(\bar{x}_t)+\sum_{i=0}^m \nu_i c(u_t^i)\right]dt\right\}\right] \\ &= \sqrt{\varepsilon}\log C_3 + J^{0,\sqrt{\varepsilon}}(\bar{U}_\cdot).\end{aligned} \tag{4.8}$$

Since $\bar{U}. \in \mathcal{A}^0$ is arbitrary we have

$$v^{\varepsilon,\sqrt{\varepsilon}} \leq v^{0,\sqrt{\varepsilon}} + \sqrt{\varepsilon} \log C_3. \tag{4.9}$$

Now it suffices to show the opposite inequality, i.e.,

$$v^{\varepsilon,\sqrt{\varepsilon}} \geq v^{0,\sqrt{\varepsilon}} - C\sqrt{\varepsilon}. \tag{4.10}$$

The proof of this part is similar to that of [20]. We provide the proof in the appendix for the sake of completeness.

Combining (4.9) and (4.10), we conclude

$$|v^{\varepsilon,\sqrt{\varepsilon}} - v^{0,\sqrt{\varepsilon}}| \leq C\sqrt{\varepsilon}. \qquad \square$$

COROLLARY 4.1. *Let $\bar{U}. = (u_.^0, \ldots, u_.^m) \in \mathcal{A}^0$ be an $\sqrt{\varepsilon}$-optimal (stochastic open loop) control for $\mathcal{P}^{0,\sqrt{\varepsilon}}$. Then, $u_t^\varepsilon := \sum_{i=0}^m \chi_{\{\alpha(\varepsilon,t)=i\}} u_t^i \in \mathcal{A}^\varepsilon$ is an asymptotically optimal (stochastic open loop) control for $\mathcal{P}^{\varepsilon,\sqrt{\varepsilon}}$, i.e.,*

$$|J^{\varepsilon,\sqrt{\varepsilon}}(u_.^\varepsilon) - v^{\varepsilon,\sqrt{\varepsilon}}| \leq C\sqrt{\varepsilon}.$$

Proof. By (4.8) and the choice of $\bar{U}. \in \mathcal{A}^0$ ($\sqrt{\varepsilon}$-optimal), we have

$$\begin{aligned} 0 &\leq J^{\varepsilon,\sqrt{\varepsilon}}(u_.^\varepsilon) - v^{\varepsilon,\sqrt{\varepsilon}} \\ &\leq [J^{\varepsilon,\sqrt{\varepsilon}}(u_.^\varepsilon) - J^{0,\sqrt{\varepsilon}}(\bar{U}.)] + [J^{0,\sqrt{\varepsilon}}(\bar{U}.) - v^{0,\sqrt{\varepsilon}}] + [v^{0,\sqrt{\varepsilon}} - v^{\varepsilon,\sqrt{\varepsilon}}] \\ &\leq C\sqrt{\varepsilon} + \sqrt{\varepsilon} + C\sqrt{\varepsilon}. \qquad \square \end{aligned}$$

5. Averaged problems. In this section we show that $\mathcal{P}^{0,\sqrt{\varepsilon}}$ can be approximated by $\mathcal{P}^{0,0}$, and the value function of $\mathcal{P}^{0,0}$ is a viscosity solution to the Isaacs equation of a zero sum, two-player differential game. To simplify the notation, we take $\delta = \sqrt{\varepsilon}$ and consider the following control problem $\mathcal{P}^{0,\delta}$:

$$\mathcal{P}^{0,\delta}: \begin{cases} \text{min.} & J^{0,\delta}(U.) \\ & = \delta \log E\left[\exp\left\{\dfrac{1}{\delta}\displaystyle\int_0^\infty e^{-\rho t}\left[h(x_t) + \sum_{i=0}^m \nu_i c(u_t^i)\right] dt\right\}\right], \\ \text{s.t.} & \dot{x}_t = \displaystyle\sum_{i=0}^m \nu_i u_t^i - z_t,\ x_0 = a,\ U. = (u_.^0, \ldots, u_.^m) \in \mathcal{A}^0, \\ \text{value fn.} & v^{0,\delta} = \displaystyle\inf_{U. \in \mathcal{A}^0} J^{0,\delta}(U.). \end{cases} \tag{5.1}$$

Then, we have the following theorem.

THEOREM 5.1. *$v^{0,\delta}$ is a monotone increasing function of $\delta > 0$ and*

$$\lim_{\delta \to 0} v^{0,\delta} = v^{0,0}.$$

Proof. First of all, note that $J^{0,\delta}(U.) \leq J^{0,0}(U.)$ for all $\delta > 0$ and $U. \in \mathcal{A}^0$. This implies

$$v^{0,\delta} \leq v^{0,0}. \tag{5.2}$$

By Lemma A.4, for each $U_\cdot \in \mathcal{A}^0$,

$$J^{0,\delta}(U_\cdot) \uparrow J^{0,0}(U_\cdot) \text{ as } \delta \downarrow 0. \tag{5.3}$$

Now, for any given $\delta > 0$, let $U_\cdot^\delta = (u_\cdot^{0\delta}, u_\cdot^{1\delta}, \dots, u_\cdot^{m\delta}) \in \mathcal{A}^0$ denote a δ-optimal control for problem $\mathcal{P}^{0,\delta}$, i.e.,

$$v^{0,\delta} \le J^{0,\delta}(U_\cdot^\delta) \le v^{0,\delta} + \delta.$$

Combining these two inequalities and (5.2), we obtain

$$\limsup_{\delta\to 0} J^{0,\delta}(U_\cdot^\delta) \le \limsup_{\delta\to 0} v^{0,\delta} \le v^{0,0}. \tag{5.4}$$

Let $\delta_1 > \delta > 0$. Then, again by Lemma A.4,

$$J^{0,\delta}(U_\cdot^\delta) \ge J^{0,\delta_1}(U_\cdot^\delta). \tag{5.5}$$

We now show that there exists a sequence of δ such that $\{U_\cdot^\delta\}$ converges weakly to a control $U_\cdot^* \in \mathcal{A}^0$. To this end, we consider the following functional space.

$$\begin{aligned}\mathcal{H}_N = \quad & \{f(t) = f(t)(\omega) \in R^1 : \text{ measurable functions on } [0,N] \times \Omega,\\ & f(t) \text{ is } \mathcal{Z}_t \text{ adapted and } \langle f, f\rangle_N < \infty\},\end{aligned}$$

where $N = 1, 2, \dots$, and $\langle f, g\rangle_N := E\int_0^N f(t)g(t)dt$ for any measurable functions f and g on $[0,N] \times \Omega$. Then, it is easy to check that $\mathcal{H}_N$ is a Hilbert space with the inner product $\langle f, g\rangle_N$.

By [26, Thm. 1, p. 126], for each fixed N there exists a sequence of $\delta \to 0$ such that $U_\cdot^\delta \longrightarrow U_\cdot^* \in \mathcal{H}_N$ weakly. By using the Cauchy diagonalization method, we can show that there exists a further subsequence of $\delta \to 0$ (still denoted by δ) and a measurable $U_t^* = (u_t^{0*}, u_t^{1*}, \dots, u_t^{m*})$ for all $t \ge 0$ such that $U_\cdot^* \in \mathcal{H}_N$ for each N and

$$U_\cdot^\delta \longrightarrow U_\cdot^* \in \mathcal{H}_N \text{ weakly on } \mathcal{H}_N \text{ for each } N.$$

This means, in particular, for all bounded measurable and $\mathcal{Z}_t$ adapted functions $f(t)$ on $[0, \infty) \times \Omega$ and for any $t \ge 0$,

$$E\int_0^t f(s)U_s^\delta ds \longrightarrow E\int_0^t f(s)U_s^* ds. \tag{5.6}$$

We now show that $U_\cdot^* \in \mathcal{A}^0$. Since $U_\cdot^* \in \mathcal{H}_N$, U_t^* is $\mathcal{Z}_t$ adapted. It suffices to show that $u_s^{i*} \in \mathcal{U}(i)$ a.s. If not, then for some $\eta > 0$, the set $D_0 = \{(s,\omega) : p \cdot u_s^{i*} > i + \eta\}$ has a positive measure, i.e., $(l \times P)(D_0) > 0$, where l denotes the Lebesgue measure. Now, let $f(t) = E[\chi_{D_0}|\mathcal{Z}_t]p$, where $p \in R^{n_0}$ is the vector used in defining $\mathcal{U}(i)$. Then,

$$E\int_0^t f(s)p \cdot u_s^{i\delta} ds = E\int_0^t \chi_{D_0} p \cdot u_s^{i\delta} ds \to E\int_0^t \chi_{D_0} p \cdot u_s^{i*} ds, \text{ as } \delta \to 0. \tag{5.7}$$

On the other hand,

$$E\int_0^t \chi_{D_0} p \cdot u_s^{i\delta} ds \le iE\int_0^t \chi_{D_0} ds = i(l \times P)(D_0)$$

and

$$E\int_0^t \chi_{D_0} p\cdot u_s^{i*} ds \geq (i+\eta) E\int_0^t \chi_{D_0} ds = (i+\eta)(l\times P)(D_0).$$

This contradicts (5.7). Hence, $U^*_\cdot \in \mathcal{A}^0$.

Let x_t^* denote the trajectory of $\mathcal{P}^{0,0}$ under control $U^*_\cdot$. Recall the convexities of h and c to obtain

$$\begin{aligned} & h(x_t) \geq h(x_t^*) + h_{x^+}(x_t^*)(x_t - x_t^*) \\ \text{and}\quad & c(u_t^{i\delta}) \geq c(u_t^{i*}) + c_{u^+}(u_t^{i*})(u_t^{i\delta} - u_t^{i*}), \end{aligned}$$

where h_{x^+} and c_{u^+} denote the right-hand derivatives of h and c. Then,

$$\begin{aligned} J^{0,\delta_1}(U^\delta_\cdot) &= \delta_1 \log E \exp\left\{ \frac{1}{\delta_1}\int_0^\infty e^{-\rho t}\left[h(x_t) + \sum_{i=0}^m \nu_i c(u_t^{i\delta})\right] dt\right\} \\ &\geq \delta_1 \log E\left[\left(\exp\left\{ \frac{1}{\delta_1}\int_0^\infty e^{-\rho t}\left[h(x_t^*) + \sum_{i=0}^m \nu_i c(u_t^{i*})\right] dt\right\}\right) R(\delta,\delta_1)\right], \end{aligned}$$

where

$$R(\delta,\delta_1) = \exp \frac{1}{\delta_1}\int_0^\infty e^{-\rho t}\left[h_{x^+}(x_t^*)(x_t - x_t^*) + \sum_{i=0}^m \nu_i c_{u^+}(u_t^{i*})(u_t^{i\delta} - u_t^{i*})\right] dt.$$

For each $\eta > 0$, let

$$D = \left\{ \int_0^\infty e^{-\rho t}\left[h(x_t^*) + \sum_{i=0}^m \nu_i c(u_t^{i*})\right] dt \geq J^{0,0}(U^*_\cdot) - \eta\right\}.$$

Then, $P(D) > 0$. Moreover,

$$\begin{aligned} J^{0,\delta_1}(U^\delta_\cdot) &\geq \delta_1 \log E\left[\chi_D \left[\exp\frac{1}{\delta_1}(J^{0,0}(U^*_\cdot) - \eta)\right] R(\delta,\delta_1)\right] \\ &= J^{0,0}(U^*_\cdot) - \eta + \delta_1 \log E\{\chi_D R(\delta,\delta_1)\}. \end{aligned}$$

We now show that

$$\liminf_{\delta\to 0}\, [\log E\{\chi_D R(\delta,\delta_1)\}] \geq \log P(D). \tag{5.8}$$

To this end, we observe in view of assumption (A1) that

$$\begin{aligned} & h_{x^+}(x_t^*) = O(1 + |x_t^*|^{k_h}) = O(1 + t^{k_h}), \\ & x_t - x_t^* = O(1+t), \\ \text{and}\quad & c_{u^+}(u_t^{i*})(u_t^{i\delta} - u_t^{i*}) = O(1). \end{aligned}$$

These imply, for any N,

$$\begin{aligned} & E\chi_D \int_N^\infty e^{-\rho t}\left[h_{x^+}(x_t^*)(x_t - x_t^*) + \sum_{i=0}^m \nu_i c_{u^+}(u_t^{i*})(u_t^{i\delta} - u_t^{i*})\right] dt \\ & \leq \int_N^\infty e^{-\rho t} O(1)[(1+t^{k_h})(1+t) + 1] dt \\ & \to 0 \quad \text{as } N\to\infty. \end{aligned} \tag{5.9}$$

Moreover, in view of the convergence of $U_\cdot^\delta \to U_\cdot^*$,

$$(5.10)\quad \begin{aligned} &E\chi_D \int_0^N e^{-\rho t}\left[h_{x^+}(x_t^*)(x_t - x_t^*) + \sum_{i=0}^m \nu_i c_{u^+}(u_t^{i*})(u_t^{i\delta} - u_t^{i*})\right] dt \\ &= E\int_0^N e^{-\rho t}(E[\chi_D|\mathcal{Z}_t])\left[h_{x^+}(x_t^*)(x_t - x_t^*)\right. \\ &\qquad \left. + \sum_{i=0}^m \nu_i c_{u^+}(u_t^{i*})(u_t^{i\delta} - u_t^{i*})\right] dt \to 0, \text{ as } \delta \to 0. \end{aligned}$$

Combining (5.9) and (5.10), we obtain

$$E\left[\chi_D \int_0^\infty e^{-\rho t}\left[h_{x^+}(x_t^*)(x_t - x_t^*) + \sum_{i=0}^m \nu_i c_{u^+}(u_t^{i*})(u_t^{i\delta} - u_t^{i*})\right]\right] dt \to 0.$$

Now, by Jensen's inequality, we have

$$\begin{aligned} &E\{\chi_D R(\delta, \delta_1)\} \\ &\geq P(D) \exp \frac{1}{\delta_1 P(D)} E\left[\chi_D \int_0^\infty e^{-\rho t}\left[h_{x^+}(x_t^*)(x_t - x_t^*)\right.\right. \\ &\qquad \left.\left. + \sum_{i=0}^m \nu_i c_{u^+}(u_t^{i*})(u_t^{i\delta} - u_t^{i*})\right] dt\right]. \end{aligned}$$

Then, (5.8) follows from the fact that

$$\begin{aligned} &\liminf_{\delta\to 0}\left[\log E\{\chi_D R(\delta, \delta_1)\}\right] \\ &\geq \log P(D) + \frac{1}{\delta_1 P(D)} E\left[\chi_D \int_0^\infty e^{-\rho t}\left[h_{x^+}(x_t^*)(x_t - x_t^*)\right.\right. \\ &\qquad \left.\left. + \sum_{i=0}^m \nu_i c_{u^+}(u_t^{i*})(u_t^{i\delta} - u_t^{i*})\right] dt\right] \to \log P(D) \end{aligned}$$

as $\delta \to 0$. Therefore,

$$\liminf_{\delta\to 0} J^{0,\delta_1}(U_\cdot^\delta) \geq J^{0,0}(U_\cdot^*) - \eta + \delta_1 \log P(D).$$

Since η is arbitrary, we have

$$\liminf_{\delta\to 0} J^{0,\delta_1}(U_\cdot^\delta) \geq J^{0,0}(U_\cdot^*) + \delta_1 \log P(D).$$

In view of the monotonicity of $J^{0,\delta}$, we obtain

$$\liminf_{\delta\to 0} J^{0,\delta}(U_\cdot^\delta) \geq \liminf_{\delta\to 0} J^{0,\delta_1}(U_\cdot^\delta) \geq J^{0,0}(U_\cdot^*) + \delta_1 \log P(D) \geq v^{0,0} + \delta_1 \log P(D).$$

Sending $\delta_1 \to 0$, we have $\liminf_{\delta\to 0} J^{0,\delta}(U_\cdot^\delta) \geq v^{0,0}$. Combining this inequality and (5.4), we obtain

$$\lim_{\delta\to 0} J^{0,\delta}(U_\cdot^\delta) = v^{0,0}.$$

Finally, since $J^{0,\delta}(U^{\delta}_{\cdot}) - \delta \le v^{0,0} \le J^{0,\delta}(U^{\delta}_{\cdot})$,

$$\lim_{\delta\to 0} v^{0,\delta} = v^{0,0}.$$

This completes the proof. □

Remark 5.1. Theorem 5.1 says that the risk-sensitive control problem $\mathcal{P}^{0,\delta}$ can be approximated by the limiting problem $\mathcal{P}^{0,0}$. This result can be easily extended to any linear systems with convex costs and compact control spaces.

Remark 5.2. A related result on the connection of risk-sensitive control and differential games is given in Barron and Jensen [2] in the context of financial economics.

In the next theorem, we will show that the value function of $\mathcal{P}^{0,0}$ satisfies the Isaacs equation of a zero-sum, two-player differential game. The advantage of considering $\mathcal{P}^{0,0}$ instead of $\mathcal{P}^{0,\delta}$ is that the theory of differential games (cf. Basar and Bernhard [1]) can be used to obtain an optimal control for $\mathcal{P}^{0,0}$, and such optimal control is independent of the choice of $\delta(=\sqrt{\varepsilon})$.

We write $v^{0,0}(x)$ as the value function of $\mathcal{P}^{0,0}$ with the initial value $x_0 = x$. Note that $\|\xi\|_\infty = \inf_{P(F)=0} \sup_{\omega\in\Omega-F} |\xi(\omega)|$ for any random variable ξ. It is not difficult to show that $0 \le v^{0,0}(x) \le C(1+|x|^{k_h})$ and $v^{0,0}(x)$ is convex and locally Lipschitz; see [20] for details of the proof.

Let $\Gamma_u = \{U = (u^0, u^1, \ldots, u^m) \in R^{n_0\times(m+1)}$ such that $u^i \in \mathcal{U}(i)\}$, and let Γ_z denote a compact subset of R^{n_0}. We consider functions $z_t \in \Gamma_z$ $(t \ge 0)$ that are right continuous and have left-hand limits. Let $\mathcal{Z}$ denote the metric space of such functions that is equipped with the Skorohod topology $d(\cdot,\cdot)$.

We make another assumption on the probability distribution of the demand process $z_t = z_t(\omega)$.

(A4) $z_{\cdot}(\omega) \in \Gamma_z$ a.s., and for each $z_{\cdot}^0 \in \mathcal{Z}$ and any $\delta_0 > 0$,

$$P(d(z_{\cdot}(\omega), z_{\cdot}^0) \le \delta_0) > 0.$$

Assumption (A4) says that the probability of $z_{\cdot}$ is continuously distributed on $\mathcal{Z}$. An immediate example that satisfies assumption (A4) can be given as a finite state Markov chain.

Note that under assumption (A4),

$$\operatorname*{ess\,sup}_{\omega\in\Omega} F(z_{\cdot}(\omega)) = \sup_{z_{\cdot}\in\mathcal{Z}} F(z_{\cdot})$$

for any continuous function $F(z_{\cdot})$ on $\mathcal{Z}$. It can be shown that

$$\int_0^\infty e^{-\rho t}\left[h(x_t) + \sum_{i=0}^m \nu_i c(u_t^i)\right] dt$$

is continuous on $\mathcal{Z}$ for each given $U_{\cdot} \in \mathcal{A}^0$. Therefore,

$$J^{0,0}(U_{\cdot}) = \sup_{z_{\cdot}\in\mathcal{Z}} \int_0^\infty e^{-\rho t}\left[h(x_t) + \sum_{i=0}^m \nu_i c(u_t^i)\right] dt.$$

THEOREM 5.2. *Assume* (A1) *and* (A4). *Then,* $v^{0,0}(x)$ *is the only viscosity solu-*

tion (*see Fleming and Soner* [10] *for definitions*) *to the following Isaacs equation*:

$$
(5.11)\qquad
\begin{aligned}
\rho v^{0,0}(x) &= \min_{U\in\Gamma_u}\max_{z\in\Gamma_z}\left\{\left(\sum_{i=0}^{m}\nu_i u^i - z\right)v_x^{0,0}(x) + h(x) + \sum_{i=0}^{m}\nu_i c(u^i)\right\}\\
&= \max_{z\in\Gamma_z}\min_{U\in\Gamma_u}\left\{\left(\sum_{i=0}^{m}\nu_i u^i - z\right)v_x^{0,0}(x) + h(x) + \sum_{i=0}^{m}\nu_i c(u^i)\right\}.
\end{aligned}
$$

Proof. First of all, note that

$$
\min_{U\in\Gamma_u}\max_{z\in\Gamma_z}\left[\left(\sum_{i=0}^{m}\nu_i u^i - z\right)p + h(x) + \sum_{i=0}^{m}\nu_i c(u^i)\right]
$$

is Lipschitz in p. The uniqueness of viscosity solution follows from Ishii [13, Thm. 1.6, p. 731]. It remains to show that $v^{0,0}$ is a viscosity solution to (5.11). The proof of this part is similar to that of Evans and Souganidis [6]. We only sketch the proof below. Let $x_0 \in R^{n_0}$ and let $\phi(x) \in C^1$ such that $v(x) - \phi(x)$ has a local maximum at $x = x_0$. Then, by using the Dynkin formula, we can show that for $\theta > 0$ small enough,

$$
(5.12)\qquad e^{-\rho\theta}v^{0,0}(x_\theta)-v^{0,0}(x_0) \le \int_0^\theta e^{-\rho t}\left[-\rho v^{0,0}(x_t)+\left(\sum_{i=0}^{m}\nu_i u_t^i - z_t\right)\phi_x(x_t)\right]dt.
$$

On the other hand, it can be shown that

$$
(5.13)\qquad v^{0,0}(x) \le \left\|\int_0^\theta e^{-\rho t}\left[h(x_t)+\sum_{i=0}^{m}\nu_i c(u_t^i)\right]dt\right\|_\infty + e^{-\rho\theta}\|v^{0,0}(x_\theta)\|_\infty.
$$

Combine (5.12) and (5.13) to obtain

$$
(5.14)\qquad
\begin{aligned}
&\left\|\int_0^\theta e^{-\rho t}\left[h(x_t)+\sum_{i=0}^{m}\nu_i c(u_t^i)\right]dt\right\|_\infty\\
&+\sup_{z.}\int_0^\theta e^{-\rho t}\left[-\rho v^{0,0}(x_t)+\left(\sum_{i=0}^{m}\nu_i u_t^i - z_t\right)\phi_x(x_t)\right]dt \ge 0.
\end{aligned}
$$

Let $\bar{x}_t = x_0+\int_0^t(\sum_{i=0}^m \nu_i u_s^i - z_0)ds$. Then $|x_t-\bar{x}_t| \le Ct$ for some C. This implies, by assumption (A1) as $\theta \to 0$,

$$
\frac{1}{\theta}\left\{\left\|\int_0^\theta e^{-\rho t}\left[h(x_t)+\sum_{i=0}^{m}\nu_i c(u_t^i)\right]dt\right\|_\infty - \int_0^\theta e^{-\rho t}\left[h(\bar{x}_t)+\sum_{i=0}^{m}\nu_i c(u_t^i)\right]dt\right\} \to 0.
$$

Using this fact together with (5.14) we can show

$$
\rho v^{0,0}(x_0) \le \min_{U\in\Gamma_u}\max_{z\in\Gamma_z}\left[\left(\sum_{i=0}^{m}\nu_i u^i - z\right)\phi_x(x_0) + h(x_0) + \sum_{i=0}^{m}\nu_i c(u^i)\right].
$$

Thus, $v^{0,0}$ is a viscosity subsolution.

Similarly, we can show $v^{0,0}$ is a viscosity supersolution. Therefore, $v^{0,0}$ is a viscosity solution to the Isaacs equation (5.11). □

6. Asymptotically optimal controls. In this section we state and prove our main results, which are summarized in the next theorem.

THEOREM 6.1. *Under assumptions* (A1)–(A3), *the following hold.*

(1) (convergence).

$$\lim_{\varepsilon\to 0} v^{\varepsilon,\sqrt{\varepsilon}} = v^{0,0}. \tag{6.1}$$

(2) (stochastic open loop control). *Let* $U_{\cdot} = (u_{\cdot}^0, \dots, u_{\cdot}^m) \in \mathcal{A}^0$ *denote a stochastic open loop* ε'*-optimal control for* $\mathcal{P}^{0,0}$, *i.e.,*

$$0 \le J^{0,0}(U_{\cdot}) - v^{0,0} \le \varepsilon'.$$

Let $u_t^\varepsilon = \sum_{i=0}^m \chi_{\{\alpha(\varepsilon,t)=i\}} u_t^i$. *Then,* $u_{\cdot}^\varepsilon \in \mathcal{A}^\varepsilon$ *and*

$$\limsup_{\varepsilon\to 0} |J^{\varepsilon,\sqrt{\varepsilon}}(u_{\cdot}^\varepsilon) - v^{\varepsilon,\sqrt{\varepsilon}}| \le \varepsilon'. \tag{6.2}$$

(3) (feedback control). *Let* $U_{\cdot} = U(z_{\cdot}, x_{\cdot}) = (u^0(z_{\cdot}, x_{\cdot}), \dots, u^m(z_{\cdot}, x_{\cdot}))$ *denote a feedback* ε'*-optimal control for* $\mathcal{P}^{0,0}$, *i.e.,* $0 \le J^{0,0}(U_{\cdot}) - v^{0,0} \le \varepsilon'$. *Let*

$$u_{\cdot}^\varepsilon = u^\varepsilon(\alpha(\varepsilon,\cdot), z_{\cdot}, x_{\cdot}) = \sum_{i=0}^m \chi_{\{\alpha(\varepsilon,\cdot)=i\}} u^i(z_{\cdot}, x_{\cdot}).$$

Assume that $U(z,x)$ *is locally Lipschitz in* x, *i.e., for some* $k_5 > 0$,

$$|U(z,x) - U(z,x')| \le C(1 + |x|^{k_5} + |x'|^{k_5})|x - x'|.$$

Then, $u_{\cdot}^\varepsilon = u^\varepsilon(\alpha(\varepsilon,\cdot), z_{\cdot}, x_{\cdot}) \in \mathcal{A}^\varepsilon$ *and*

$$\limsup_{\varepsilon\to 0} |J^{\varepsilon,\sqrt{\varepsilon}}(u_{\cdot}^\varepsilon) - v^{\varepsilon,\sqrt{\varepsilon}}| \le \varepsilon'. \tag{6.3}$$

Proof. Note that, by Theorems 4.1 and 5.1,

$$\begin{aligned} |v^{\varepsilon,\sqrt{\varepsilon}} - v^{0,0}| &\le |v^{\varepsilon,\sqrt{\varepsilon}} - v^{0,\sqrt{\varepsilon}}| + |v^{0,\sqrt{\varepsilon}} - v^{0,0}| \\ &\le C\sqrt{\varepsilon} + |v^{0,\sqrt{\varepsilon}} - v^{0,0}| \to 0 \text{ as } \varepsilon \to 0. \end{aligned}$$

Thus, $\lim_{\varepsilon\to 0} v^{\varepsilon,\sqrt{\varepsilon}} = v^{0,0}$. (1) holds.

To show (2), we first observe that

$$\begin{aligned} &|J^{\varepsilon,\sqrt{\varepsilon}}(u_{\cdot}^\varepsilon) - v^{\varepsilon,\sqrt{\varepsilon}}| \\ &= J^{\varepsilon,\sqrt{\varepsilon}}(u_{\cdot}^\varepsilon) - v^{\varepsilon,\sqrt{\varepsilon}} \\ &= (J^{\varepsilon,\sqrt{\varepsilon}}(u_{\cdot}^\varepsilon) - J^{0,\sqrt{\varepsilon}}(U_{\cdot})) + (J^{0,\sqrt{\varepsilon}}(U_{\cdot}) - J^{0,0}(U_{\cdot})) \\ &\quad + (J^{0,0}(U_{\cdot}) - v^{0,0}) + (v^{0,0} - v^{\varepsilon,\sqrt{\varepsilon}}). \end{aligned} \tag{6.4}$$

Note that

$$\begin{aligned} &J^{\varepsilon,\sqrt{\varepsilon}}(u_{\cdot}^\varepsilon) - J^{0,\sqrt{\varepsilon}}(U_{\cdot}) \le C\sqrt{\varepsilon} \text{ (by (4.8))}, \\ &J^{0,\sqrt{\varepsilon}}(U_{\cdot}) - J^{0,0}(U_{\cdot}) \to 0 \text{ as } \varepsilon \to 0 \text{ (by Lemma A.4)}, \\ \text{and}\quad &v^{\varepsilon,\sqrt{\varepsilon}} - v^{0,0} \to 0 \text{ (by (1) of this theorem)}. \end{aligned}$$

Then, (6.4) yields

$$\begin{aligned}&\limsup_{\varepsilon\to 0} |J^{\varepsilon,\sqrt{\varepsilon}}(u^\varepsilon_\cdot) - v^{\varepsilon,\sqrt{\varepsilon}}|\\ &\quad\le \limsup_{\varepsilon\to 0}[C\sqrt{\varepsilon} + (J^{0,\sqrt{\varepsilon}}(U_\cdot) - J^{0,0}(U_\cdot)) + \varepsilon' + (v^{0,0} - v^{\varepsilon,\sqrt{\varepsilon}})] = \varepsilon'.\end{aligned}$$

This implies (2).

We now show (3). In view of (6.4), it suffices to show that

$$\limsup_{\varepsilon\to 0}[J^{\varepsilon,\sqrt{\varepsilon}}(u^\varepsilon(\alpha(\varepsilon,\cdot),z_\cdot,x_\cdot)) - J^{0,\sqrt{\varepsilon}}(U(z_\cdot,\bar{x}_\cdot))] \le 0, \tag{6.5}$$

where x_t and $\bar{x}_t$ are the trajectories of systems $\mathcal{P}^{\varepsilon,\sqrt{\varepsilon}}$ and $\mathcal{P}^{0,\sqrt{\varepsilon}}$ under the controls $u^\varepsilon_\cdot = u^\varepsilon(\alpha(\varepsilon,\cdot),z_\cdot,x_\cdot)$ and $U_\cdot = U(z_\cdot,\bar{x}_\cdot)$, respectively. We can show, by the local Lipschitz property of $U(z,\cdot)$, that, for some k_6 and C_4,

$$|x_t - \bar{x}_t| \le R(t) + C_4\int_0^t (1+s^{k_6})|x_s - \bar{x}_s|ds, \tag{6.6}$$

where $R(t) = |\int_0^t \sum_{i=0}^m (\chi_{\{\alpha(\varepsilon,s)=i\}} - \nu_i)u^i(z_s,\bar{x}_s)ds|$. By Gronwall's inequality,

$$|x_t - \bar{x}_t| \le R(t) + C_4\int_0^t (1+s^{k_6})R(s)\exp\left\{C_4\int_s^t (1+r^{k_6})dr\right\}ds.$$

Let T be any fixed number. Then, for $0\le t\le T$ and some C_5,

$$|x_t - \bar{x}_t| \le R(t) + C_5\int_0^t R(s)ds.$$

Note that $|x_t| \le C_6(1+t)$ for some constant C_6. Thus, for all $t\ge 0$,

$$h(x_t) \le C_0(1+|x_t|^{k_h}) \le C_7(1+t^{k_h}).$$

We have

$$\begin{aligned}\int_0^\infty e^{-\rho t}h(x_t)dt \le\ & \int_0^\infty e^{-\rho t}h(\bar{x}_t)dt + C_8\int_0^T e^{-\rho t}(1+t^{k_h})|x_t - \bar{x}_t|dt\\ & + C_7\int_T^\infty e^{-\rho t}(1+t^{k_h})dt.\end{aligned} \tag{6.7}$$

Now, (6.6) implies

$$\begin{aligned}&C_8\int_0^T e^{-\rho t}(1+t^{k_h})|x_t - \bar{x}_t|dt\\ &\le C_9\int_0^T e^{-\rho t}\left[R(t) + C_5\int_0^t R(s)ds\right]dt\\ &= C_{10}\left\{\int_0^T e^{-\rho t}R(t)dt + \frac{1}{\rho}\int_0^T e^{-\rho t}R(t)dt - \frac{e^{-\rho T}}{\rho}\int_0^T R(t)dt\right\}\\ &\le C_{11}\int_0^T e^{-\rho t}R(t)dt.\end{aligned} \tag{6.8}$$

Combine (6.7) and (6.8) to obtain

$$\begin{aligned}(6.9)\qquad &\int_0^\infty e^{-\rho t}h(x_t)dt\\ &\leq \int_0^\infty e^{-\rho t}h(\bar{x}_t)dt + C_{11}\int_0^T e^{-\rho t}R(t)dt + C_7\int_T^\infty e^{-\rho t}(1+t^{k_h})dt.\end{aligned}$$

Let $R_1(\varepsilon) = \int_0^\infty e^{-\rho t}\sum_{i=0}^m (\chi_{\{\alpha(\varepsilon,t)=i\}} - \nu_i)c(u^i(z_t,\bar{x}_t))dt$. Then,

$$\begin{aligned}&\int_0^\infty e^{-\rho t}c(u^\varepsilon(\alpha(\varepsilon,t),z_t,x_t))dt\\ &= \int_0^\infty e^{-\rho t}\sum_{i=0}^m \chi_{\{\alpha(\varepsilon,t)=i\}}c(u^i(z_t,x_t))dt\\ &\leq \int_0^\infty e^{-\rho t}\sum_{i=0}^m \nu_i c(u^i(z_t,\bar{x}_t))dt + R_1(\varepsilon)\\ &+\int_0^T e^{-\rho t}\sum_{i=0}^m \chi_{\{\alpha(\varepsilon,t)=i\}}[c(u^i(z_t,x_t)) - c(u^i(z_t,\bar{x}_t))]dt \quad + C_{11}\int_T^\infty e^{-\rho t}dt\end{aligned}$$

for any fixed T and some constant C_{11}. Note, by the local Lipschitz property of U., that

$$\begin{aligned}\int_0^T e^{-\rho t}\sum_{i=0}^m \chi_{\{\alpha(\varepsilon,t)=i\}}[c(u^i(z_t,x_t)) - c(u^i(z_t,\bar{x}_t))]dt &\leq C_{13}\int_0^T e^{-\rho t}|x_t-\bar{x}_t|dt\\ &\leq C_{14}\int_0^T e^{-\rho t}R(t)dt.\end{aligned}$$

This yields

$$\begin{aligned}(6.10)\quad \int_0^\infty e^{-\rho t}c(u^\varepsilon(\alpha(\varepsilon,t),z_t,x_t))dt &\leq \int_0^\infty e^{-\rho t}\sum_{i=0}^m \nu_i c(u^i(z_t,\bar{x}_t))dt + R_1(\varepsilon)\\ &+C_{14}\int_0^T e^{-\rho t}R(t)dt + C_{12}\int_T^\infty e^{-\rho t}dt.\end{aligned}$$

Let

$$\begin{aligned}I_2(\varepsilon) = \exp\frac{1}{\sqrt{\varepsilon}}\Bigg[&(C_{11}+C_{14})\int_0^T e^{-\rho t}R(t)dt\\ &+(C_7+C_{12})\int_T^\infty e^{-\rho t}(1+t^{k_h})dt + R_1(\varepsilon)\Bigg].\end{aligned}$$

Then, by combining (6.9) and (6.10), we have

$$\begin{aligned}&\int_0^\infty e^{-\rho t}h(x_t)dt + \int_0^\infty e^{-\rho t}c(u^\varepsilon(\alpha(\varepsilon,t),z_t,x_t))dt\\ &\leq I_2(\varepsilon) + \int_0^\infty e^{-\rho t}h(\bar{x}_t)dt + \int_0^\infty e^{-\rho t}\sum_{i=0}^m \nu_i c(u^i(z_t,\bar{x}_t))dt.\end{aligned}$$

Moreover,

$$
\begin{aligned}
E[I_2(\varepsilon)|\mathcal{Z}_t] \le & \exp\left\{\frac{(C_7+C_{12})}{\sqrt{\varepsilon}}\int_T^\infty e^{-\rho t}(1+t^{k_h})dt\right\} \\
& \cdot\left(E\left[\exp\left\{\frac{2(C_{11}+C_{14})}{\sqrt{\varepsilon}}\int_0^T e^{-\rho t}R(t)dt\right\}\middle|\mathcal{Z}_t\right]\right)^{\frac{1}{2}} \\
& \cdot\left(E\left[\exp\left\{\frac{2}{\sqrt{\varepsilon}}R_1(\varepsilon)\right\}\middle|\mathcal{Z}_t\right]\right)^{\frac{1}{2}}.
\end{aligned}
$$

Similarly, as in §4, we can show that, for some C_{15},

$$
\left(E\left[\exp\left\{\frac{2(C_{11}+C_{14})}{\sqrt{\varepsilon}}\int_0^T e^{-\rho t}R(t)dt\right\}\middle|\mathcal{Z}_t\right]\right)^{\frac{1}{2}} \le C_{15},
$$
$$
\text{and } \left(E\left[\exp\left\{\frac{2}{\sqrt{\varepsilon}}R_1(\varepsilon)\right\}\middle|\mathcal{Z}_t\right]\right)^{\frac{1}{2}} \le C_{15}.
$$

Therefore,

$$
E[I_2(\varepsilon)|\mathcal{Z}_t] \le C_{15}^2 \exp\frac{(C_7+C_{12})}{\sqrt{\varepsilon}}\int_T^\infty e^{-\rho t}(1+t^{k_h})dt.
$$

Now, we have

$$
\begin{aligned}
& J^{\varepsilon,\sqrt{\varepsilon}}(u^\varepsilon_\cdot) \\
&= \sqrt{\varepsilon}\log E\exp\left\{\frac{1}{\sqrt{\varepsilon}}\int_0^\infty e^{-\rho t}[h(x_t)+c(u^\varepsilon_t)]dt\right\} \\
&\le \sqrt{\varepsilon}\log E\left\{\left(\exp\left\{\frac{1}{\sqrt{\varepsilon}}\int_0^\infty e^{-\rho t}\left(h(\bar{x}_t)+\sum_{i=0}^m \nu_i c(u^i_t)\right)dt\right\}\right)\cdot I_2(\varepsilon)\right\} \\
&= \sqrt{\varepsilon}\log E\left\{\left(\exp\left\{\frac{1}{\sqrt{\varepsilon}}\int_0^\infty e^{-\rho t}\left(h(\bar{x}_t)+\sum_{i=0}^m \nu_i c(u^i_t)\right)dt\right\}\right)\cdot E[I_2(\varepsilon)|\mathcal{Z}_t]\right\} \\
&\le \sqrt{\varepsilon}\log E\left\{\exp\left\{\frac{1}{\sqrt{\varepsilon}}\int_0^\infty e^{-\rho t}\left(h(\bar{x}_t)+\sum_{i=0}^m \nu_i c(u^i_t)\right)dt\right\}\right\} \\
&\quad + \sqrt{\varepsilon}\log C_{15}^2 + (C_7+C_{12})\int_T^\infty e^{-\rho t}(1+t^{k_h})dt \\
&= J^{0,\sqrt{\varepsilon}}(U_\cdot) + \sqrt{\varepsilon}\log C_{15}^2 + (C_7+C_{12})\int_T^\infty e^{-\rho t}(1+t^{k_h})dt.
\end{aligned}
$$

Thus, for any fixed $0 < T < \infty$,

$$
\begin{aligned}
& \limsup_{\varepsilon\to 0}[J^{\varepsilon,\sqrt{\varepsilon}}(u^\varepsilon(\alpha(\varepsilon,\cdot),z_\cdot,x_\cdot)) - J^{0,\sqrt{\varepsilon}}(U(z_\cdot,\bar{x}_\cdot))] \\
& \le \ (C_7+C_{12})\int_T^\infty e^{-\rho t}(1+t^{k_h})dt
\end{aligned}
$$

which converges to 0 as $T \to \infty$. Therefore,

$$
\limsup_{\varepsilon\to 0}[J^{\varepsilon,\sqrt{\varepsilon}}(u^\varepsilon(\alpha(\varepsilon,\cdot),z_\cdot,x_\cdot)) - J^{0,\sqrt{\varepsilon}}(U(z_\cdot,x_\cdot))] \le 0.
$$

This completes the proof. □

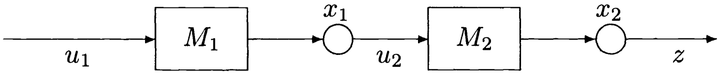


FIG. 1. *A manufacturing system with two machines in tandem.*

7. Extensions to problems with state constraints. In this section, we extend the results in the previous sections to incorporate problems of manufacturing systems with state constraints. These constraints are inherent in systems with internal buffers, i.e., buffers between any two machines, as the inventories in each of them cannot be allowed to become negative. These state constraints cause great difficulty to the analysis. In fact, if we follow Theorem 6.1 to construct a control for the original problem from a near optimal control for the limiting problem, then we find that the constructed control may not even be admissible, i.e., the corresponding trajectory may not satisfy the state constraints. In [22], a method of "lifting" and "modification" is introduced to overcome the difficulty. The basic idea behind it is as follows: First, we modify a given near optimal control of the limiting problem by increasing the inventory in the buffer by a small amount. Then, we use this resulting control to construct a "control" for the original problem in the same way as in Theorem 6.1. The constructed control is not necessarily admissible for the original problem, so we modify it whenever the state constraints are violated.

In this section, we only consider the risk-sensitive control problem of manufacturing systems with two tandem machines and an internal buffer. For risk-sensitive controls with more general manufacturing systems, the results can be extended similarly; see Remark 7.2.

We consider a manufacturing system with two machines arranged in tandem (see Fig. 1). Each machine has a finite number of states resulting in a finite state machine capacity process denoted by $\alpha(\varepsilon,t) = (\alpha_1(\varepsilon,t), \alpha_2(\varepsilon,t))$.

We use $u_1(t)$ and $u_2(t)$ to denote the input rates to the first and second machines, respectively. We denote the number of parts in the buffer between the first and second machine as $x_1(t) \geq 0$ and the difference between cumulative production and cumulative demand, called surplus, as $x_2(t)$. Then, the system can be written as follows:

$$\begin{cases} \dot{x}_1(t) = u_1(t) - u_2(t), & x_1(0) = a_1, \\ \dot{x}_2(t) = u_2(t) - z_t, & x_2(0) = a_2, \end{cases}$$

where

$$x_1(t) \geq 0,\ 0 \leq u_j(t) \leq \alpha_j(\varepsilon,t),\ t \geq 0,\ j = 1,2. \tag{7.1}$$

Let $S = [0,\infty) \times R^1 \subset R^2$ denote the state constraint domain.

DEFINITION 7.1. *A control* $u_t = (u_1(t), u_2(t))$ *is* admissible *with respect to the initial state value* $a = (a_1, a_2) \in S$ *if* (i) $u(t)$ *is adapted to* $\sigma\{\alpha(\varepsilon,s), z_s : 0 \leq s \leq t\}$, (ii) $0 \leq u_j(t) \leq \alpha_j(\varepsilon,t)$ for $t \geq 0$ and $j = 1,2$, and (iii) *the corresponding state*

$$x_t = (x_1(t), x_2(t)) \in S\, for\ \ all\, t \geq 0. \tag{7.2}$$

We use $\mathcal{A}^\varepsilon$ to denote the set of admissible controls.

The problem is to find an admissible control $u(t)$ that minimizes the risk-sensitive cost function

$$J^{\varepsilon,\sqrt[4]{\varepsilon}}(u.) = \sqrt[4]{\varepsilon}\log E\left[\exp\left\{\frac{1}{\sqrt[4]{\varepsilon}}\int_0^\infty e^{-\rho t}[h(x_t)+c(u_t)]dt\right\}\right].$$

Note that the risk-sensitive parameter in the above cost function is chosen to be $\sqrt[4]{\varepsilon}$ rather than $\sqrt{\varepsilon}$. This is because a certain degree of sharpness in estimation is lost due to the presence of state constraints. We refer the reader to [22] for more discussion on this point.

We use $\mathcal{M} = \{\alpha^0,\ldots,\alpha^m\}$ to denote the machine capacity process, where $\alpha^i = (\alpha_1^i,\alpha_2^i)$ with α_j^i denoting the capacity of the jth machine in state i for $j = 1,2$.

We use $\mathcal{P}^{\varepsilon,\sqrt[4]{\varepsilon}}$ to denote our control problem, i.e.,

$$(7.3)\quad \mathcal{P}^{\varepsilon,\sqrt[4]{\varepsilon}}:\begin{cases}\text{min.} & J^{\varepsilon,\sqrt[4]{\varepsilon}}(u.)\\ & = \sqrt[4]{\varepsilon}\log E\left[\exp\left\{\frac{1}{\sqrt[4]{\varepsilon}}\int_0^\infty e^{-\rho t}[h(x_t)+c(u_t)]dt\right\}\right],\\ \text{s.t.} & \begin{cases}\dot{x}_1(t) = u_1(t)-u_2(t), & x_1(0)=a_1,\\ \dot{x}_2(t) = u_2(t)-z_t, & x_2(0)=a_2,\ u.\in\mathcal{A}^\varepsilon\end{cases}\\ \text{value fn.} & v^{\varepsilon,\sqrt[4]{\varepsilon}} = \inf\limits_{u.\in\mathcal{A}^\varepsilon} J^{\varepsilon,\sqrt[4]{\varepsilon}}(u.).\end{cases}$$

Similarly, as in §2, we define $\mathcal{A}^0$ as a set of the following $\mathcal{Z}_t$ adapted controls $U.$.

$$U. = (u^0(\cdot),\ldots,u^m(\cdot)) = ((u_1^0(\cdot),u_2^0(\cdot)),\ldots,(u_1^m(\cdot),u_2^m(\cdot)))$$

such that $0 \le u_j^i(t) \le \alpha_j^i$ for all $t \ge 0$, $j = 1,2$, and $i = 0,1,\ldots,m$, and the corresponding solutions $x. = (x_1(\cdot),x_2(\cdot))$ of the following system:

$$(7.4)\quad \begin{cases}\dot{x}_1(t) &= \sum\limits_{i=0}^m \nu_i u_1^i(t) - \sum\limits_{i=0}^m \nu_i u_2^i(t), & x_1(0)=a_1,\\ \dot{x}_2(t) &= \sum\limits_{i=0}^m \nu_i u_2^i(t) - z_t, & x_2(t)=a_2,\end{cases}$$

satisfy $x_t \in S$ for all $t \ge 0$.

We use $\mathcal{P}^{0,0}$ to denote the limiting problem.

$$\mathcal{P}^{0,0}:\begin{cases}\text{min.} & J^{0,0}(U.) = \left\|\int_0^\infty e^{-\rho t}\left[h(x(t)) + \sum\limits_{i=0}^m \nu_i c(u^i(t))\right]dt\right\|_\infty,\\ \text{s.t.} & \begin{cases}\dot{x}_1(t) = \sum\limits_{i=0}^m \nu_i u_1^i(t) - \sum\limits_{i=0}^m \nu_i u_2^i(t), & x_1(0)=a_1,\\ \dot{x}_2(t) = \sum\limits_{i=0}^m \nu_i u_2^i(t) - z_t, & x_2(0)=a_2,\\ U.\in\mathcal{A}^0,\end{cases}\\ \text{vaule fn.} & v^{0,0} = \inf\limits_{U.\in\mathcal{A}^0} J^{0,0}(U.).\end{cases}$$

Next, we describe the flow of constructing an asymptotic optimal control $u^\varepsilon_\cdot \in \mathcal{A}^\varepsilon$ of the original problem $\mathcal{P}^{\varepsilon,\sqrt[4]{\varepsilon}}$ beginning with a near optimal control $\bar{U}_\cdot \in \mathcal{A}^0$ of the problem $\mathcal{P}^{0,0}$.

Construction of nearly optimal controls. Let $\bar{U}_\cdot = (u^0(\cdot), \dots, u^m(\cdot)) \in \mathcal{A}^0$, where $u^j(t) = (u_1^j(t), u_2^j(t))$, be an ε'-optimal control for $\mathcal{P}^{0,0}$, and let

$$t^* = \inf\left\{t : \int_0^t \left[\sum_{i=0}^m \nu_i(\alpha_1^i - u_1^i(s) + u_2^i(s))\right] ds \geq \sqrt[4]{\varepsilon}\right\}.$$

We define another control process $\tilde{U}_t = (\tilde{u}^0(\cdot), \dots, \tilde{u}^m(\cdot)) \in \mathcal{A}^0$ as follows: for $j = 0, 1, \dots, m$,

$$\tilde{u}^i(t) = (\tilde{u}_1^i(t), \tilde{u}_2^i(t)) = \begin{cases} (\alpha_1^i, 0) & \text{if } t < t^*, \\ (u_1^i(t), u_2^i(t)) & \text{if } t \geq t^*. \end{cases} \tag{7.5}$$

Let

$$w(t) = (w_1(t), w_2(t)) = \sum_{i=0}^m \chi_{\{\alpha(\varepsilon,t)=\alpha^i\}}(\tilde{u}_1^i(t), \tilde{u}_2^i(t)), \tag{7.6}$$

and let $y(t) = (y_1(t), y_2(t))$ be the corresponding trajectory defined as

$$\begin{cases} y_1(t) = a_1 + \displaystyle\int_0^t (w_1(s) - w_2(s))ds, \\ y_2(t) = a_2 + \displaystyle\int_0^t (w_2(s) - z_s)ds. \end{cases}$$

Note that $w(t)$ satisfies the machine capacity constraints. However, it does not necessarily satisfy the state constraints, i.e., $y(t)$ may not be in S for some $t \geq 0$. To obtain an admissible control for $\mathcal{P}^{\varepsilon,\sqrt[4]{\varepsilon}}$, we need to modify $w(t)$ so that the state trajectory stays in S. We define

$$u_t^\varepsilon = w(t)\chi_{\{y_1(t)\geq 0\}}. \tag{7.7}$$

Then, $u^\varepsilon_\cdot \in \mathcal{A}^\varepsilon$. Moreover, we have the following theorem.

THEOREM 7.1. *Let $\bar{U}_\cdot \in \mathcal{A}^0$ be an open loop ε'-optimal for $\mathcal{P}^{0,0}$. Then, for the control $u^\varepsilon_\cdot \in \mathcal{A}^\varepsilon$ constructed above, there exist constants ε_0 and C such that, for all $0 < \varepsilon \leq \varepsilon_0$,*

$$\limsup_{\varepsilon\to 0} |J^{\varepsilon,\sqrt{\varepsilon}}(u^\varepsilon_\cdot) - v^{\varepsilon,\sqrt{\varepsilon}}| \leq C\varepsilon'. \tag{7.8}$$

Proof. We only sketch the proof as follows. We can define $\mathcal{P}^{0,\sqrt[4]{\varepsilon}}$, $J^{0,\sqrt[4]{\varepsilon}}$, and $v^{0,\sqrt[4]{\varepsilon}}$ similar to their definitions in previous sections.

Step 1. Following the proof of Theorem 4.1 and [22], we can show

$$J^{\varepsilon,\sqrt[4]{\varepsilon}}(u^\varepsilon_\cdot) \leq J^{0,\sqrt[4]{\varepsilon}}(\bar{U}_\cdot) + C\sqrt[4]{\varepsilon}.$$

In this step, Corollary 3.3 must be used to replace Corollary 3.2, which was used in [22].

Step 2. We can show as in the second part of the proof of Theorem 4.1 that

$$v^{\varepsilon,\sqrt[4]{\varepsilon}} \geq v^{0,\sqrt[4]{\varepsilon}} - C\sqrt[4]{\varepsilon}.$$

Step 3. We then check that Theorem 5.1 holds even with the presence of the state constraints.

Step 4. An estimate similar to the one in (6.4) can be obtained. Finally, (7.8) follows from the proof of Theorem 6.1. □

Remark 7.1. Note that only the construction of open loop controls is discussed in the above theorem. In fact, construction of feedback controls is much more difficult in the presence of state constraints. The reader is referred to [18] for some discussions on feedback controls.

Remark 7.2. As we mentioned, Theorem 7.1 can actually be extended to incorporate flowshops that consist of more than two machines (cf. [22]), or even more general jobshops that are formulated in [23]. The proofs are similar to the ones we sketched above. Since the procedures are much more involved, the statements, proofs, and results are omitted.

8. Concluding remarks. In this paper, we have carried out an asymptotic analysis of risk-sensitive production of stochastic manufacturing systems as the rates of machines breakdown and repairs become arbitrarily large. Based on this analysis we have constructed open loop and feedback production rate decisions for the original problem $\mathcal{P}^{\varepsilon,\sqrt{\varepsilon}}$ from a near-optimal control of the limit problem $\mathcal{P}^{0,0}$. We have shown that the constructed production decisions are asymptotically optimal as the fluctuation rate of the machine capacities goes to infinity, i.e., $\varepsilon \to 0$.

9. Appendix. In this section, we give five lemmas that are needed in the previous sections and the second part of the proof of Theorem 4.1.

LEMMA A.1. *There exist constants C and $k_0 > 0$ such that for all $t \geq 0$,*

$$|P(\alpha(\varepsilon,t)=i) - \nu_i| \leq C(\varepsilon + e^{-k_0 t/\varepsilon})\ \forall i \in \mathcal{M}.$$

Proof. We refer the reader to [20] for the proof. □

LEMMA A.2. *Let $\mathcal{Y} = \{y(\cdot) : [0,\infty) \to R^1, |y(t)| \leq N_0$ a.s.$\}$ and let*

$$f(t, y(\cdot), \alpha(\varepsilon,\cdot)) = \exp\left\{\frac{1}{\sqrt{\varepsilon}(t+1)^{\frac{3}{2}}}\left|\int_0^t (\chi_{\{\alpha(\varepsilon,t)=i\}} - \nu_i) y(s) ds\right|\right\}.$$

For each $y(\cdot) \in \mathcal{Y}$, define

$$\Psi_t(y(\cdot)) = Ef(t, y(\cdot), \alpha(\varepsilon,\cdot)).$$

Then, for any $\mathcal{Z}_t$ adapted process $\beta(\cdot) \in \mathcal{Y}$ a.s.,

$$\Psi_t(\beta(\cdot)) = E[f(t,\beta(\cdot),\alpha(\varepsilon,\cdot))|\mathcal{Z}_t] \text{ a.s.}$$

Proof. The proof is similar to the proof of [4, Lem. 14.18]. □

LEMMA A.3. *Let ξ denote a nonnegative random variable. Then,*

$$Ee^{\xi} \leq e + (e-1)\sum_{j=1}^{\infty} e^j P(\xi \geq j).$$

Proof. The proof is given as follows:

$$\begin{aligned}
Ee^{\xi} =&\sum_{j=0}^{\infty}\int_{\{j\leq\xi<j+1\}} e^{\xi}dP\\
\leq&\sum_{j=0}^{\infty}\int_{\{j\leq\xi<j+1\}} e^{j+1}dP\\
=&\sum_{j=0}^{\infty} e^{j+1}P(j\leq\xi<j+1)\\
=&\sum_{j=0}^{\infty} e^{j+1}[P(\xi\geq j)-P(\xi\geq j+1)]\\
\leq&\ \ e+(e-1)\sum_{j=1}^{\infty} e^{j}P(\xi\geq j). \qquad \square
\end{aligned}$$

LEMMA A.4. *Let ξ denote a nonnegative random variable such that $||\xi||_\infty < \infty$, where $||\xi||_\infty := \text{esssup}|\xi|$. Then, $\phi(x) := x^{-1}\log Ee^{x\xi}$ is a monotone increasing* function on $(0,\infty)$ and

$$\lim_{x\to\infty}\phi(x)=||\xi||_\infty.$$

Proof. First of all,

$$\phi'(x)=-\frac{1}{x^2}\log Ee^{x\xi}+\frac{1}{x}\frac{E\xi e^{x\xi}}{Ee^{x\xi}}=\frac{1}{x^2}\left[-\log Ee^{x\xi}+x\frac{E\xi e^{x\xi}}{Ee^{x\xi}}\right].$$

Let

$$\psi(x)=-\log Ee^{x\xi}+x\frac{E\xi e^{x\xi}}{Ee^{x\xi}}.$$

Then, $\phi'(x)=\frac{1}{x^2}\psi(x)$. Note that $\psi(0)=0$ and

$$\begin{aligned}
\psi'(x)=&-\frac{E\xi e^{x\xi}}{Ee^{x\xi}}+\frac{E\xi e^{x\xi}}{Ee^{x\xi}}+x\frac{E\xi^2e^{x\xi}Ee^{x\xi}-(E\xi e^{x\xi})^2}{(Ee^{x\xi})^2}\\
=&\,x\left(\frac{E\xi^2e^{x\xi}Ee^{x\xi}-(E\xi e^{x\xi})^2}{(Ee^{x\xi})^2}\right)\geq 0.
\end{aligned}$$

Therefore, $\psi(x)\geq 0$ for all $x\in(0,\infty)$. This implies $\phi'(x)\geq 0$ for all $x\in(0,\infty)$. Thus, $\phi(x)$ is a monotone increasing function.

To see the limit of $\phi(x)$ as $x\to\infty$, we observe that

$$\phi(x)\leq\frac{1}{x}\log Ee^{x||\xi||_\infty}=||\xi||_\infty<\infty.$$

This implies that $\lim_{x\to\infty}\phi(x)$ exists. Now, for any $\eta>0$, let $D=\{\xi\geq||\xi||_\infty-\eta\}$. Then $P(D)>0$.

$$\begin{aligned}
\phi(x)\geq&\frac{1}{x}\log E\chi_De^{x(||\xi||_\infty-\eta)}\\
=&\frac{1}{x}\log e^{x(||\xi||_\infty-\eta)}+\frac{1}{x}\log E\chi_D\\
=&\,||\xi||_\infty-\eta+\frac{1}{x}\log P(D).
\end{aligned}$$

Thus,

$$\|\xi\|_\infty \geq \lim_{x\to\infty} \phi(x) \geq \|\xi\|_\infty - \eta.$$

Since η is arbitrary, $\lim_{x\to\infty} \phi(x) = \|\xi\|_\infty$. □

LEMMA A.5. *Let $\beta(t)$ denote a $\sigma\{\alpha(\varepsilon, s), z_s : s \leq t\}$ adapted process. Then $E[\beta(t)|\mathcal{Z}_t] = E[\beta(t)| \vee_{t\geq 0} \mathcal{Z}_t]$. In particular, $E[\beta(t)|\mathcal{Z}_t] = E[\beta(t)|\mathcal{Z}_{t'}]$ for all $t' \geq t$.*

Proof. We refer the reader to [20] for the proof. □

Proof of Theorem 4.1 (*cont.*). To show (4.10), we first show that for any control $u^\varepsilon_\cdot \in \mathcal{A}^\varepsilon$, there exists a control $U_\cdot = (u^0_\cdot, \dots, u^m_\cdot) \in \mathcal{A}^0$ such that $|E[u^\varepsilon_t|\mathcal{Z}_t] - \sum_{i=0}^m \nu_i u^i_t|$ is small. In fact, for each $i \in \mathcal{M}$, let $u^i_t = E[u^\varepsilon_t|\alpha(\varepsilon, t) = i, \mathcal{Z}_t]$. Then $U_\cdot = (u^0_\cdot, \dots, u^m_\cdot) \in \mathcal{A}^0$. Moreover, note that u^i_t is $\mathcal{Z}_t$ adapted and $\mathcal{Z}_t$ is independent of $\alpha(\varepsilon, t)$.

$$\begin{aligned}
E[u^\varepsilon_t|\mathcal{Z}_t] &= E\left\{\sum_{i=0}^m \chi_{\{\alpha(\varepsilon,t)=i\}} u^\varepsilon_t \middle| \mathcal{Z}_t\right\}\\
&= E\left\{\sum_{i=0}^m \chi_{\{\alpha(\varepsilon,t)=i\}} E[u^\varepsilon_t|\alpha(\varepsilon,t), \mathcal{Z}_t] \middle| \mathcal{Z}_t\right\}\\
&= E\left\{\sum_{i=0}^m \chi_{\{\alpha(\varepsilon,t)=i\}} E[u^\varepsilon_t|\alpha(\varepsilon,t) = i, \mathcal{Z}_t] \middle| \mathcal{Z}_t\right\}\\
&= \sum_{i=0}^m E[\chi_{\{\alpha(\varepsilon,t)=i\}}|\mathcal{Z}_t] E[u^\varepsilon_t|\alpha(\varepsilon,t) = i, \mathcal{Z}_t]\\
&= \sum_{i=0}^m P(\alpha(\varepsilon,t) = i) E[u^\varepsilon_t|\alpha(\varepsilon,t) = i, \mathcal{Z}_t]\\
&= \sum_{i=0}^m P(\alpha(\varepsilon,t) = i) u^i_t.
\end{aligned}$$

By Lemma A.1,

$$\begin{aligned}
E[u^\varepsilon_t|\mathcal{Z}_t] &= \sum_{i=0}^m \nu_i u^i_t + \sum_{i=0}^m (P(\alpha(\varepsilon,t) = i) - \nu_i) u^i_t\\
&= \sum_{i=0}^m \nu_i u^i_t + O(\varepsilon + e^{-k_0 t/\varepsilon}).
\end{aligned} \tag{9.1}$$

We repeat the above argument with u^ε_t replaced by $c(u^\varepsilon_t)$. Then,

$$E[c(u^\varepsilon_t)|\mathcal{Z}_t] = \sum_{i=0}^m P(\alpha(\varepsilon,t) = i) E[c(u^\varepsilon_t)|\alpha(\varepsilon,t) = i, \mathcal{Z}_t].$$

Now, by the convexity of $c(u)$, we have

$$E[c(u^\varepsilon_t)|\alpha(\varepsilon,t) = i, \mathcal{Z}_t] \geq c(E[u^\varepsilon_t|\alpha(\varepsilon,t) = i, \mathcal{Z}_t]) = c(u^i_t).$$

Thus,

$$\begin{aligned}
E[c(u^\varepsilon_t)|\mathcal{Z}_t] &\geq \sum_{i=0}^m P(\alpha(\varepsilon,t) = i) c(u^i_t)\\
&= \sum_{i=0}^m \nu_i c(u^i_t) + O(\varepsilon + e^{-k_0 t/\varepsilon}).
\end{aligned} \tag{9.2}$$

Let $x.$ and $\bar{x}.$ be corresponding state trajectories under controls $u^\varepsilon_\cdot$ and $U.$, respectively. Then, by Lemma A.5,

$$\begin{aligned} E[x_t|\mathcal{Z}_t] &= x + \int_0^t [E[u^\varepsilon_s|\mathcal{Z}_s] - z_s]ds \\ \bar{x}_t \quad &= x + \int_0^t \left[\sum_{i=0}^m \nu_i u^i_s - z_s\right] ds. \end{aligned}$$

Subtracting the above two equations, we have

$$E[x_t|\mathcal{Z}_t] - \bar{x}_t = \int_0^t \left[E[u^\varepsilon_s|\mathcal{Z}_s] - \sum_{i=0}^m \nu_i u^i_t\right] ds. \tag{9.3}$$

Then, applying (9.1), we obtain

$$|E[x_t|\mathcal{Z}_t] - \bar{x}_t| = O(\varepsilon(1+t)).$$

Now, by the convex and locally Lipschitz assumptions on h,

$$\begin{aligned} J^{\varepsilon,\sqrt{\varepsilon}}(u^\varepsilon_\cdot) &= \sqrt{\varepsilon}\log E\left\{E\left[\exp\left\{\frac{1}{\sqrt{\varepsilon}}\int_0^\infty e^{-\rho t}[h(x_t) + c(u^\varepsilon_t)]dt\right\}\middle|\mathcal{Z}_t\right]\right\} \\ &\geq \sqrt{\varepsilon}\log E\left\{\exp\left\{\frac{1}{\sqrt{\varepsilon}}\int_0^\infty e^{-\rho t}[h(E[x_t|\mathcal{Z}_t]) + E[c(u^\varepsilon_t)|\mathcal{Z}_t]]dt\right\}\right\} \\ &= \sqrt{\varepsilon}\log E\left\{\exp\left\{\frac{1}{\sqrt{\varepsilon}}\int_0^\infty e^{-\rho t}\left[h(\bar{x}_t) + \sum_{i=0}^m \nu_i c(u^i_t)\right]dt\right\}\cdot I_1(\varepsilon)\right\}, \end{aligned}$$

where

$$I_1(\varepsilon) = \exp\frac{1}{\sqrt{\varepsilon}}\int_0^\infty e^{-\rho t}\left[h(E[x_t|\mathcal{Z}_t]) - h(\bar{x}_t) + E[c(u^\varepsilon_t)|\mathcal{Z}_t] - \sum_{i=0}^m \nu_i c(u^i_t)\right] dt.$$

Combining (9.2) and (9.3), we obtain

$$\begin{aligned} I_1(\varepsilon) &\geq \exp\frac{-k_3}{\sqrt{\varepsilon}}\int_0^\infty e^{-\rho t}\left[\varepsilon(1+t^{k_h})(1+t) + \varepsilon + \exp\left\{-\frac{k_0 t}{\varepsilon}\right\}\right] dt \\ &\geq \exp\{-k_4\sqrt{\varepsilon}\} \end{aligned}$$

for some $k_3 > 0$ and $k_4 > 0$. Therefore,

$$\begin{aligned} J^{\varepsilon,\sqrt{\varepsilon}}(u^\varepsilon_\cdot) &\geq \sqrt{\varepsilon}\log\exp\{-k_4\sqrt{\varepsilon}\} + J^{0,\sqrt{\varepsilon}}(U.) \\ &= -k_4\varepsilon + J^{0,\sqrt{\varepsilon}}(U.). \qquad \square \end{aligned}$$

REFERENCES

[1] T. Basar and P. Bernhard, *H^∞-Optimal Control and Related Minimax Design Problems*, Birkhäuser, Boston, 1991.

[2] E. N. Barron and R. Jensen, *Total risk aversion, stochastic optimal control, and differential games*, Appl. Math. Optim., 19 (1989), pp. 313–327.

[3] Committee on the Next Decade in Operations Research, *Operations research: the next decade*, Oper. Res., 36 (1988), pp. 619–637.

[4] R. J. Elliott, *Stochastic Calculus and Applications*, Springer-Verlag, New York, 1982.

[5] ———, *Smoothing for a finite state Markov process*, Lecture Notes in Control and Inform. Sci., 69 (1985), pp. 199–206.

[6] L. C. Evans and P. E. Souganidis, *Differential games and representation formulas for solutions of Hamilton–Jacobi–Isaacs equations*, Indiana Univ. Math. J., 33 (1984), pp. 773–797.

[7] W. H. Fleming, Chair, *Future directions in control theory: A mathematical perspective*, SIAM Reports on Issues in the Mathematical Sciences, Society for Industrial and Applied Mathematics, Philadelphia, PA, 1988.

[8] W. H. Fleming and W. M. McEneaney, *Risk sensitive optimal control and differential games*, in Lecture Notes in Control and Information Sciences 184, T. E. Duncan and B. Pasik-Duncan eds., Springer-Verlag, New York, 1992, pp. 185–197.

[9] ———, *Risk sensitive control with ergodic cost criteria*, in Proc. 31st IEEE Conference on Decision and Control, Tucson, 1992.

[10] W. H. Fleming and H. M. Soner, *Controlled Markov Processes and Viscosity Solutions*, Springer-Verlag, New York, 1993.

[11] S. Gershwin, *Hierarchical flow control: A framework for scheduling and planning discrete events in manufacturing systems*, in Proc. IEEE, Special Issue on Dynamics of Discrete Event Systems, 77 (1989), pp. 195–209.

[12] K. Glover and J. C. Doyle, *State space formulae for all stabilizing controllers that satisfy an H^∞ norm bound and relations to risk sensitivity*, Systems Control Lett., 11 (1988), pp. 167–172.

[13] H. Ishii, *Uniqueness of unbounded viscosity solutions of Hamilton–Jacobi equations*, Indiana Univ. Math. J., 26 (1984), pp. 721–748.

[14] M. R. James, *Asymptotic analysis of nonlinear stochastic risk-sensitive control and differential games*, Australian National University, 1992, working paper.

[15] J. Lehoczky, S. P. Sethi, H. M. Soner, and M. Taksar, *An asymptotic analysis of hierarchical control of manufacturing systems under uncertainty*, Math. Oper. Res., 16 (1992), pp. 596–608.

[16] D. F. Rogers, H. R. Evans, R. D. Plante, and R. T. Wong, *Aggregation and disaggregation techniques and methodology in optimization*, Oper. Res., 39 (1991), pp. 558–582.

[17] V. R. Saksena, J. O'Reilly, and P. V. Kokotovic, *Singular perturbations and time-scale methods in control theory: Survey* 1976–1983, Automatica, 20 (1984), pp. 273–293.

[18] S. P. Sethi, H. Yan, Q. Zhang, and X. Y. Zhou, *Feedback production planning in a stochastic two-machine flowshop: Asymptotic optimality and computation results*, Internat. J. Production Economics, 30–31 (1993), pp. 79–93.

[19] S. P. Sethi and Q. Zhang, *Hierarchical Decision Making in Stochastic Manufacturing Systems*, Birkhäuser, Boston, 1994.

[20] ———, *Hierarchical production planning in dynamic stochastic manufacturing systems: Asymptotic optimality and error bounds*, J. Math. Anal. Appl., 181 (1994), pp. 285–319.

[21] S. P. Sethi, Q. Zhang, and X. Y. Zhou, *Hierarchical controls in stochastic manufacturing systems with convex costs*, J. Optim. Theory Appl., 80 (1994), pp. 299–317.

[22] ———, *Hierarchical controls in stochastic manufacturing systems with machines in tandem*, Stochastics Stochastics Rep., 41 (1992), pp. 89–118.

[23] S. P. Sethi and X. Y. Zhou, *Dynamic stochastic jobshops and hierarchical controls*, IEEE Trans. Automat. Control, AC-39 (1994), pp. 2061–2076.

[24] H. M. Soner, *Singular perturbations in manufacturing systems*, SIAM J. Control Optim., 31 (1993), pp. 132–146.

[25] P. Whittle, *Risk-Sensitive Optimal Control*, John Wiley, New York, 1990.

[26] K. Yosida, *Functional Analysis*, 6th ed., Springer-Verlag, New York, 1980.

[27] Q. Zhang, *An asymptotic analysis of controlled diffusions with rapidly oscillating parameters*, Stochastics Stochastics Rep., 42 (1993), pp. 67–92.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 33, No. 2, pp. 528–553, March 1995

© 1995 Society for Industrial and Applied Mathematics

010

MULTILEVEL HIERARCHICAL DECISION MAKING IN STOCHASTIC MARKETING–PRODUCTION SYSTEMS*

S. P. SETHI† AND Q. ZHANG†

Abstract. This paper presents an asymptotic analysis of hierarchical marketing–production systems with stochastic demand and stochastic production capacity modelled as finite state Markov processes. The decision variables used are advertising and production rates which influence capacity, demand, and inventory levels. The objective of this paper is to maximize the expected total discounted profit over an infinite horizon. The authors are interested in situations in which the rate of change in capacity states is an order of magnitude different from the rate of change in demand states. These give rise to upper-level problems in which the stochastic capacity is replaced by the average capacity and/or the random demand is replaced by the average demand. Controls for the corresponding lower-level problems in different cases can be constructed from nearly optimal controls of the upper-level problems in a way that guarantees their asymptotic optimality.

Key words. stochastic manufacturing systems, marketing–production planning, hierarchical control, Markov processes, dynamic programming, viscosity solutions

AMS subject classifications. 93E20, 93A13, 90B30, 90C39

1. Introduction. One of the most important methods for dealing with optimization of large complex systems is that of hierarchical decomposition. The idea is to reduce the overall complex problem into manageable approximate problems or subproblems, solve these problems, and construct a solution of the original problem from solutions of these simpler problems. Development of such approaches for large complex systems has been identified as a particularly fruitful research area by the Committee on the Next Decade in Operations Research [6] and by the Panel on Future Directions in Control Theory [11].

Most manufacturing firms are large complex systems characterized by several decision subsystems such as finance, personnel, marketing, and operations. Furthermore, they may have a number of plants and warehouses and produce a large number of different products using a wide variety of machines and equipment. Moreover, these systems are subject to discrete events such as construction of new facilities; purchase of new equipment and scrappage of old; machine setups, failures, and repairs; and new product introductions. These events could be deterministic or stochastic. Management must recognize and react to these events. Because of the large size of these systems and the presence of these events, obtaining exact optimal policies to run these systems is nearly impossible both theoretically and computationally.

In practice, therefore, these systems, largely due to their complexity, are managed in a hierarchical fashion. In this context, Herbert Simon [29] writes, "My thesis has been that one path to the construction of a nontrivial theory of complex systems is by way of a theory of hierarchy. Empirically, a large proportion of the complex systems we observe in nature exhibit hierarchic structure." The literature provides little additional justification beyond the fact that these systems are complex for the practice of treating them hierarchically, especially when the environment is uncertain.

* Received by the editors June 16, 1993; accepted for publication (in revised form) November 23, 1993. This work was partly supported by Natural Sciences and Engineering Research Council of Canada grant A4619.

† Faculty of Management, University of Toronto, Toronto, Ontario, M5S 1V4, Canada. Present address: Department of Mathematics, University of Georgia, Athens, Georgia 30602.

There are several different and not mutually exclusive ways to reduce the complexity. These include decomposing the problem into problems of smaller subsystems with a proper coordinating mechanism; aggregating products and subsequently disaggregating them; replacing random processes with their averages and possibly other moments; and so on. For further details on hierarchical approaches in production planning systems and their importance in practice, we refer the reader to the surveys of the literature by Libosvar [18], Bitran and Tirupati [4], and Sethi and Zhang [22], [24], books by Stadtler [33], Switalski [34], and Sethi and Zhang [26], and a bibliography compiled by Bukh [5].

In this paper, we focus on the problem of a profit-maximizing manufacturing firm that must decide over time on the rate of promotional expenditures that create additional demand for its products and the rate of production to meet the demand. Problems incorporating promotional and production decisions have also been addressed by Abad [1], Sogomonian and Tang [30], and Sethi and Zhang [23]. Sethi, Taksar, and Zhang [21] treat a problem dealing with capacity expansion decisions along with production decisions and Zhou and Sethi [36] consider a problem with with production and personnel decisions.

The problem under consideration, termed the global problem, is formulated as a dynamic stochastic optimization problem with finite state Markovian demand and production capacity processes that depend, respectively, on the advertising and the production rates over time. In general, such problems are intractable. Either because of this intractability or because of some organizational considerations, such as the presence of a hierarchical structure within the firm, in practice the advertising and production planning decisions are made at different levels of the organization; see Meal [19] or Kistner and Switalski [16]. The former decisions are usually medium- or long-term decisions and are in the domain of marketing management. The latter are short- to medium-term decisions and are usually the concern of operations management. The two-level decision-making procedure works roughly as follows. Marketing (the upper level) bases its promotional decisions on some aggregated, rather than detailed, information from the shop floor. Subsequently, operations management (the lower level) makes production planning decisions given the advertising decisions already made at the upper level.

An important and obvious question that arises is whether there is a two-level decision procedure, such as the above, that is simpler than solving the global problem and is, at the same time, a good approximation of the optimal solution of the global problem. That such a procedure is usually simpler has been discussed in the literature; see Meal [19], for example. The theory developed in this paper answers the second part of the question in the affirmative under reasonable assumptions.

We shall develop several different two-level procedures, such as the above, that are simpler than solving the global problem and are, at the same time, good approximations of the optimal solution of the global problem. One such two-level procedure can be described as follows. The upper level solves a limiting problem obtained by replacing random capacities by their averages. The solution of this limiting problem yields an advertising decision as well as an average production plan. The upper level implements the advertising decision and informs the lower level of it. It is clear that the average production plan is not feasible for the global problem. However, one could construct a feasible production plan from it at the lower level that takes into account the stochastic demand resulting from the upper level's advertising decision. We are able to prove that the two-level decision procedure provides an asymptotic optimal solution to the global problem as the rates of transition between the various possible capacity states become very large in comparison to those between the demand states. The striking novelty of our approach is that this can be done without solving for the optimal solution, which, as stated earlier, is an insurmountable task.

It is important to note that the model we formulate is sufficiently rich and representative to illustrate the idea of asymptotic optimality in hierarchical manufacturing organizations in which medium-term and short-term decisions are made by different organizational units. Moreover, the processes taking place in the short term are much faster than those in the medium term. By a fast-changing process, we mean a process that is changing so rapidly that from any initial condition, it reaches its stationary distribution in a time during which there are few, if any, fluctuations in the other processes. The reader is referred to Lehoczky et al. [17, pp. 597 and 605] and Gershwin [14], [15] for further details on this point.

So far in our discussion, we have assumed that the capacity process is much faster than the demand process. In other words, the mean times between changes in capacity states is much smaller than those between changes in demand states. In the cases when the opposite might hold, a proper hierarchy would have operations management at the upper level and marketing at the lower level. In this paper we treat all possible hierarchies, including the case when capacity and demand processes have comparable frequencies with both of them being faster than the discounting process.

The model developed here represents an extension of preceding papers by Soner [31] and Sethi and Zhang [25] in the sense that it incorporates the possibility of influencing demand using advertising. In more general terms, the model has *two* explicit decision-making levels not present in the other two papers.

Furthermore, the results provide a rigorous theoretical justification, although in the case of a mathematically tractable model, for the common practice of hierarchical decision making as elaborated in the classical work of Anthony [2]. More importantly, they elicit deep insights into the structure of hierarchy and suggest near-optimal procedures of hierarchical decision making in more general contexts. Also, by establishing a criterion for determining the quality of hierarchical solutions, these results and further research could identify existing "right" and "wrong" practices. Viewed in this way, the results may have profound implications for the design of hierarchical structures within manufacturing organizations.

Finally, our model contains two parameters signifying the orders of magnitude associated with the rates of change of the capacity process and the demand process. We deal with cases in which only one of the two parameters is small and cases in which both parameters are small. In the latter cases, the arguments used in [25] to obtain the limiting problem in a single parameter case do not work because the two parameters in our problem may converge to zero at different rates. Thus, we need to modify the method used in [25] to suit our situation. In the former cases, the presence of the two parameters implies a number of different limiting problems. When one parameter is fixed while the other is small, the associated limiting problem involves a stochastic process associated with the fixed parameter. While this does not complicate the asymptotic analysis of (partial) open-loop controls, it requires additional care in dealing with feedback controls. Specifically, the uniqueness of a solution for such a stochastic limiting problem needs to be specified via a solution to an associated martingale problem. Consequently, various concepts in probability and stochastic processes such as tightness, convergence in distribution, Skorohod's representation, and weak convergence in the space of square integrable functions are used for obtaining the desired asymptotic optimality results.

The plan of the paper is as follows. In §2, we formulate the model of the marketing–production system under consideration and the related global stochastic optimization problem. Also developed are possible hierarchies and relevant limiting problems. In §3, we discuss some elementary properties of the associated value functions and show that the value function of our problem converges to the value functions of appropriate limiting problems. In §4, we study the asymptotic behavior of the capacity and demand processes. Then, in §5, we provide a construction of the asymptotically optimal open-loop controls. Asymptotically optimal feedback controls are studied in §6. Finally, §7 concludes the paper.

2. Problem formulation and possible hierarchies.

In §2.1 we develop the marketing–production problem under consideration. The stochastic processes and control variables involved in the problem will be specified precisely in §2.2. In §2.3, we develop possible hierarchies by specifying three different limiting problems.

2.1. The marketing–production problem.

Let us consider a marketing–production system that produces n distinct product types using a random production capacity $\alpha(\varepsilon, t) \in R^1$ (parameter ε to be specified later). Let $\mathbf{u}_t \in R^n$ denote the production rate. Clearly $\mathbf{u}_t \geq 0$ and, in addition, $\mathbf{u}_t$ will be subject to the available production capacity $\alpha(\varepsilon, t)$ in a way defined later. With the total surplus $\mathbf{x}_t \in R^n$ and a stochastic demand rate $\mathbf{z}(\delta, t) \in R^n$ (parameter δ to be specified later), the system dynamic is

$$\dot{\mathbf{x}}_t = \mathbf{u}_t - \mathbf{z}(\delta, t), \ \mathbf{x}_0 = \mathbf{x}, \tag{2.1}$$

where $\mathbf{x} \in R^n$ is the initial surplus. Note that surplus represents inventories when positive and shortages when negative. Here and elsewhere we use boldface letters to represent vectors.

We consider the profit functional J defined by

$$J(\mathbf{x}, \alpha, \mathbf{z}, \mathbf{u}., \mathbf{w}.) = E\left[\int_0^\infty e^{-\rho t} G(\mathbf{x}_t, \mathbf{z}(\delta, t), \mathbf{u}_t, \mathbf{w}_t) dt\right], \tag{2.2}$$

where $\rho > 0$ is the discount rate; $\mathbf{w}_t \in R^{n_1}$ is the rate of advertising; G is the net income function of $\mathbf{x}, \mathbf{z}, \mathbf{u}$, and $\mathbf{w}$; and $\mathbf{x}$, α, and $\mathbf{z}$ are the initial surplus, the initial capacity, and the initial demand, respectively. The problem is to find a control $(\mathbf{u}_t, \mathbf{w}_t)$, $t \geq 0$, that maximizes $J(\mathbf{x}, \alpha, \mathbf{z}, \mathbf{u}., \mathbf{w}.)$.

Example 2.1. Let $x_t \in R^1$ denote the total surplus at time t, $u_t \in R^1$ denote the production rate, $w_t \in R^1$ denote the advertising rate, and $z(\delta, t) \in R^1$ denote the demand rate. These variables satisfy (2.1). The decision variables u_t and w_t satisfy the constraints $0 \leq u_t \leq \alpha(\varepsilon, t)$, $0 \leq w_t \leq K < \infty$, where K represents an upper bound on the advertising rate. The objective is to choose admissible decision $(u., w.)$ to maximize the expected total discounted profit

$$J(x, \alpha, z, u., w.) = E\int_0^\infty e^{-\rho t}[\pi z(\delta, t) - (h_1(x_t) + cu_t + w_t)]dt,$$

where π is the revenue per unit sale, $h_1(\cdot)$ is the inventory/shortage cost function, and $c < \pi$ is the unit production cost. See Sethi [20] and Feichtinger, Hartl, and Sethi [10] for surveys on dynamic optimal control models in advertising.

We now specify the production and advertising constraints. For each possible capacity state i, let

$$\mathcal{U}_i = \{\mathbf{u} = (u_1, \ldots, u_n) \geq 0 : \mathbf{p} \cdot \mathbf{u} \leq i\} \subset R^n, \tag{2.3}$$

where $\mathbf{p} = (p_1, \ldots, p_n) \geq 0$ is a given constant vector with p_j representing the amount of capacity needed to produce product type j at rate 1. Let $\mathcal{W}$ denote a convex compact subset of the positive orthant of R^{n_1}. With this definition, the advertising and production constraints at time t, respectively, are

$$\mathbf{w}_t \in \mathcal{W} \text{ and } \mathbf{u}_t \in \mathcal{U}_{\alpha(\varepsilon,t)}.$$

2.2. Specification of capacity and demand processes. Our purpose in this section is to specify precisely the joint process $(\alpha(\varepsilon,t), \mathbf{z}(\delta,t))$ as a Markov process constructed from an infinitesimal generator that depends on production and advertising rates.

We begin with a standard probability space $(\Omega, \mathcal{F}, P)$. Let $\mathcal{Z} = \{\mathbf{z}^0, \mathbf{z}^1, \ldots, \mathbf{z}^d\}$ denote the set of demand states and let $\mathbf{w}_t \in R^{n_1}$, $\mathbf{w}_t \geq 0$, $\mathbf{w}_t$ bounded, denote the advertising rate at time t. We shall assume that the demand process $\mathbf{z}(\delta,t)$ takes values in $\mathcal{Z}$ and that transitions between the demand states depend on the rate of advertising. Assume that the random process $\alpha(\varepsilon,t) \in \mathcal{M} = \{0, 1, \ldots, m\}$, and that the transitions between the capacity states may depend on the production rate.

The dependence of $\alpha(\varepsilon,t)$ on the production rate will be given by a generator $\varepsilon^{-1}Q^m(\mathbf{u})$, $\varepsilon > 0$, where $Q^m(\mathbf{u})$ is an $(m+1) \times (m+1)$-matrix such that $Q^m(\mathbf{u}) = \{q^m_{ij}(\mathbf{u})\}$ with $q^m_{ij}(\mathbf{u}) \geq 0$ if $i \neq j$ and $q^m_{ii}(\mathbf{u}) = -\sum_{j \neq i} q^m_{ij}(\mathbf{u})$. To model the dependence of $\mathbf{z}(\delta,t)$ on the advertising decision, we let $\delta > 0$ and $\mathbf{z}(\delta,t) \in \mathcal{Z}$ be governed by a generator $\delta^{-1}Q^d(\mathbf{w})$. Here $Q^d(\mathbf{w})$ is a $(d+1) \times (d+1)$-matrix such that $Q^d(\mathbf{w}) = \{q^d_{ij}(\mathbf{w})\}$ with $q^d_{ij}(\mathbf{w}) \geq 0$ if $i \neq j$ and $q^d_{ii}(\mathbf{w}) = -\sum_{j \neq i} q^d_{ij}(\mathbf{w})$.

Next, we define the average distributions associated with the generators. For this purpose, let

$$\begin{aligned} &\Gamma^m = \{U = (\mathbf{u}^0, \mathbf{u}^1, \ldots, \mathbf{u}^m) : \text{ such that } \mathbf{u}^i \in \mathcal{U}_i\} \\ \text{and}\quad &\Gamma^d = \{W = (\mathbf{w}^0, \mathbf{w}^1, \ldots, \mathbf{w}^d) : \text{ such that } \mathbf{w}^i \in \mathcal{W}\}. \end{aligned} \tag{2.4}$$

We define two matrices $\bar{Q}^m(U)$ and $\bar{Q}^d(W)$ as functions of $(U, W) \in \Gamma^m \times \Gamma^d$:

$$\bar{Q}^m(U) = \begin{pmatrix} q^m_{00}(\mathbf{u}^0) & q^m_{01}(\mathbf{u}^0) & \cdots & q^m_{0m}(\mathbf{u}^0) \\ q^m_{10}(\mathbf{u}^1) & q^m_{11}(\mathbf{u}^1) & \cdots & q^m_{0m}(\mathbf{u}^1) \\ \vdots & \vdots & \cdots & \vdots \\ q^m_{m0}(\mathbf{u}^m) & q^m_{m1}(\mathbf{u}^m) & \cdots & q^m_{mm}(\mathbf{u}^m) \end{pmatrix},$$

$$\bar{Q}^d(W) = \begin{pmatrix} q^d_{00}(\mathbf{w}^0) & q^d_{01}(\mathbf{w}^0) & \cdots & q^d_{0d}(\mathbf{w}^0) \\ q^d_{10}(\mathbf{w}^1) & q^d_{11}(\mathbf{w}^1) & \cdots & q^d_{1d}(\mathbf{w}^1) \\ \vdots & \vdots & \cdots & \vdots \\ q^d_{d0}(\mathbf{w}^d) & q^d_{d1}(\mathbf{w}^d) & \cdots & q^d_{dd}(\mathbf{w}^d) \end{pmatrix}.$$

For each $U \in \Gamma^m$ and $W \in \Gamma^d$, let

$$\begin{aligned} \boldsymbol{\nu}^m(U) &= (\nu_0^m(U), \nu_1^m(U), \dots, \nu_m^m(U)), \\ \text{and} \quad \boldsymbol{\nu}^d(W) &= (\nu_0^d(W), \nu_1^d(W), \dots, \nu_d^d(W)) \end{aligned}$$

denote nonnegative solutions, respectively, to

$$\begin{aligned} &\boldsymbol{\nu}^m(U)\bar{Q}^m(U) = 0 \text{ and } \textstyle\sum_{i=0}^m \nu_i^m(U) = 1, \\ \text{and} \quad &\boldsymbol{\nu}^d(W)\bar{Q}^d(W) = 0 \text{ and } \textstyle\sum_{i=0}^d \nu_i^d(W) = 1. \end{aligned} \tag{2.5}$$

The vectors $\boldsymbol{\nu}^m(U)$ and $\boldsymbol{\nu}^d(W)$ will be called the average distributions of $\bar{Q}^m(U)$ and $\bar{Q}^d(W)$ for any given $U \in \Gamma^m$ and $W \in \Gamma^d$, respectively.

We make the following assumptions on the function G and the generators $Q^m(\mathbf{u})$ and $Q^d(\mathbf{w})$.

(A1) There exists a constant c_g, such that for all $\mathbf{x}, \mathbf{x}', \mathbf{u}, \mathbf{u}', \mathbf{w}$, and $\mathbf{w}'$, we have

$$\begin{aligned} &0 \le G(\mathbf{x}, \mathbf{z}, \mathbf{u}, \mathbf{w}) \le c_g(1 + |\mathbf{x}|), \\ &|G(\mathbf{x}, \mathbf{z}, \mathbf{u}, \mathbf{w}) - G(\mathbf{x}', \mathbf{z}, \mathbf{u}', \mathbf{w}')| \\ &\quad \le c_g(1 + |\mathbf{x}| + |\mathbf{x}'|)|\mathbf{x} - \mathbf{x}'| + c_g|\mathbf{u} - \mathbf{u}'| + c_g|\mathbf{w} - \mathbf{w}'|. \end{aligned}$$

(A2) $Q^m(\mathbf{u})$ and $Q^d(\mathbf{w})$ are continuous functions of $\mathbf{u}$ and $\mathbf{w}$. Moreover, for each $U \in \Gamma^m$ and $W \in \Gamma_d$, the systems of equations (2.5) have unique nonnegative solutions $\boldsymbol{\nu}^m(U)$ and $\boldsymbol{\nu}^d(W)$, which are assumed to be continuous with respect to $U \in \Gamma^m$ and $W \in \Gamma^d$. Furthermore, there exist $U \in \Gamma^m$ and $W \in \Gamma^d$ such that $\bar{Q}^m(U)$ and $\bar{Q}^d(W)$ are both irreducible.

We give two examples in which assumption (A2) holds.

Example 2.2. $Q^m(\mathbf{u}) = Q^m$ and $Q^d(\mathbf{w}) = Q^d$, where Q^m and Q^d are constant irreducible matrices. In this case, $\boldsymbol{\nu}^m(U) = \boldsymbol{\nu}^m$ and $\boldsymbol{\nu}^d(W) = \boldsymbol{\nu}^d$ are the familiar equilibrium distributions.

Example 2.3. Let $Q^m(\mathbf{u})$ and $Q^d(\mathbf{w})$ denote generators of birth–death processes, i.e.,

$$Q^m(\mathbf{u}) = \begin{pmatrix} -\mu_0 & \mu_0 & 0 & 0 & \cdots & 0 \\ \lambda_1(\mathbf{u}) & -(\lambda_1(\mathbf{u}) + \mu_1) & \mu_1 & 0 & \cdots & 0 \\ 0 & \lambda_2(\mathbf{u}) & -(\lambda_2(\mathbf{u}) + \mu_2) & \mu_2 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \cdots & \vdots \\ 0 & 0 & 0 & \cdots & \lambda_m(\mathbf{u}) & -\lambda_m(\mathbf{u}) \end{pmatrix},$$

$$Q^d(\mathbf{w}) =$$

$$\begin{pmatrix} -\bar{\mu}_0(\mathbf{w}) & \bar{\mu}_0(\mathbf{w}) & 0 & 0 & \cdots & 0 \\ \bar{\lambda}_1(\mathbf{w}) & -(\bar{\lambda}_1(\mathbf{w}) + \bar{\mu}_1(\mathbf{w})) & \bar{\mu}_1(\mathbf{w}) & 0 & \cdots & 0 \\ 0 & \bar{\lambda}_2(\mathbf{w}) & -(\bar{\lambda}_2(\mathbf{w}) + \bar{\mu}_2(\mathbf{w})) & \bar{\mu}_2(\mathbf{w}) & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \cdots & \vdots \\ 0 & 0 & 0 & \cdots & \bar{\lambda}_d(\mathbf{w}) & -\bar{\lambda}_d(\mathbf{w}) \end{pmatrix}.$$

We assume that $\mu_i > 0$, $\bar{\lambda}_i > 0$, and that $\lambda_i(\mathbf{u}) \geq 0$ and $\bar{\mu}_i(\mathbf{w})$ are continuous in $\mathbf{u}, \mathbf{w}$. Thus, for any given $U \in \Gamma$ and $W \in \Gamma^d$, we have

$$\bar{Q}^m(U) = \begin{pmatrix} -\mu_0 & \mu_0 & 0 & 0 & \cdots & 0 \\ \lambda_1(\mathbf{u}^1) & -\lambda_1(\mathbf{u}^1) - \mu_1 & \mu_1 & 0 & \cdots & 0 \\ 0 & \lambda_2(\mathbf{u}^2) & -\lambda_2(\mathbf{u}^2) - \mu_2 & \mu_2 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \cdots & \vdots \\ 0 & 0 & 0 & \cdots & \lambda_m(\mathbf{u}^m) & -\lambda_m(\mathbf{u}^m) \end{pmatrix},$$

$$\bar{Q}^d(\mathbf{w}) =$$

$$\begin{pmatrix} -\bar{\mu}_0(\mathbf{w}^0) & \bar{\mu}_0(\mathbf{w}^0) & 0 & 0 & \cdots & 0 \\ \bar{\lambda}_1(\mathbf{w}^1) & -\bar{\lambda}_1(\mathbf{w}^1) - \bar{\mu}_1(\mathbf{w}^1) & \bar{\mu}_1(\mathbf{w}^1) & 0 & \cdots & 0 \\ 0 & \bar{\lambda}_2 & -\bar{\lambda}_2(\mathbf{w}^2) - \bar{\mu}_2(\mathbf{w}^2) & \bar{\mu}_2(\mathbf{w}^2) & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \cdots & \vdots \\ 0 & 0 & 0 & \cdots & \bar{\lambda}_d(\mathbf{w}^d) & -\bar{\lambda}_d(\mathbf{w}^d) \end{pmatrix}.$$

The average distributions $\boldsymbol{\nu}^m(U)$ and $\boldsymbol{\nu}^d(W)$ are given as follows:

$$\nu_0^m(U) = \frac{\lambda_1(\mathbf{u}^1) \cdots \lambda_m(\mathbf{u}^m)}{\lambda_1(\mathbf{u}^1) \cdots \lambda_m(\mathbf{u}^m) + \mu_0 \lambda_2(\mathbf{u}^2) \cdots \lambda_m(\mathbf{u}^m) + \cdots + \mu_0 \cdots \mu_{m-1}},$$

and for $i \geq 1$,

$$\nu_i^m(U) = \nu_0^m(U) \frac{\mu_0 \cdots \mu_{i-1}}{\lambda_1(\mathbf{u}^1) \cdots \lambda_m(\mathbf{u}^m)};$$

$$\nu_0^d(W) =$$

$$\frac{\bar{\lambda}_1(\mathbf{w}^1) \cdots \bar{\lambda}_d(\mathbf{w}^d)}{\bar{\lambda}_1(\mathbf{w}^1) \cdots \bar{\lambda}_d(\mathbf{w}^d) + \bar{\mu}_0(\mathbf{w}^0) \bar{\lambda}_2(\mathbf{w}^2) \cdots \bar{\lambda}_d(\mathbf{w}^d) + \cdots + \bar{\mu}_0(\mathbf{w}^0) \cdots \bar{\mu}_{d-1}(\mathbf{w}^{d-1})},$$

and for $i \geq 1$,

$$\nu_i^d(W) = \nu_0^d(W) \frac{\bar{\mu}_0(\mathbf{w}^0) \cdots \bar{\mu}_{i-1}(\mathbf{w}^{i-1})}{\bar{\lambda}_1(\mathbf{w}^1) \cdots \bar{\lambda}_d(\mathbf{w}^d)}.$$

Note that $Q^m(\mathbf{u})$ in the example could be thought of as a production capacity process consisting of m identical machines of unit capacity each. The state $i \in \mathcal{M}$ would correspond to the situation when i machines are up and $(m - i)$ machines are down. The machine breakdown rate depends on the production rate $\mathbf{u}$, whereas their repair rates are independent of it.

Let us now return to the general setup. Since $Q^m(\mathbf{u})$ and $Q^d(\mathbf{w})$ depend on the control variables of the system under consideration, the processes $\alpha(\varepsilon, t)$ and $\mathbf{z}(\delta, t)$ need to be defined simultaneously by using the piecewise-deterministic process approach introduced by Davis [8]. To this end, we need to define the generator of the joint process $(\alpha(\varepsilon, t), \mathbf{z}(\delta, t))$.

First of all, note that the state space of $(\alpha(\varepsilon,t),\mathbf{z}(\delta,t))$ is given by

$$\mathcal{M}\times\mathcal{Z}=\{(0,\mathbf{z}^0),\dots,(0,\mathbf{z}^d),(1,\mathbf{z}^0),\dots,(1,\mathbf{z}^d),\dots,(m,\mathbf{z}^0),\dots,(m,\mathbf{z}^d)\}.$$

For ease of notation, we represent $\mathcal{M}\times\mathcal{Z}$ by the set

$$\{(00),\dots,(0d),(10),\dots,(1d),\dots,(m0),\dots,(md)\}$$

such that (ij) corresponds to $(i,\mathbf{z}^j)$.

We define an $[(m+1)\times(d+1)]\times[(m+1)\times(d+1)]$-matrix $Q(\mathbf{u},\mathbf{w})$ as follows:

$$Q(\mathbf{u},\mathbf{w})=\frac{1}{\varepsilon}\begin{pmatrix} q^m_{00}(\mathbf{u})I & q^m_{01}(\mathbf{u})I & \cdots & q^m_{0m}(\mathbf{u})I\\ q^m_{10}(\mathbf{u})I & q^m_{11}(\mathbf{u})I & \cdots & q^m_{1m}(\mathbf{u})I\\ \vdots & \vdots & \cdots & \vdots\\ q^m_{m0}(\mathbf{u})I & q^m_{m1}(\mathbf{u})I & \cdots & q^m_{mm}(\mathbf{u})I\end{pmatrix} +\frac{1}{\delta}\begin{pmatrix} Q^d(\mathbf{w}) & 0 & \cdots & 0\\ 0 & Q^d(\mathbf{w}) & \cdots & 0\\ \vdots & \vdots & \cdots & \vdots\\ 0 & 0 & \cdots & Q^d(\mathbf{w})\end{pmatrix},$$

where I is the $(d+1)\times(d+1)$-identity matrix. If we let $q_{(ii')(jj')}(\mathbf{u},\mathbf{w})$ denote the $(ii')(jj')$th entry of $Q(\mathbf{u},\mathbf{w})$, then one can see that

$$q_{(ii')(jj')}(\mathbf{u},\mathbf{w})=\begin{cases}\varepsilon^{-1}q^m_{ii}(\mathbf{u})+\delta^{-1}q^d_{i'i'}(\mathbf{w}) & \text{if } i=j,\ i'=j',\\ \delta^{-1}q^d_{i'j'}(\mathbf{w}) & \text{if } i=j,\ i'\neq j',\\ \varepsilon^{-1}q^m_{ij}(\mathbf{u}) & \text{if } i\neq j,\ i'=j',\\ 0 & \text{if } i\neq j,\ i'\neq j'.\end{cases}$$

We can also define a matrix $\bar{Q}(U,W)$ as follows:

$$\bar{Q}(U,W)=\frac{1}{\varepsilon}\begin{pmatrix} q^m_{00}(\mathbf{u}^0)I & q^m_{01}(\mathbf{u}^0)I & \cdots & q^m_{0m}(\mathbf{u}^0)I\\ q^m_{10}(\mathbf{u}^1)I & q^m_{11}(\mathbf{u}^1)I & \cdots & q^m_{1m}(\mathbf{u}^1)I\\ \vdots & \vdots & \cdots & \vdots\\ q^m_{m0}(\mathbf{u}^m)I & q^m_{m1}(\mathbf{u}^m)I & \cdots & q^m_{mm}(\mathbf{u}^m)I\end{pmatrix} +\frac{1}{\delta}\begin{pmatrix} \bar{Q}^d(W) & 0 & \cdots & 0\\ 0 & \bar{Q}^d(W) & \cdots & 0\\ \vdots & \vdots & \cdots & \vdots\\ 0 & 0 & \cdots & \bar{Q}^d(W)\end{pmatrix}. \tag{2.6}$$

It can be shown that if $\bar{Q}^m(U)$ and $\bar{Q}^d(W)$ are irreducible for some U and W, respectively, then $\bar{Q}(U,W)$ is also irreducible for some (U,W).

With $\bar{Q}(U, W)$ thus specified, we can construct the desired Markov process $(\alpha(\varepsilon, t), \mathbf{z}(\delta, t))$ by the same approach used in [25]. Moreover, the process

$$(2.7)\quad f(\alpha(\varepsilon,t),\mathbf{z}(\delta,t)) - f(\alpha(\varepsilon,s),\mathbf{z}(\delta,s)) - \int_s^t Q(\mathbf{u}_r,\mathbf{w}_r) f(\alpha(\varepsilon,r),\mathbf{z}(\delta,r))dr$$

is a martingale for any bounded function f on $\mathcal{M} \times \mathcal{Z}$.

2.3. Possible hierarchies and associated problems. In this section, we shall state precisely the various optimization problems that will be studied in this paper. These problems arise in the process of specifying various possible hierarchies that are involved.

We begin with a precise statement of the marketing–production problem developed in §§2.1 and 2.2.

DEFINITION 2.1. *We say that a control* $(\mathbf{u}.,\mathbf{w}.) = \{(\mathbf{u}_t,\mathbf{w}_t) : t \geq 0\}$ *is* admissible *if* $(\mathbf{u}.,\mathbf{w}.)$ *is right-continuous having left-hand limit* (RCLL), *is* $\sigma\{(\alpha(\varepsilon,s),\mathbf{z}(\delta,s)) : s \leq t\}$ *adapted, and satisfies* $\mathbf{u}_t \in \mathcal{U}_{\alpha(\varepsilon,t)}$ *and* $\mathbf{w}_t \in \mathcal{W}$ *for all* $t \geq 0$. *We use* $\mathcal{A}^{\varepsilon,\delta}$ *to denote the set of all admissible controls. Then our control problem can be written as follows*:

$$(2.8)\ \mathcal{P}^{\varepsilon,\delta} : \begin{cases} \text{max.} & J^{\varepsilon,\delta}(\mathbf{x},\alpha,\mathbf{z},\mathbf{u}.,\mathbf{w}.) = E\displaystyle\int_0^\infty e^{-\rho t} G(\mathbf{x}_t,\mathbf{z}(\delta,t),\mathbf{u}_t,\mathbf{w}_t)dt \\ \text{s.t.} & \begin{cases} \dot{\mathbf{x}}_t = \mathbf{u}_t - \mathbf{z}(\delta,t), \ \mathbf{x}_0 = \mathbf{x}, \\ (\alpha(\varepsilon,t),\mathbf{z}(\delta,t)) \sim \bar{Q}(\mathbf{u}_t,\mathbf{w}_t), \\ (\alpha(\varepsilon,0),\mathbf{z}(\delta,0)) = (\alpha,\mathbf{z}), \\ (\mathbf{u}.,\mathbf{w}.) \in \mathcal{A}^{\varepsilon,\delta}, \end{cases} \\ \text{value fn.} & v^{\varepsilon,\delta}(\mathbf{x},\alpha,\mathbf{z}) = \displaystyle\sup_{(\mathbf{u}.,\mathbf{w}.)\in\mathcal{A}^{\varepsilon,\delta}} J^{\varepsilon,\delta}(\mathbf{x},\alpha,\mathbf{z},\mathbf{u}.,\mathbf{w}.), \end{cases}$$

where by $(\alpha(\varepsilon,t),\mathbf{z}(\delta,t)) \sim \bar{Q}(\mathbf{u}_t,\mathbf{w}_t)$, *we mean that the Markov process* $(\alpha(\varepsilon,t),\mathbf{z}(\delta,t))$ *has the generator* $\bar{Q}(\mathbf{u}_t,\mathbf{w}_t)$.

We use $\mathcal{A}^{0,0}$, $\mathcal{A}^{\varepsilon,0}$, and $\mathcal{A}^{0,\delta}$ to denote the admissible control spaces

$$\mathcal{A}^{0,0} = \{(U_t,W_t) \in \Gamma^m \times \Gamma^d : (U_t,W_t) \text{ is deterministic and RCLL}\},$$
$$\mathcal{A}^{\varepsilon,0} = \{(\mathbf{u}_t,W_t) \in \mathcal{U}_{\alpha(\varepsilon,t)} \times \Gamma^d : (\mathbf{u}_t,W_t) \text{ is } \sigma\{\alpha(\varepsilon,s) : s \leq t\} \text{ adapted and RCLL}\},$$
$$\mathcal{A}^{0,\delta} = \{(U_t,\mathbf{w}_t) \in \Gamma^m \times \mathcal{W} : (U_t,\mathbf{w}_t) \text{ is } \sigma\{\mathbf{z}(\delta,s) : s \leq t\} \text{ adapted and RCLL}\},$$

respectively, for the optimal control problems $\mathcal{P}^{0,0}, \mathcal{P}^{\varepsilon,0}$, and $\mathcal{P}^{0,\delta}$ given below:

$$(2.9)\qquad \mathcal{P}^{\varepsilon,0}: \begin{cases} \text{max.} & J^{\varepsilon,0}(\mathbf{x}, \alpha, \mathbf{u}., W.) \\ & \quad = \displaystyle\int_0^\infty e^{-\rho t} \sum_{j=0}^{d} \nu_j^d(W_t) G(\mathbf{x}_t, \mathbf{z}^j, \mathbf{u}_t, \mathbf{w}_t^j)\, dt \\ \text{s.t.} & \begin{cases} \dot{\mathbf{x}}_t = \mathbf{u}_t - \displaystyle\sum_{j=0}^{d} \nu_j^d(W_t)\mathbf{z}^j, & \mathbf{x}_0 = \mathbf{x}, \\ \alpha(\varepsilon, t) \sim \dfrac{1}{\varepsilon} Q^m(\mathbf{u}_t), & \alpha(\varepsilon, 0) = \alpha, \\ (\mathbf{u}., W.) \in \mathcal{A}^{\varepsilon,0}, \end{cases} \\ \text{value fn.} & v^{\varepsilon,0}(\mathbf{x}, \alpha) = \displaystyle\sup_{(\mathbf{u}., W.) \in \mathcal{A}^{\varepsilon,0}} J^{\varepsilon,0}(\mathbf{x}, \alpha, \mathbf{u}., W.); \end{cases}$$

$$(2.10)\qquad \mathcal{P}^{0,\delta}: \begin{cases} \text{max.} & J^{0,\delta}(\mathbf{x}, \mathbf{z}, U., \mathbf{w}.) \\ & \quad = \displaystyle\int_0^\infty e^{-\rho t} \sum_{i=0}^{m} \nu_i^m(U_t) G(\mathbf{x}_t, \mathbf{z}(\delta, t), \mathbf{u}_t^i, \mathbf{w}_t)\, dt \\ \text{s.t.} & \begin{cases} \dot{\mathbf{x}}_t = \displaystyle\sum_{i=0}^{m} \nu_i^m(U_t)\mathbf{u}_t^i - \mathbf{z}(\delta, t), & \mathbf{x}_0 = \mathbf{x}, \\ \mathbf{z}(\delta, t) \sim \dfrac{1}{\delta} Q^d(\mathbf{w}_t), & \mathbf{z}(\delta, 0) = \mathbf{z}, \\ (U., \mathbf{w}.) \in \mathcal{A}^{0,\delta}, \end{cases} \\ \text{value fn.} & v^{0,\delta}(\mathbf{x}, \mathbf{z}) = \displaystyle\sup_{(U., \mathbf{w}.) \in \mathcal{A}^{0,\delta}} J^{0,\delta}(\mathbf{x}, \mathbf{z}, U., \mathbf{w}.); \end{cases}$$

$$(2.11)\qquad \mathcal{P}^{0,0}: \begin{cases} \text{max.} & J^{0,0}(\mathbf{x}, U., W.) \\ & \quad = \displaystyle\int_0^\infty e^{-\rho t} \sum_{i=0, j=0}^{m,d} \nu_i^m(U_t)\nu_j^d(W_t) G(\mathbf{x}_t, \mathbf{z}^j, \mathbf{u}_t^i, \mathbf{w}_t^j)\, dt \\ \text{s.t.} & \begin{cases} \dot{\mathbf{x}}_t = \displaystyle\sum_{i=0}^{m} \nu_i^m(U_t)\mathbf{u}_t^i - \sum_{j=0}^{d} \nu_j^d(W_t)\mathbf{z}^j, & \mathbf{x}_0 = \mathbf{x}, \\ (U., W.) \in \mathcal{A}^{0,0}, \end{cases} \\ \text{value fn.} & v^{0,0}(\mathbf{x}) = \displaystyle\sup_{(U., W.) \in \mathcal{A}^{0,0}} J^{0,0}(\mathbf{x}, U., W.). \end{cases}$$

Here we refer to $\mathcal{P}^{0,0}$ as the corporate-level problem, $\mathcal{P}^{\varepsilon,0}$ as the production-level problem, $\mathcal{P}^{0,\delta}$ as the marketing-level problem, and $\mathcal{P}^{\varepsilon,\delta}$ as the operational-level problem. Figure 1 shows the structure of this multilevel hierarchy. Note further that when we consider any two of the above problems (except $\mathcal{P}^{\varepsilon,0}$ vs. $\mathcal{P}^{0,\delta}$), we always use *upper level* to refer to the simpler problem and *lower level* to refer to the other. For example, between $\mathcal{P}^{0,0}$ and $\mathcal{P}^{\varepsilon,0}$, we say $\mathcal{P}^{0,0}$ is the upper-level problem and $\mathcal{P}^{\varepsilon,0}$ is the lower-level problem. The structure of the multilevel hierarchy is shown in Fig. 1, in which we use $(U., W.)^{0,0}$, $(\mathbf{u}., W)^{\varepsilon,0}$, and $(U., \mathbf{w})^{0,\delta}$ to denote near-optimal controls for $\mathcal{P}^{0,0}$, $\mathcal{P}^{\varepsilon,0}$, and $\mathcal{P}^{0,\delta}$, respectively. We also use $(\mathbf{u}., W.)^{0,0}_{\varepsilon,0} \in \mathcal{A}^{\varepsilon,0}$ and $(U., \mathbf{w}.)^{0,0}_{0,\delta} \in \mathcal{A}^{0,\delta}$ to

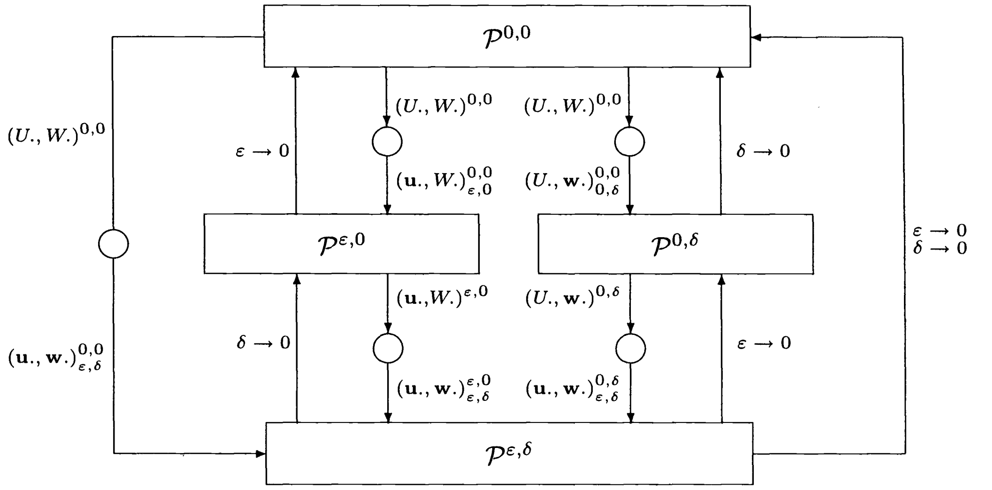

$\mathcal{P}^{0,0}$
$(U., W.)^{0,0}$
$(U., W.)^{0,0}$
$(U., W.)^{0,0}$
$\varepsilon \to 0$
$\delta \to 0$
$(\mathbf{u}., W.)^{0,0}_{\varepsilon,0}$
$(U., \mathbf{w}.)^{0,0}_{0,\delta}$
$\mathcal{P}^{\varepsilon,0}$
$\mathcal{P}^{0,\delta}$
$\varepsilon \to 0$
$\delta \to 0$
$(\mathbf{u}., W.)^{\varepsilon,0}$
$(U., \mathbf{w}.)^{0,\delta}$
$(\mathbf{u}., \mathbf{w}.)^{0,0}_{\varepsilon,\delta}$
$\delta \to 0$
$\varepsilon \to 0$
$(\mathbf{u}., \mathbf{w}.)^{\varepsilon,0}_{\varepsilon,\delta}$
$(\mathbf{u}., \mathbf{w}.)^{0,\delta}_{\varepsilon,\delta}$
$\mathcal{P}^{\varepsilon,\delta}$

FIG. 1. *Possible multilevel hierarchies.*

denote asymptotic optimal controls constructed from $(U., W.)^{0,0}$ for $\mathcal{P}^{\varepsilon,0}$ and $\mathcal{P}^{0,\delta}$, respectively. Similarly, $(\mathbf{u}., \mathbf{w}.)^{0,0}_{\varepsilon,\delta} \in \mathcal{A}^{\varepsilon,\delta}$, $(\mathbf{u}., \mathbf{w}.)^{\varepsilon,0}_{\varepsilon,\delta} \in \mathcal{A}^{\varepsilon,0}$, and $(\mathbf{u}., \mathbf{w}.)^{0,\delta}_{\varepsilon,\delta} \in \mathcal{A}^{\varepsilon,0}$, are asymptotic optimal controls constructed from $(U., W.)^{0,0}$, $(\mathbf{u}., W)^{\varepsilon,0}$, and $(U., \mathbf{w}.)^{0,\delta}$, respectively, for $\mathcal{P}^{\varepsilon,\delta}$. It should be noted that the construction of asymptotic optimal controls for the lower problems are certainly not unique.

3. Analysis of the value functions. In this section, we first state some elementary properties of the value functions and then study their asymptotic behavior as $\varepsilon \to 0$ and/or $\delta \to 0$.

We begin with the Lipschitz property of the value functions.

THEOREM 3.1. *Let $v(\mathbf{x}, \alpha, \mathbf{z})$ denote any of the four value functions $v^{\varepsilon,\delta}(\mathbf{x}, \alpha, \mathbf{z})$, $v^{\varepsilon,0}(\mathbf{x}, \alpha)$, $v^{0,\delta}(\mathbf{x}, \mathbf{z})$, and $v^{0,0}(\mathbf{x})$. Then for each $(\alpha, \mathbf{z}) \in \mathcal{M} \times \mathcal{Z}$, $v(\mathbf{x}, \alpha, \mathbf{z})$ has at most linear growth and is locally Lipschitz with the Lipschitz constant independent of ε and δ, i.e., for some constant C,*

$$0 \le v(\mathbf{x}, \alpha, \mathbf{z}) \le C(1 + |\mathbf{x}|) \text{ for all } \varepsilon \text{ and } \mathbf{x}, \tag{3.1}$$

and

$$|v(\mathbf{x}_1, \alpha, \mathbf{z}) - v(\mathbf{x}_2, \alpha, \mathbf{z})| \le C(1 + |\mathbf{x}_1| + |\mathbf{x}_2|)|\mathbf{x}_1 - \mathbf{x}_2| \tag{3.2}$$

for all $\mathbf{x}_1$ and $\mathbf{x}_2$.

Proof. The proof is similar to the one in [25, Thm. 3.1]. □

Next, we write Hamilton–Jacobi–Bellman (HJB) equations for the problems defined in the last section and claim that the value functions of the problems satisfy these equations. In writing these equations, however, we note that the notation $f_{\mathbf{x}}$ means the gradient of a function f with respect to $\mathbf{x}$ when it exists. Otherwise, the equations are understood to be in the sense of the viscosity solutions (see, e.g., [13] and [7]).

THEOREM 3.2. *The value functions* $v^{\varepsilon,\delta}(\mathbf{x},\alpha,\mathbf{z})$, $v^{\varepsilon,0}(\mathbf{x},\alpha)$, $v^{0,\delta}(\mathbf{x},\mathbf{z})$, *and* $v^{0,0}(\mathbf{x})$ *are viscosity solutions to the following* HJB *equations, respectively*:

(3.3)
$$\rho v^{\varepsilon,\delta}(\mathbf{x},\alpha,\mathbf{z}) = \max_{(\mathbf{u},\mathbf{w})\in\mathcal{U}_\alpha\times\mathcal{W}}\left\{(\mathbf{u}-\mathbf{z})\cdot v^{\varepsilon,\delta}_{\mathbf{x}}(\mathbf{x},\alpha,\mathbf{z}) + G(\mathbf{x},\mathbf{z},\mathbf{u},\mathbf{w}) \right.$$
$$\left. + \frac{1}{\varepsilon}Q^m(\mathbf{u})v^{\varepsilon,\delta}(\mathbf{x},\cdot,\mathbf{z})(\alpha) + \frac{1}{\delta}Q^d(\mathbf{w})v^{\varepsilon,\delta}(\mathbf{x},\alpha,\cdot)(\mathbf{z})\right\}$$

for any $(\alpha,\mathbf{z})\in\mathcal{M}\times\mathcal{Z}$;

(3.4)
$$\rho v^{\varepsilon,0}(\mathbf{x},\alpha) = \max_{(\mathbf{u},W)\in\mathcal{U}_\alpha\times\Gamma^d}\left\{(\mathbf{u}-\mathbf{z})\cdot v^{\varepsilon,0}_{\mathbf{x}}(\mathbf{x},\alpha) + \sum_{j=0}^{d}\nu^d_j(W)G(\mathbf{x},\mathbf{z}^j,\mathbf{u},\mathbf{w}^j)\right.$$
$$\left. + \frac{1}{\varepsilon}Q^m(\mathbf{u})v^{\varepsilon,0}(\mathbf{x},\cdot)(\alpha)\right\}$$

for any $\alpha\in\mathcal{M}$;

(3.5)
$$\rho v^{0,\delta}(\mathbf{x},\mathbf{z}) = \max_{(U,\mathbf{w})\in\Gamma^m\times\mathcal{W}}\left\{\sum_{i=0}^{m}\nu^m_i(U)(\mathbf{u}^i-\mathbf{z})\cdot v^{0,\delta}_{\mathbf{x}}(\mathbf{x},\mathbf{z})\right.$$
$$\left. + \sum_{i=0}^{m}\nu^m_i(U)G(\mathbf{x},\mathbf{z},\mathbf{u}^i,\mathbf{w}) + \frac{1}{\delta}Q^d(\mathbf{w})v^{0,\delta}(\mathbf{x},\cdot)(\mathbf{z})\right\}$$

for any $\mathbf{z}\in\mathcal{Z}$;

(3.6)
$$\rho v^{0,0}(\mathbf{x}) = \max_{(U,W)\in\Gamma^m\times\Gamma^d}\left\{\left(\sum_{i=0}^{m}\nu^m_i(U)\mathbf{u}^i - \sum_{j=0}^{d}\nu^d_j(W)\mathbf{z}^j\right)\cdot v^{0,0}_{\mathbf{x}}(\mathbf{x})\right.$$
$$\left. + \sum_{i=0,j=0}^{m,d}\nu^m_i(U)\nu^d_j(W)G(\mathbf{x},\mathbf{z}^j,\mathbf{u}^i,\mathbf{w}^j)\right\}.$$

Proof. The proof follows directly from Soner [32]. □

We now examine the asymptotic behavior of the value functions $v^{\varepsilon,\delta}(\mathbf{x},\alpha,\mathbf{z})$, $v^{\varepsilon,0}(\mathbf{x},\alpha)$, and $v^{0,\delta}(\mathbf{x},\mathbf{z})$, as ε goes to 0 and/or δ goes to 0.

THEOREM 3.3. *Under assumptions* (A1) *and* (A2), *the following hold*:

(i) $\lim_{\varepsilon,\delta\to 0} v^{\varepsilon,\delta}(\mathbf{x},\alpha,\mathbf{z}) = v^{0,0}(\mathbf{x})$;

(ii) $\lim_{\delta\to 0} v^{\varepsilon,\delta}(\mathbf{x},\alpha,\mathbf{z}) = v^{\varepsilon,0}(\mathbf{x},\alpha)$;

(iii) $\lim_{\varepsilon\to 0} v^{\varepsilon,\delta}(\mathbf{x},\alpha,\mathbf{z}) = v^{0,\delta}(\mathbf{x},\mathbf{z})$;

(iv) $\lim_{\varepsilon\to 0} v^{\varepsilon,0}(\mathbf{x},\alpha) = v^{0,0}(\mathbf{x})$;

(v) $\lim_{\delta\to 0} v^{0,\delta}(\mathbf{x},\mathbf{z}) = v^{0,0}(\mathbf{x})$.

Proof. We only prove (i), since the proofs for the others are similar. By [25, Thm. 4.1], it is sufficient to show the following: if there exists a sequence $(\varepsilon_l,\delta_l)\to 0$

$(l = 1, 2, \ldots)$ such that $v^{\varepsilon_l,\delta_l}(\mathbf{x}, \alpha, \mathbf{z}) \to v(\mathbf{x}, \alpha, \mathbf{z})$ as $l \to \infty$, then the limit function $v(\mathbf{x}, \alpha, \mathbf{z})$ is independent of $(\alpha, \mathbf{z})$.

If we examine the ratio of ε_l and δ_l, there are only three cases that may occur.

Case (a). $\liminf_{l\to\infty} \varepsilon_l/\delta_l = 0$.

Case (b). $\limsup_{l\to\infty} \varepsilon_l/\delta_l = \infty$.

Case (c). There exist constants c_1 and c_2 such that $0 < c_1 \le \varepsilon_l/\delta_l \le c_2 < \infty$ for all l.

We deal with each of these cases separately. We show that in each case, the limiting function $v(\mathbf{x}, \alpha, \mathbf{z})$ is independent of $(\alpha, \mathbf{z})$.

We consider Case (a) first. Note that Case (a) implies the existence of a further subsequence of $\{l\}$, still denoted by $\{l\}$, such that $\varepsilon_l/\delta_l \to 0$ as $l \to \infty$.

By assumption (A2), there exist $U^0 = (\mathbf{u}^0, \mathbf{u}^1, \ldots, \mathbf{u}^m) \in \Gamma^m$ and $W^0 = (\mathbf{w}^0, \mathbf{w}^1, \ldots, \mathbf{w}^d) \in \Gamma^d$ such that $\bar{Q}^m(U^0)$ and $\bar{Q}^d(W^0)$ are irreducible. Moreover, $\bar{Q}(U^0, W^0)$ defined in (2.6) is also irreducible.

For each fixed $\mathbf{x}$, it follows from Theorem 3.1 that $v^{\varepsilon_l,\delta_l}(\mathbf{x}, \alpha, \mathbf{z})$ is bounded on $(\alpha, \mathbf{z}) \in \mathcal{M} \times \mathcal{Z}$. Moreover, the subgradient $\partial_{\mathbf{x}} v^{\varepsilon_l,\delta_l}(\mathbf{x}, \alpha, \mathbf{z})$ is also a bounded set. Therefore by (3.3), we have

$$\begin{aligned}\rho v^{\varepsilon_l,\delta_l}(\mathbf{x}, i, \mathbf{z}^j) \ge\ & (\mathbf{u}^i - \mathbf{z}^j)\cdot\xi + G(\mathbf{x}, \mathbf{z}^j, \mathbf{u}^i, \mathbf{w}^j) \\ & + \frac{1}{\varepsilon_l}\bar{Q}^m(U^0)v^{\varepsilon_l,\delta_l}(\mathbf{x}, \cdot, \mathbf{z}^j)(i) + \frac{1}{\delta_l}\bar{Q}^d(W^0)v^{\varepsilon_l,\delta_l}(\mathbf{x}, i, \cdot)(\mathbf{z}^j)\end{aligned} \tag{3.7}$$

for all $\xi \in \partial_{\mathbf{x}} v^{\varepsilon_l,\delta_l}(\mathbf{x}, i, \mathbf{z}^j)$. Since $\varepsilon_l/\delta_l \to 0$, it follows that

$$\liminf_{l\to\infty} \bar{Q}^m(U^0)v^{\varepsilon_l,\delta_l}(\mathbf{x}, \cdot, \mathbf{z})(i) \le 0$$

for each $i \in \mathcal{M}$. Since $v^{\varepsilon_l,\delta_l}(\mathbf{x}, i, \mathbf{z}) \to v(\mathbf{x}, i, \mathbf{z})$,

$$\bar{Q}^m(U^0)v(\mathbf{x}, \cdot, \mathbf{z})(i) \le 0. \tag{3.8}$$

This system of inequalities (3.8) implies (cf. [25, Thm. 4.1]) that $v(\mathbf{x}, i, \mathbf{z})$ is independent of i. We may now denote $v(\mathbf{x}, i, \mathbf{z})$ by $v(\mathbf{x}, \mathbf{z})$ with a slight abuse of notation.

We next show that $v(\mathbf{x}, \alpha, \mathbf{z})$ is also independent of $\mathbf{z}$. Indeed, if we multiply the inequality in (3.7) that corresponds to $(\alpha, \mathbf{z}) = (i, \mathbf{z}^j)$ by $\nu_i^m(U^0)$, sum over $i \in \mathcal{M}$, and use the fact that $\sum_{i=0}^m \nu_i^m(U^0)\bar{Q}^m(U^0)v^{\varepsilon,\delta}(\mathbf{x}, \cdot, \mathbf{z}^j)(i) = 0$, then we shall have

$$\begin{aligned}\rho\sum_{i=0}^m \nu_i^m(U^0)v^{\varepsilon_l,\delta_l}(\mathbf{x}, i, \mathbf{z}^j) \ge\ & \sum_{i=0}^m \nu_i^m(U^0)(\mathbf{u}^i - \mathbf{z}^j)\cdot\partial_{\mathbf{x}} v^{\varepsilon_l,\delta_l}(\mathbf{x}, i, \mathbf{z}^j) \\ & + \sum_{i=0}^m \nu_i^m(U^0)G(\mathbf{x}, \mathbf{z}^j, \mathbf{u}^i, \mathbf{w}^j) \\ & + \frac{1}{\delta_l}\bar{Q}^d(W^0)\sum_{i=0}^m \nu_i^m(U^0)v^{\varepsilon_l,\delta_l}(\mathbf{x}, i, \cdot)(\mathbf{z}^j).\end{aligned}$$

Multiplying by δ_l and taking the limit yields

$$\liminf_{l\to\infty} \bar{Q}^d(W^0)\sum_{i=0}^m \nu_i^m(U^0)v^{\varepsilon_l,\delta_l}(\mathbf{x}, i, \cdot)(\mathbf{z}^j) \le 0$$

for each $\mathbf{z}^j \in \mathcal{Z}$. By using $v^{\varepsilon_l,\delta_l}(\mathbf{x}, \alpha, \mathbf{z}) \to v(\mathbf{x}, \mathbf{z})$, we have

$$\bar{Q}^d(W^0)v(\mathbf{x}, \cdot)(\mathbf{z}) \le 0.$$

This implies that $v(\mathbf{x}, \mathbf{z})$ is independent of $\mathbf{z}$. Hence, if Case (a) holds, then $v(\mathbf{x}, \alpha, \mathbf{z})$ is independent of $(\alpha, \mathbf{z})$.

If Case (b) holds, i.e., $\limsup_{l \to \infty} \varepsilon_l/\delta_l = \infty$, then we have $\liminf_{l \to \infty} \delta_l/\varepsilon_l = 0$. If we exchange the role of ε_l and δ_l, then there exists a further subsequence of $\{l\}$ such that $\delta_l/\varepsilon_l \to 0$. We can thus repeat the argument used in Case (a) to show that the limiting function $v(\mathbf{x}, \alpha, \mathbf{z})$ is independent of $(\alpha, \mathbf{z})$, also in Case (b).

Finally, in Case (c), there exists a further subsequence of $\{l\}$, still denoted by $\{l\}$, such that $\varepsilon_l/\delta_l \to a > 0$ as $l \to \infty$. Without loss of generality, we may assume $a = 1$. Then, as in Case (a), we have

$$\bar{Q}(U^0, W^0)v(\mathbf{x}, \cdot, \cdot)(\alpha, \mathbf{z}) \leq 0$$

for all $(\alpha, \mathbf{z}) \in \mathcal{M} \times \mathcal{Z}$. Then by the irreducibility of $\bar{Q}(U^0, W^0)$, we conclude that $v(\mathbf{x}, \alpha, \mathbf{z})$ is independent of $(\alpha, \mathbf{z})$. This completes the proof. □

4. Asymptotic behavior of capacity and demand processes. The purpose of this section is to analyze the asymptotic behavior of the process $(\alpha(\varepsilon, t), \mathbf{z}(\delta, t))$ as $\varepsilon \to 0$ and/or $\delta \to 0$. The analysis will allow us to use optimal controls of lower-level problems for constructing asymptotically optimal controls for the corresponding upper-level problems.

Let χ_D denote the indicator function of set D and let

$$\begin{aligned}
&\psi^\varepsilon(t) = (\chi_{\{\alpha(\varepsilon,t)=0\}}, \ldots, \chi_{\{\alpha(\varepsilon,t)=m\}}),\\
&\psi^\delta(t) = (\chi_{\{\mathbf{z}(\delta,t)=\mathbf{z}^1\}}, \ldots, \chi_{\{\mathbf{z}(\delta,t)=\mathbf{z}^d\}}),\\
\text{and}\quad &\psi^{\varepsilon,\delta}(t) = (\chi_{\{\alpha(\varepsilon,t)=0\}}\psi^\delta(t), \ldots, \chi_{\{\alpha(\varepsilon,t)=m\}}\psi^\delta(t)).
\end{aligned}$$

From (2.7) it is easy to see that

$$\psi^{\varepsilon,\delta}(t) - \psi^{\varepsilon,\delta}(s) - \int_s^t \psi^{\varepsilon,\delta}(r)\bar{Q}(U_r, W_r)dr \tag{4.1}$$

is a martingale. But

$$\begin{aligned}
\psi^{\varepsilon,\delta}(t)\bar{Q}(U_t, W_t) = \frac{1}{\varepsilon}\Big\{ &(q_{00}^m(\mathbf{u}^0)\chi_{\{\alpha(\varepsilon,t)=0\}} + \cdots + q_{m0}^m(\mathbf{u}^m)\chi_{\{\alpha(\varepsilon,t)=m\}})\psi^\delta(t),\\
&\ldots, (q_{0m}^m(\mathbf{u}^0)\chi_{\{\alpha(\varepsilon,t)=0\}} + \cdots + q_{mm}^m(\mathbf{u}^m)\chi_{\{\alpha(\varepsilon,t)=m\}})\psi^\delta(t)\Big\}\\
&+ \frac{1}{\delta}\Big\{\chi_{\{\alpha(\varepsilon,t)=0\}}\psi^\delta(t)\bar{Q}^d(W_t), \ldots, \chi_{\{\alpha(\varepsilon,t)=m\}}\psi^\delta(t)\bar{Q}^d(W_t)\Big\}.
\end{aligned}$$

This implies that for each $i \in \mathcal{M}$,

$$\begin{aligned}
&\chi_{\{\alpha(\varepsilon,t)=i\}}\psi^\delta(t) - \chi_{\{\alpha(\varepsilon,s)=i\}}\psi^\delta(s)\\
&\quad - \int_s^t \{\varepsilon^{-1}[(q_{0i}^m(\mathbf{u}_r^0)\chi_{\{\alpha(\varepsilon,r)=0\}} + \cdots + q_{mi}^m(\mathbf{u}_r^m)\chi_{\{\alpha(\varepsilon,r)=m\}})\psi^\delta(r)]\\
&\qquad + \delta^{-1}\chi_{\{\alpha(\varepsilon,t)=i\}}\psi^\delta(t)\bar{Q}^d(W_r)\}\, dr
\end{aligned}$$

is a martingale. Summing up over $i \in \mathcal{M}$, we conclude that

$$\psi^\delta(t) - \psi^\delta(s) - \int_s^t \frac{1}{\delta}\psi^\delta(r)\bar{Q}^d(W_r)dr \tag{4.2}$$

is a martingale.

Similarly, we can show that

$$\psi^\varepsilon(t) - \psi^\varepsilon(s) - \int_s^t \frac{1}{\varepsilon}\psi^\varepsilon(r)\bar{Q}^m(U_r)dr \tag{4.3}$$

is also a martingale.

Now define $P^{\varepsilon,\delta}(t) = E\psi^{\varepsilon,\delta}(t)$. Then,

$$\begin{aligned}P^{\varepsilon,\delta}(t) = (\,&P((\alpha(\varepsilon,t),\mathbf{z}(\delta,t)) = (0,\mathbf{z}^1)),\dots,P((\alpha(\varepsilon,t),\mathbf{z}(\delta,t)) = (0,\mathbf{z}^d)),\\ &\dots,P((\alpha(\varepsilon,t),\mathbf{z}(\delta,t)) = (m,\mathbf{z}^1)),\dots,P((\alpha(\varepsilon,t),\mathbf{z}(\delta,t)) = (m,\mathbf{z}^d))).\end{aligned}$$

By (4.1), we see that

$$P^{\varepsilon,\delta}(t) = P^{\varepsilon,\delta}(s) + \int_s^t P^{\varepsilon,\delta}(r)\bar{Q}(U_r,W_r)dr. \tag{4.4}$$

Similarly, if we let $P^\varepsilon(t) = E\psi^\varepsilon(t)$ and $P^\delta(t) = E\psi^\delta(t)$ so that

$$\begin{aligned}&P^\varepsilon(t) = (P(\alpha(\varepsilon,t) = 0),\dots,P(\alpha(\varepsilon,t) = m))\\ \text{and}\quad &P^\delta(t) = (P(\mathbf{z}(\delta,t) = \mathbf{z}^1),\dots,P(\mathbf{z}(\delta,t) = \mathbf{z}^d)),\end{aligned}$$

we can derive

$$P^\varepsilon(t) = P^\varepsilon(s) + \int_s^t P^\varepsilon(r)\bar{Q}^m(U_r)dr \text{ and } P^\delta(t) = P^\delta(s) + \int_s^t P^\delta(r)\bar{Q}^d(W_r)dr.$$

Let $\boldsymbol{\nu}^m(t) := \boldsymbol{\nu}^m(U_t)$, $\boldsymbol{\nu}^d(t) := \boldsymbol{\nu}^d(W_t)$, and $\boldsymbol{\nu}(t) := (\nu_0^m(t)\boldsymbol{\nu}^d(t),\dots,\nu_m^m(t)\boldsymbol{\nu}^d(t))$. Let $L^2([s,T])$, $0 \le s \le T$, denote the space of all functions $f : [s,T] \to R^1$ that are square integrable. Then, we have the following two lemmas.

LEMMA 4.1. *For each $s \in [0,T]$, $P^{\varepsilon,\delta}(t)$, $P^\varepsilon(t)$, and $P^\delta(t)$ converge weakly to $\boldsymbol{\nu}(t)$, $\boldsymbol{\nu}^m(t)$, and $\boldsymbol{\nu}^d(t)$ on $L^2([s,T])$, respectively. That is, for each $f(t) \in L^2([0,T])$, we have*

$$\begin{aligned}&\text{(i)}\quad \lim_{\varepsilon\to 0}\int_s^T [P^\varepsilon(t) - \boldsymbol{\nu}^m(t)]f(t)dt = 0;\\ &\text{(ii)}\quad \lim_{\delta\to 0}\int_s^T [P^\delta(t) - \boldsymbol{\nu}^d(t)]f(t)dt = 0;\\ &\text{(iii)}\quad \lim_{\varepsilon,\delta\to 0}\int_s^T [P^{\varepsilon,\delta}(t) - \boldsymbol{\nu}(t)]f(t)dt = 0.\end{aligned}$$

Proof. First of all, (i) and (ii) can be proved as in [25, Lem. 5.1]. Thus we need only to show (iii). Let $(\varepsilon_l,\delta_l,\ l = 1,2,\dots)$ denote a subsequence of $(\varepsilon,\delta) \to 0$. We have the same three possible cases as in Theorem 3.3. We deal with these cases one by one.

If Case (a) holds, then there exists a further subsequence of $\{l\}$, still denoted by $\{l\}$, such that $\varepsilon_l/\delta_l \to 0$ as $l \to \infty$. Note that since $P^{\varepsilon,\delta}(t) \in L^2([0,T])$, there exists (cf. [35, Thm. 1, p. 126]) yet another subsequence of $(\varepsilon_l,\delta_l) \to 0$ such that $\{P^{\varepsilon_l,\delta_l}(t)\}_{[0,T]}$ converges weakly to some

$$P^{0,0}(t) = (p_{00}(t),\dots,p_{0d}(t),\dots,p_{m0}(t),\dots,p_{md}(t)) \in L^2([0,T]),$$

i.e.,

$$\int_0^T [P^{\varepsilon,\delta}(r) - P^{0,0}(r)]f(r)dr \to 0$$

for any $f(\cdot) \in L^2([0,T])$. Moreover, $0 \leq P^{0,0}(t) \leq 1$ and $\sum_{i,j}^{m,d} p_{ij}(t) = 1$ a.e.

Let

$$\tilde{Q}^m(U_t) = \begin{pmatrix} q_{00}^m(\mathbf{u}_t^0)I & q_{01}^m(\mathbf{u}_t^0)I & \cdots & q_{0m}^m(\mathbf{u}_t^0)I \\ q_{10}^m(\mathbf{u}_t^1)I & q_{11}^m(\mathbf{u}_t^1)I & \cdots & q_{1m}^m(\mathbf{u}_t^1)I \\ \vdots & \vdots & \cdots & \vdots \\ q_{m0}^m(\mathbf{u}_t^m)I & q_{m1}^m(\mathbf{u}_t^m)I & \cdots & q_{mm}^m(\mathbf{u}_t^m)I \end{pmatrix}.$$

It is easy to see that $\tilde{Q}^m(U_t) \in L^2([0,T])$. This implies that for $0 \leq s \leq t \leq T$, we have

$$\int_s^t [P^{\varepsilon,\delta}(r) - P^{0,0}(r)]\tilde{Q}^m(U_r)dr \to 0.$$

Thus, by noting (4.4) and the fact that $\varepsilon_l/\delta_l \to 0$, we obtain

$$\int_s^t P^{0,0}(r)\tilde{Q}^m(U_r)dr = 0.$$

Since s and t are arbitrary, it follows immediately that

$$P^{0,0}(t)\tilde{Q}^m(U_t) = 0, \text{ a.e.}$$

From this and the irreducibility of $\bar{Q}^m(U_t)$, which is related to $\tilde{Q}^m(U_t)$, one can conclude

$$\begin{cases} (p_{00}(t), \ldots, p_{m0}(t)) = p_0(t)\boldsymbol{\nu}^m(t), \\ \qquad \cdots \\ (p_{0d}(t), \ldots, p_{md}(t)) = p_d(t)\boldsymbol{\nu}^m(t), \\ \qquad p_0(t) + \cdots + p_d(t) = 1 \end{cases}$$

for some $p_j(t) \geq 0$. By (i) of Lemma 4.1,

$$P(\mathbf{z}(\delta,t) = \mathbf{z}^j) \to \nu_j^d(t).$$

Since

$$P(\mathbf{z}(\delta,t) = \mathbf{z}^j) = P((\alpha(\varepsilon,t), \mathbf{z}(\delta,t)) = (0, \mathbf{z}^j)) + \cdots + P((\alpha(\varepsilon,t), \mathbf{z}(\delta,t)) = (m, \mathbf{z}^j)),$$

it follows that $p_{0j}(t) + \cdots + p_{mj}(t) = \nu_j^d(t)$. This implies $p_j(t) = \nu_j^d(t)$. Thus,

$$p_{ij}(t) = \nu_i^m(t)\nu_j^d(t) \text{ a.e. for all } i,j.$$

Since the weak limit of $P^{\varepsilon,\delta}(t)$ equals $\boldsymbol{\nu}(t)$, which is independent of the choice of the subsequences of $(\varepsilon_l, \delta_l)$, it holds that $P^{\varepsilon,\delta}(t) \to \boldsymbol{\nu}(t)$ weakly.

Similarly, we can show that Lemma 4.1(iii) holds in Case (b) if we reverse the role of ε and δ. Finally, if Case (c) holds, the above argument goes through by noting the irreducibility of $\bar{Q}(U_t, W_t)$. □

LEMMA 4.2. *For any bounded deterministic process $\beta(s)$ and for each $i \in \mathcal{M}$ and $t \geq 0$,*

$$E\left|\int_0^t (\chi_{\{\alpha(\varepsilon,s)=i\}} - \nu_i^m(s))\beta(s)ds\right|^2 \to 0 \text{ as } \varepsilon \to 0,$$

$$E\left|\int_0^t (\chi_{\{\mathbf{z}(\delta,s)=\mathbf{z}^j\}} - \nu_i^d(s))\beta(s)ds\right|^2 \to 0 \text{ as } \delta \to 0.$$

Moreover, for any $0 \leq s \leq T$,

$$P(\alpha(\varepsilon,t) = i|\sigma\{\mathbf{z}(\delta,r) : \, r \leq s\}) \to \nu_j^m(U_t) \textit{weakly on } L^2([s,T]) \textit{as } \varepsilon \to 0 \text{ a.s.},$$
$$P(\mathbf{z}(\delta,t) = \mathbf{z}^j|\sigma\{\alpha(\varepsilon,r) : r \leq s\}) \to \nu_j^d(W_t) \textit{weakly on } L^2([s,T]) \textit{as } \delta \to 0 \text{ a.s.}$$

Proof. The proof is as in [25, Lem. 5.1]. □

5. Asymptotically optimal open-loop controls. In this section, we construct asymptotically optimal open-loop controls for the lower-level problems from optimal controls of the corresponding upper-level problems. Here open-loop controls refer to partially open-loop controls, i.e., controls that respond to the machine state but not to the surplus state. Feedback controls are considered in the next section.

THEOREM 5.1 (open-loop controls; $\mathcal{P}^{\varepsilon,\delta}$ vs. $\mathcal{P}^{0,0}$). *Let $(U_\cdot, W_\cdot) \in \mathcal{A}^{0,0}$ denote an optimal control for the upper-level problem $\mathcal{P}^{0,0}$. Construct*

$$(\mathbf{u}_t^{\varepsilon,\delta}, \mathbf{w}_t^{\varepsilon,\delta}) = \left(\sum_{i=0}^m \chi_{\{\alpha(\varepsilon,t)=i\}}\mathbf{u}_t^i, \sum_{j=0}^d \chi_{\{\mathbf{z}(\delta,t)=\mathbf{z}^j\}}\mathbf{w}_t^j\right).$$

Then, $(\mathbf{u}_\cdot^{\varepsilon,\delta}, \mathbf{w}_\cdot^{\varepsilon,\delta}) \in \mathcal{A}^{\varepsilon,\delta}$. Furthermore, it is asymptotically optimal for the original lower-level problem $\mathcal{P}^{\varepsilon,\delta}$, i.e.,

$$\lim_{\varepsilon,\delta\to 0} |J^{\varepsilon,\delta}(\mathbf{x}, \alpha, \mathbf{z}, \mathbf{u}_\cdot^{\varepsilon,\delta}, \mathbf{w}_\cdot^{\varepsilon,\delta}) - v^{\varepsilon,\delta}(\mathbf{x}, \alpha, \mathbf{z})| = 0.$$

Proof. The proof of this theorem can be given as in [25, Thm. 6.1]. We provide an outline of the proof here, however, for the sake of completeness. From the procedure of constructing piecewise-deterministic processes described in [25], it is not difficult to see that the generators $Q(\mathbf{u}_t^{\varepsilon,\delta}, \mathbf{w}_t^{\varepsilon,\delta})$ and $\bar{Q}(U_t, W_t)$ both generate the same process $(\alpha(\varepsilon,t), \mathbf{z}(\delta,t))$. Thus, $(\mathbf{u}_t^{\varepsilon,\delta}, \mathbf{w}_t^{\varepsilon,\delta}) \in \mathcal{A}^{\varepsilon,\delta}$. Since $\lim_{\varepsilon,\delta\to 0} v^{\varepsilon,\delta}(\mathbf{x}, \alpha, \mathbf{z}) = v^{0,0}(\mathbf{x})$ and $J^{\varepsilon,\delta}(\mathbf{x}, \alpha, \mathbf{z}, \mathbf{u}_\cdot^{\varepsilon,\delta}, \mathbf{w}_\cdot^{\varepsilon,\delta}) \leq v^{\varepsilon,\delta}(\mathbf{x}, \alpha, \mathbf{z})$, it suffices to show that

$$\liminf_{\varepsilon,\delta\to 0} J^{\varepsilon,\delta}(\mathbf{x}, \alpha, \mathbf{z}, \mathbf{u}_\cdot^{\varepsilon,\delta}, \mathbf{w}_\cdot^{\varepsilon,\delta}) \geq v^{0,0}(\mathbf{x}). \tag{5.1}$$

Let $(U_\cdot, W_\cdot) = ((\mathbf{u}_\cdot^0, \ldots, \mathbf{u}_\cdot^m), (\mathbf{w}_\cdot^0, \ldots, \mathbf{w}_\cdot^d)) \in \mathcal{A}^{0,0}$ be an open-loop optimal control for $\mathcal{P}^{0,0}$ and let $(\mathbf{u}_t^{\varepsilon,\delta}, \mathbf{w}_t^{\varepsilon,\delta}) = (\sum_{i=0}^m \chi_{\{\alpha(\varepsilon,t)=i\}}\mathbf{u}_t^i, \sum_{j=0}^d \chi_{\{\mathbf{z}(\delta,t)=\mathbf{z}^j\}}\mathbf{w}^j)$. Let $\mathbf{x}_\cdot$ and $\bar{\mathbf{x}}_\cdot$ denote the corresponding state trajectories of the systems $\mathcal{P}^{\varepsilon,\delta}$ and $\mathcal{P}^{0,0}$ with

TABLE 1
Asymptotically optimal open-loop controls.

Hierarchy: lower vs. upper	Optimal control for upper level	Asymptotically optimal control constructed for lower-level problems
$\mathcal{P}^{\varepsilon,\delta}$ vs. $\mathcal{P}^{0,0}$	$(U_\cdot, W_\cdot) \in \mathcal{A}^{0,0}$	$\left(\sum_{i=0}^{m} \chi_{\{\alpha=i\}}\mathbf{u}_t^i, \sum_{j=0}^{d} \chi_{\{\mathbf{z}=\mathbf{z}^j\}}\mathbf{w}_t^j\right) \in \mathcal{A}^{\varepsilon,\delta}$
$\mathcal{P}^{\varepsilon,\delta}$ vs. $\mathcal{P}^{\varepsilon,0}$	$(\mathbf{u}_\cdot, W_\cdot) \in \mathcal{A}^{\varepsilon,0}$	$\left(\mathbf{u}_t, \sum_{j=0}^{d} \chi_{\{\mathbf{z}=\mathbf{z}^j\}}\mathbf{w}_t^j\right) \in \mathcal{A}^{\varepsilon,\delta}$
$\mathcal{P}^{\varepsilon,\delta}$ vs. $\mathcal{P}^{0,\delta}$	$(U_\cdot, \mathbf{w}_\cdot) \in \mathcal{A}^{0,\delta}$	$\left(\sum_{i=0}^{m} \chi_{\{\alpha=i\}}\mathbf{u}_t^i, \mathbf{w}_t\right) \in \mathcal{A}^{\varepsilon,\delta}$
$\mathcal{P}^{\varepsilon,0}$ vs. $\mathcal{P}^{0,0}$	$(U_\cdot, W_\cdot) \in \mathcal{A}^{0,0}$	$\left(\sum_{i=0}^{m} \chi_{\{\alpha=i\}}\mathbf{u}_t^i, W_t\right) \in \mathcal{A}^{\varepsilon,0}$
$\mathcal{P}^{0,\delta}$ vs. $\mathcal{P}^{0,0}$	$(U_\cdot, W_\cdot) \in \mathcal{A}^{0,0}$	$\left(U_t, \sum_{j=0}^{d} \chi_{\{\mathbf{z}=\mathbf{z}^j\}}\mathbf{w}_t^j\right) \in \mathcal{A}^{0,\delta}$

the same initial $\mathbf{x}$, respectively. Then by Lemma 4.2, we can show that $E|\mathbf{x}_t - \bar{\mathbf{x}}_t| \to 0$ as $\varepsilon, \delta \to 0$.

Moreover, by assumption (A1),

$$\liminf_{\varepsilon,\delta\to 0} E\int_0^\infty e^{-\rho t}[G(\mathbf{x}_t, \mathbf{z}(\delta,t), \mathbf{u}_t^{\varepsilon,\delta}, \mathbf{w}_t^{\varepsilon,\delta}) - G(\bar{\mathbf{x}}_t, \mathbf{z}(\delta,t), \mathbf{u}_t^{\varepsilon,\delta}, \mathbf{w}_t^{\varepsilon,\delta})]dt \geq 0.$$

It follows that

$$\liminf_{\varepsilon,\delta\to 0} J^{\varepsilon,\delta}(\mathbf{x}, \alpha, \mathbf{z}, \mathbf{u}_\cdot^{\varepsilon,\delta}, \mathbf{w}_\cdot^{\varepsilon,\delta}) \geq \liminf_{\varepsilon,\delta\to 0} E\int_0^\infty e^{-\rho t} G(\bar{\mathbf{x}}_t, \mathbf{z}(\delta,t), \mathbf{u}_t^{\varepsilon,\delta}, \mathbf{w}_t^{\varepsilon,\delta})dt.$$

Now, by our construction of $(\mathbf{u}_\cdot^{\varepsilon,\delta}, \mathbf{w}_\cdot^{\varepsilon,\delta})$ and by Lemma 4.1, we have

$$\begin{aligned} & E\int_0^\infty e^{-\rho t} G(\bar{\mathbf{x}}_t, \mathbf{z}(\delta,t), \mathbf{u}_t^{\varepsilon,\delta}, \mathbf{w}_t^{\varepsilon,\delta})dt \\ & \to \int_0^\infty e^{-\rho t}\left(\sum_{i=0,j=0}^{m,d} \nu_i^m(U_t)\nu_j^d(W_t)G(\bar{\mathbf{x}}_t, \mathbf{z}^j, \mathbf{u}_t^i, \mathbf{w}_t^j)\right)dt = v^{0,0}(\mathbf{x}). \end{aligned}$$

This implies (5.1) and completes the proof. □

In Theorem 5.1, the lower-level problem is $\mathcal{P}^{\varepsilon,\delta}$ and the upper-level problem is $\mathcal{P}^{0,0}$. The idea is to construct an asymptotically optimal solution for the lower-level problem from an optimal solution of the upper-level problem. The result is summarized in Table 1 in the $\mathcal{P}^{\varepsilon,\delta}$ vs. $\mathcal{P}^{0,0}$ row, which is the first row. Similar results can be proved in similar ways for other combinations of lower- and upper-level problems. These are also summarized in Table 1 in rows 2–4.

Remark 5.1. Another possible method for obtaining a lower-level decision is to resolve the lower-level problem given the upper-level decision. In view of the results proved in this section, it should be obvious that this method would also provide an asymptotically optimal solution; see also [21].

6. Asymptotically optimal feedback controls. We now consider feedback controls. We begin with the hierarchy $\mathcal{P}^{\varepsilon,\delta}$ vs. $\mathcal{P}^{0,0}$. Let

$$(U(\mathbf{x}), W(\mathbf{x})) = (\mathbf{u}^0(\mathbf{x}), \dots, \mathbf{u}^m(\mathbf{x}), \mathbf{w}^0(\mathbf{x}), \dots, \mathbf{w}^d(\mathbf{x}))$$

denote an optimal feedback control for $\mathcal{P}^{0,0}$. This is obtained by maximizing the right-hand side of (3.6), i.e.,

$$\begin{aligned} &\left(\sum_{i=0}^{m} \nu_i^m(U(\mathbf{x}))\mathbf{u}^i(\mathbf{x}) - \sum_{j=0}^{d} \nu_j^d(W(\mathbf{x}))\mathbf{z}^j\right) \cdot v_{\mathbf{x}}^{0,0}(\mathbf{x}) \\ &\quad + \sum_{i=0,j=0}^{m,d} \nu_i^m(U(\mathbf{x}))\nu_j^d(W(\mathbf{x}))G(\mathbf{x}, \mathbf{z}^j, \mathbf{u}^i(\mathbf{x}), \mathbf{w}^j(\mathbf{x})) \\ &= \max_{(U,W)\in\Gamma^m\times\Gamma^d} \left\{ \left(\sum_{i=0}^{m} \nu_i^m(U)\mathbf{u}^i - \sum_{j=0}^{d} \nu_j^d(W)\mathbf{z}^j\right) \cdot v_{\mathbf{x}}^{0,0}(\mathbf{x}) \right. \\ &\qquad \left. + \sum_{i=0,j=0}^{m,d} \nu_i^m(U)\nu_j^d(W)G(\mathbf{x}, \mathbf{z}^j, \mathbf{u}^i, \mathbf{w}^j) \right\}. \end{aligned} \tag{6.1}$$

We then construct a control

$$\begin{cases} \mathbf{u}(\mathbf{x}, \alpha, \mathbf{z}) = \displaystyle\sum_{i=0}^{m} \chi_{\{\alpha=i\}}\mathbf{u}^i(\mathbf{x}), \\ \mathbf{w}(\mathbf{x}, \alpha, \mathbf{z}) = \displaystyle\sum_{j=0}^{d} \chi_{\{\mathbf{z}=\mathbf{z}^j\}}\mathbf{w}^j(\mathbf{x}), \end{cases} \tag{6.2}$$

which is clearly feasible (satisfies the control constraints) for $\mathcal{P}^{\varepsilon,\delta}$. Moreover, if $(U(\mathbf{x}), W(\mathbf{x}))$ is locally Lipschitz, then the system $\dot{\mathbf{x}}_t = \mathbf{u}(\mathbf{x}_t, \alpha(\varepsilon, t), \mathbf{z}(\delta, t)) - \mathbf{z}(\delta, t)$, $\mathbf{x}_0 = \mathbf{x}$, has a unique solution and, therefore,

$$(\mathbf{u}_t, \mathbf{w}_t) = (\mathbf{u}(\mathbf{x}_t, \alpha(\varepsilon, t), \mathbf{z}(\delta, t)), \mathbf{w}(\mathbf{x}_t, \alpha(\varepsilon, t), \mathbf{z}(\delta, t)))$$

is also admissible for $\mathcal{P}^{\varepsilon,\delta}$. Discussions concerning the existence of such locally Lipschitz feedback controls can be found in [12], [28].

We need an additional assumption to show that $(\mathbf{u}(\mathbf{x}, \alpha, \mathbf{z}), \mathbf{w}(\mathbf{x}, \alpha, \mathbf{z}))$ is asymptotically optimal for $\mathcal{P}^{\varepsilon,\delta}$.

(A3) The following equation has a unique solution:

$$\bar{\mathbf{x}}_t = \mathbf{x} + \int_0^t \left(\sum_{i=0}^{m} \nu_i^m(U(\bar{\mathbf{x}}_r))\mathbf{u}^i(\bar{\mathbf{x}}_r) - \sum_{j=0}^{d} \nu_j^d(W(\bar{\mathbf{x}}_r))\mathbf{z}^j\right) dr, \ \bar{\mathbf{x}}_0 = \mathbf{x}. \tag{6.3}$$

A sufficient condition for this is that $(\boldsymbol{\nu}^m(U), \boldsymbol{\nu}^d(W))$ is locally Lipschitz in (U, W).

THEOREM 6.1 (feedback controls; $\mathcal{P}^{\varepsilon,\delta}$ vs. $\mathcal{P}^{0,0}$). *Assume* (A1)–(A3). *Suppose the feedback control* $(U(\mathbf{x}), W(\mathbf{x})) \in \mathcal{A}^{0,0}$ *is locally Lipschitz. Then, the feedback control constructed in* (6.2) *is asymptotically optimal for* $\mathcal{P}^{\varepsilon,\delta}$, *i.e.*,

$$\lim_{\varepsilon,\delta\to 0} |J^{\varepsilon,\delta}(\mathbf{x}, \alpha, \mathbf{z}, \mathbf{u}(\mathbf{x}_\cdot, \alpha(\varepsilon, \cdot), \mathbf{z}(\delta, \cdot)), \mathbf{w}(\mathbf{x}_\cdot, \alpha(\varepsilon, \cdot), \mathbf{z}(\delta, \cdot))) - v^{\varepsilon,\delta}(\mathbf{x}, \alpha, \mathbf{z})| = 0.$$

Proof. Since $\bar{\mathbf{x}}_t$ is the only solution to (6.3), we can show as in Lemma 4.1 and [25, Lem. 5.2] that

$$\psi^{\varepsilon,\delta}(t) \to \nu_i^m(U(\bar{\mathbf{x}}_t))\nu_j^d(W(\bar{\mathbf{x}}_t)) \text{ as } \varepsilon, \delta \to 0.$$

The rest of the proof follows exactly as in [25, Thm. 6.2]. □

Next we consider the hierarchy $\mathcal{P}^{\varepsilon,\delta}$ vs. $\mathcal{P}^{\varepsilon,0}$, which arises when ε is fixed and δ is small. While asymptotically optimal controls for $\mathcal{P}^{\varepsilon,\delta}$ can be constructed from the optimal control of $\mathcal{P}^{\varepsilon,0}$ in a way similar to that in (6.2), the proof of asymptotic optimality is quite different from that in [25] due to the fact that the limiting problem is no longer deterministic as it is in [25]. The uniqueness of the solution to the limiting problem, therefore, needs to be specified by the probability distribution of the underlying processes.

Let $(\mathbf{u}(\mathbf{x},\alpha), W(\mathbf{x},\alpha)) \in \mathcal{A}^{\varepsilon,0}$ denote an optimal feedback control for $\mathcal{P}^{\varepsilon,0}$ obtained by maximizing the right-hand side of (3.4). We need to make the following assumption:

(A3)$'$ The ordinary differential equation

$$\dot{\bar{\mathbf{x}}}_t = \mathbf{u}(\bar{\mathbf{x}}_t, \bar{\alpha}(\varepsilon,t)) - \sum_{j=0}^{d} \nu_j^d(W(\bar{\mathbf{x}}_t, \bar{\alpha}(\varepsilon,t)))\mathbf{z}^j, \ \bar{\mathbf{x}}_0 = \mathbf{x} \tag{6.4}$$

has a unique solution for any given $\bar{\alpha}(\varepsilon,t)$ generated by $\varepsilon^{-1}\bar{Q}^m(U(\bar{\mathbf{x}}_t))$ and $\alpha(\varepsilon,0) = \alpha$, where

$$U(\mathbf{x}) = (\mathbf{u}(\mathbf{x},0),\ldots,\mathbf{u}(\mathbf{x},m)). \tag{6.5}$$

A sufficient condition for (A3)$'$ is that $(\mathbf{u}(\mathbf{x},\alpha), W(\mathbf{x},\alpha))$ is locally Lipschitz in $\mathbf{x}$ for each α and $\boldsymbol{\nu}^d(W)$ is locally Lipschitz in W.

THEOREM 6.2 (feedback controls; $\mathcal{P}^{\varepsilon,\delta}$ vs. $\mathcal{P}^{\varepsilon,0}$). *Assume* (A1), (A2), *and* (A3)$'$. *Suppose the optimal feedback control* $(\mathbf{u}(\mathbf{x},\alpha), W(\mathbf{x},\alpha)) \in \mathcal{A}^{\varepsilon,0}$ *is locally Lipschitz for each* α. *Let*

$$\begin{cases} \mathbf{u}_t^{\varepsilon,\delta} = \mathbf{u}(\mathbf{x}_t, \alpha(\varepsilon,t), \mathbf{z}(\delta,t)), \\ \mathbf{w}_t^{\varepsilon,\delta} = \displaystyle\sum_{j=0}^{d} \chi_{\{\mathbf{z}(\delta,t)=\mathbf{z}^j\}}\mathbf{w}^j(\mathbf{x}_t, \alpha(\varepsilon,t)). \end{cases} \tag{6.6}$$

Then, the feedback control $(\mathbf{u}_\cdot^{\varepsilon,\delta}, \mathbf{w}_\cdot^{\varepsilon,\delta})$ *is asymptotically optimal for* $\mathcal{P}^{\varepsilon,\delta}$, *i.e.*,

$$\lim_{\delta\to 0} |J^{\varepsilon,\delta}(\mathbf{x},\alpha,\mathbf{z},\mathbf{u}_\cdot^{\varepsilon,\delta},\mathbf{w}_\cdot^{\varepsilon,\delta}) - v^{\varepsilon,\delta}(\mathbf{x},\alpha,\mathbf{z})| = 0.$$

Proof. In view of Theorem 3.3, it is sufficient to prove that

$$\liminf_{\delta\to 0} J^{\varepsilon,\delta}(\mathbf{x},\alpha,\mathbf{z},\mathbf{u}_\cdot^{\varepsilon,\delta},\mathbf{w}_\cdot^{\varepsilon,\delta}) \geq v^{\varepsilon,0}(\mathbf{x},\alpha,\mathbf{z}). \tag{6.7}$$

We divide the proof into four steps. These steps require the concepts of a martingale problem, the tightness of a class of processes, convergence in distribution, and the Skorohod topology. The reader is referred to [9] for details on these concepts.

Step 1. We need to show that $(\bar{\mathbf{x}}_t, \bar{\alpha}(\varepsilon,t))$ stipulated in (A3)$'$ has a uniquely determined distribution. For this purpose, we let A denote the generator defined as follows:

$$Af(\mathbf{x},\alpha) = \left(\mathbf{u}(\mathbf{x},\alpha) - \sum_{j=0}^{d} \nu_j^d(W(\mathbf{x},\alpha))\mathbf{z}^j\right)\frac{\partial}{\partial\mathbf{x}}f(\mathbf{x},\alpha) + \frac{1}{\varepsilon}\bar{Q}^m(U(\mathbf{x}))f(\mathbf{x},\cdot)(\alpha),$$

where $U(\mathbf{x})$ is given as in (6.5). As in [25, Eq. (5)], we can conclude that for $(\bar{\mathbf{x}}_t, \bar{\alpha}(\varepsilon, t))$,

$$f(\bar{\mathbf{x}}_t, \bar{\alpha}(\varepsilon, t)) - f(\mathbf{x}, \alpha) - \int_0^t Af(\bar{\mathbf{x}}_r, \bar{\alpha}(\varepsilon, r))dr \tag{6.8}$$

is a martingale for any continuously differentiable $f(\mathbf{x}, \alpha)$ which vanishes at infinity. Therefore, in order to prove that the probability distribution of $(\bar{\mathbf{x}}_t, \bar{\alpha}(\varepsilon, t))$ is uniquely determined, it is sufficient to show that the martingale problem (6.8) for A has a unique solution.

Indeed, let $(\bar{\mathbf{x}}_t^0, \bar{\alpha}^0(\varepsilon, t))$ denote a stochastic process such that

$$f(\bar{\mathbf{x}}_t^0, \bar{\alpha}^0(\varepsilon, t)) - f(\mathbf{x}, \alpha) - \int_0^t Af(\bar{\mathbf{x}}_r^0, \bar{\alpha}^0(\varepsilon, r))dr$$

is a martingale. For any $\gamma > 0$ and a bounded locally Lipschitz function $g(\mathbf{x}, \alpha)$, let

$$\phi(\mathbf{x}, \alpha) = E\int_0^\infty e^{-\gamma t} g(\bar{\mathbf{x}}_t^0, \bar{\alpha}^0(\varepsilon, t))dt.$$

Then, we can show (cf. [32]) that $\phi(\mathbf{x}, \alpha)$ is locally Lipschitz and is a viscosity solution to the HJB equation $(\gamma - A)\phi = g$. It can be shown (cf. [32]) that such an HJB equation has a unique viscosity solution. This implies, by [9, Eq. (4.29), p. 187], that the martingale problem (6.8) has a unique solution.

Step 2. Consider $(\mathbf{x}_t, \alpha(\varepsilon, t), \mathbf{z}(\delta, t))$ given by

$$\begin{cases} \dot{\mathbf{x}}_t = \mathbf{u}(\mathbf{x}_t, \alpha(\varepsilon, t)) - \mathbf{z}(\delta, t), \ \mathbf{x}_0 = \mathbf{x}, \\ (\alpha(\varepsilon, t), \mathbf{z}(\delta, t)) \sim \bar{Q}(U(\mathbf{x}_t), W(\mathbf{x}_t, \alpha(\varepsilon, t))), \ \alpha(\varepsilon, 0) = \alpha, \ \mathbf{z}(\delta, 0) = \mathbf{z}. \end{cases} \tag{6.9}$$

Let $D([0, T])$ denote the space of functions that are right-continuous having left-hand limits on $[0, T]$. We now prove that $\{(\mathbf{x}_t, \alpha(\varepsilon, t))\}$, as a sequence of processes indexed by $\delta > 0$, is tight on $D([0, T])$.

Since $\mathbf{u}(\mathbf{x}, \alpha)$ and $\mathbf{z}(\delta, t)$ are bounded, it suffices to show that $\{\alpha(\varepsilon, t)\}$ is tight. Note that for any $0 \le t_1 \le t \le t_2 \le T$ and $\eta > 0$,

$$\begin{aligned} &P(|\alpha(\varepsilon, t) - \alpha(\varepsilon, t_1)| \ge \eta, |\alpha(\varepsilon, t_2) - \alpha(\varepsilon, t)| \ge \eta) \\ &\quad \le P(\alpha(\varepsilon, t) \text{ jumps at least twice on } [t_1, t_2]). \end{aligned}$$

Let $0 = \tau_0 < \tau_1 < \tau_2 < \cdots$ denote a sequence of jump times of $\alpha(\varepsilon, t)$. If $t_1 = 0$, then

$$\begin{aligned} &P(\alpha(\varepsilon, t) \text{ jumps at least twice on } [0, t_2]) \\ &\quad \le P(\tau_1 + \tau_2 \le t_2) \\ &\quad \le 2C\int_0^{t_2}\int_0^t dsdt, \text{ for some constant } C \\ &\quad = Ct_2^2. \end{aligned}$$

If $t_1 > 0$, then by shifting the process $\alpha(\varepsilon, t)$ by t_1 units, we can show that

$$P(\alpha(\varepsilon, t) \text{ jumps at least twice on } [t_1, t_2]) \le C(t_2 - t_1)^2.$$

Let $F(t) = \sqrt{C}t$. Then,

$$P(|\alpha(\varepsilon,t) - \alpha(\varepsilon,t_1)| \geq \eta, |\alpha(\varepsilon,t_2) - \alpha(\varepsilon,t)| \geq \eta) \leq (F(t_2) - F(t_1))^2.$$

By the proof of [3, Thm. 15.6, p. 128], we conclude that $\{\alpha(\varepsilon,t)\}$ is tight.

Step 3. Let $(\mathbf{x}_t, \alpha(\varepsilon,t))$ denote a solution to (6.9) and $(\bar{\mathbf{x}}_t, \bar{\alpha}(\varepsilon,t))$ as defined in (A3)$'$. We now show that $(\mathbf{x}_t, \alpha(\varepsilon,t))$ converges in distribution to $(\bar{\mathbf{x}}_t, \bar{\alpha}(\varepsilon,t))$ as $\delta \to 0$.

Since $\{(\mathbf{x}_t, \alpha(\varepsilon,t)\}$ is tight on $D([0,T])$ for each sequence of δ converging to 0, there exists a subsequence, still denoted by δ, and $(\mathbf{x}_t^0, \alpha^0(\varepsilon,t)) \in D([0,T])$, such that $(\mathbf{x}_t, \alpha(\varepsilon,t))$ converges in distribution to $(\mathbf{x}_t^0, \alpha^0(\varepsilon,t))$. Then, by the Skorohod representation theorem [9, Thm. 1.8, p. 102], there exist a probability space $(\hat{\Omega}, \hat{\mathcal{F}}, \hat{P})$ and processes $(\tilde{\mathbf{x}}_t, \tilde{\alpha}(\varepsilon,t)) \in D([0,T])$ and $(\hat{\mathbf{x}}_t, \hat{\alpha}(\varepsilon,t)) \in D([0,T])$ defined on it, such that

$$\hat{P}((\tilde{\mathbf{x}}_t, \tilde{\alpha}(\varepsilon,t)) \in \cdot) = P((\mathbf{x}_t, \alpha(\varepsilon,t)) \in \cdot),$$
$$\hat{P}((\hat{\mathbf{x}}_t, \hat{\alpha}(\varepsilon,t)) \in \cdot) = P((\mathbf{x}_t^0, \alpha^0(\varepsilon,t)) \in \cdot),$$

and

$$(\tilde{\mathbf{x}}_t, \tilde{\alpha}(\varepsilon,t)) \to (\hat{\mathbf{x}}_t, \hat{\alpha}(\varepsilon,t)) \text{ a.s. as } \delta \to 0 \tag{6.10}$$

in the Skorohod topology. Note that $\mathbf{x}_t$ is continuous almost surely. It follows that $\tilde{\mathbf{x}}_t$ is also continuous almost surely, and so is $\hat{\mathbf{x}}_t$. Therefore, (6.10) yields that $\tilde{\mathbf{x}}_t \to \hat{\mathbf{x}}_t$ in the space $C([0,T])$ of continuous functions with the uniform convergence topology.

Since $\tilde{\mathbf{x}}_t$ has the same distribution as $\mathbf{x}_t$, it holds that $\tilde{\mathbf{x}}_t$ is differentiable almost surely. Define

$$\tilde{\mathbf{z}}(\delta,t) = \mathbf{u}(\tilde{\mathbf{x}}_t, \tilde{\alpha}(\varepsilon,t)) - \dot{\tilde{\mathbf{x}}}_t.$$

Then

$$\hat{P}((\tilde{\mathbf{x}}_t, \tilde{\alpha}(\varepsilon,t), \tilde{\mathbf{z}}(\delta,t)) \in \cdot) = P((\mathbf{x}_t, \alpha(\varepsilon,t), \mathbf{z}(\delta,t)) \in \cdot).$$

Moreover, for any f defined on $\mathcal{Z}$,

$$f(\tilde{\mathbf{z}}(\delta,t)) - f(\tilde{\mathbf{z}}(\delta,0)) - \frac{1}{\delta}\int_0^t \bar{Q}^d(W(\tilde{\mathbf{x}}_r, \tilde{\alpha}(\varepsilon,r)))f(\tilde{\mathbf{z}}(\delta,r))dr$$

is a martingale. Then, we can show as in [25, Lem. 5.1] that

$$\tilde{\mathbf{z}}(\delta,t) \to \boldsymbol{\nu}^d(W(\hat{\mathbf{x}}_t, \hat{\alpha}(\varepsilon,t))) \text{ weakly in } L^2([0,T]) \text{ a.s. as } \delta \to 0,$$

i.e.,

$$\int_a^b \chi_{\{\tilde{\mathbf{z}}(\delta,t)=\mathbf{z}^j\}} g(t)dt \to \int_a^b \nu_j^d(W(\hat{\mathbf{x}}_t, \hat{\alpha}(\varepsilon,t)))g(t)dt \text{ a.s.}$$

for all $g \in L^2([0,T])$.

Note that for any continuously differentiable $f(\mathbf{x},\alpha)$,

$$\begin{aligned} f(\tilde{\mathbf{x}}_t, \tilde{\alpha}(\varepsilon,t)) - f(\mathbf{x},\alpha) - \int_0^t & (\mathbf{u}(\tilde{\mathbf{x}}_r, \tilde{\alpha}(\varepsilon,r)) - \tilde{\mathbf{z}}(\delta,r))\frac{\partial}{\partial \mathbf{x}} f(\tilde{\mathbf{x}}_r, \tilde{\alpha}(\varepsilon,r)) \\ & + \frac{1}{\varepsilon}\bar{Q}^m(U(\tilde{\mathbf{x}}_r))f(\tilde{\mathbf{x}}_r, \cdot)(\tilde{\alpha}(\varepsilon,r))dr \end{aligned} \tag{6.11}$$

is a martingale. Recall that $(\tilde{\mathbf{x}}_t, \tilde{\alpha}(\varepsilon, t)) \to (\bar{\mathbf{x}}_t, \bar{\alpha}(\varepsilon, t))$. By sending $\delta \to 0$ in (6.11), we conclude (cf. [9]) that

$$f(\hat{\mathbf{x}}_t, \hat{\alpha}(\varepsilon, t)) - f(\mathbf{x}, \alpha) - \int_0^t \left(\mathbf{u}(\hat{\mathbf{x}}_r, \hat{\alpha}(\varepsilon, r)) - \sum_{j=0}^{d} \nu_j^d(W(\hat{\mathbf{x}}_r, \hat{\alpha}(\varepsilon, r))) \right) \frac{\partial}{\partial \mathbf{x}} f(\hat{\mathbf{x}}_r, \hat{\alpha}(\varepsilon, r))$$
$$+ \frac{1}{\varepsilon} \bar{Q}^m(U(\hat{\mathbf{x}}_r)) f(\hat{\mathbf{x}}_r, \cdot)(\hat{\alpha}(\varepsilon, r)) dr$$

is also a martingale. Therefore, by Step 1, the probability distribution of $(\hat{\mathbf{x}}_t, \hat{\alpha}(\varepsilon, t))$ is uniquely determined by (6.8). Thus, $(\tilde{\mathbf{x}}_t, \tilde{\alpha}(\varepsilon, t))$ converges to $(\hat{\mathbf{x}}_t, \hat{\alpha}(\varepsilon, t))$ in distribution as $\delta \to 0$.

By the above argument, we may assume (by the Skorohod representation) that

$$\begin{aligned} &(\mathbf{x}_t, \alpha(\varepsilon, t)) \to (\bar{\mathbf{x}}_t, \bar{\alpha}(\varepsilon, t)) \text{ a.s. in the Skorohod topology} \\ \text{and} \quad &\mathbf{z}(\delta, t) \to \boldsymbol{\nu}^d(W(\bar{\mathbf{x}}_t, \bar{\alpha}(\varepsilon, t))) \text{ weakly in } L^2([0, T]) \text{ a.s.} \end{aligned} \tag{6.12}$$

Step 4. We can now prove (6.7).

Let

$$\begin{cases} \mathbf{u}_t^{\varepsilon,0} = \mathbf{u}(\bar{\mathbf{x}}_t, \bar{\alpha}(\varepsilon, t)), \\ W_t^{\varepsilon,0} = (\mathbf{w}_t^0, \ldots, \mathbf{w}_t^d) = W(\bar{\mathbf{x}}_t, \bar{\alpha}(\varepsilon, t)). \end{cases}$$

Then, in view of assumption (A1) and a corresponding result in [25, Thm. 6.1], it remains to show that for any fixed $0 < T < \infty$,

$$\begin{aligned} &\liminf_{\delta \to 0} E \int_0^T e^{-\rho t} G(\mathbf{x}_t, \mathbf{z}(\delta, t), \mathbf{u}_t^{\varepsilon,\delta}, \mathbf{w}_t^{\varepsilon,\delta}) dt \\ &\qquad \geq E \int_0^T e^{-\rho t} \sum_{j=0}^{d} \nu_j^d(W(\bar{\mathbf{x}}_t, \bar{\alpha}(\varepsilon, t))) G(\bar{\mathbf{x}}_t, \mathbf{z}^j, \mathbf{u}_t^{\varepsilon,0}, \mathbf{w}_t^{\varepsilon,0}) dt. \end{aligned} \tag{6.13}$$

In fact,

$$\begin{aligned} &E \int_0^T e^{-\rho t} G(\mathbf{x}_t, \mathbf{z}(\delta, t), \mathbf{u}_t^{\varepsilon,\delta}, \mathbf{w}_t^{\varepsilon,\delta}) dt \\ &\quad = E \int_0^T e^{-\rho t} \chi_{\{\mathbf{z}(\delta,t) = \mathbf{z}^j\}} G(\bar{\mathbf{x}}_t, \mathbf{z}^j, \mathbf{u}_t^{\varepsilon,0}, \mathbf{w}_t^j) dt \\ &\qquad + E \int_0^T e^{-\rho t} \sum_{j=0}^{d} \chi_{\{\mathbf{z}(\delta,t) = \mathbf{z}^j\}} [G(\mathbf{x}_t, \mathbf{z}^j, \mathbf{u}_t^{\varepsilon,\delta}, \mathbf{w}^j(\mathbf{x}_t, \alpha(\varepsilon, t), \mathbf{z}(\delta, t))) \\ &\qquad\qquad - G(\bar{\mathbf{x}}_t, \mathbf{z}^j, \mathbf{u}_t^{\varepsilon,0}, \mathbf{w}_t^j)] dt. \end{aligned} \tag{6.14}$$

Since $(\mathbf{x}_t, \alpha(\varepsilon, t))$ converges to $(\bar{\mathbf{x}}_t, \bar{\alpha}(\varepsilon, t))$ almost surely, the second term in the right-hand side of (6.14) goes to 0. By (6.12), the first term goes to

$$E \int_0^T e^{-\rho t} \nu_j^d(W(\bar{\mathbf{x}}_t, \bar{\alpha}(\varepsilon, t))) G(\bar{\mathbf{x}}_t, \mathbf{z}^j, \mathbf{u}_t^{\varepsilon,0}, \mathbf{w}_t^j) dt = v^{\varepsilon,0}(\mathbf{x}, \alpha, \mathbf{z}).$$

This completes the proof. □

In view of Theorems 6.1 and 6.2, summarized, respectively, in rows 1 and 2 of Table 2, the only remaining hierarchies for which asymptotically optimal controls need

TABLE 2
Asymptotically optimal feedback controls.

Hierarchy: lower vs. upper	Optimal control for upper level	Asymptotically optimal control constructed for lower-level problems
$\mathcal{P}^{\varepsilon,\delta}$ vs. $\mathcal{P}^{0,0}$	$(U,W)\in\mathcal{A}^{0,0}$	$\left(\sum_{i=0}^{m}\chi_{\{\alpha=i\}}\mathbf{u}^i(\mathbf{x}),\sum_{j=0}^{d}\chi_{\{\mathbf{z}=\mathbf{z}^j\}}\mathbf{w}^j(\mathbf{x})\right)\in\mathcal{A}^{\varepsilon,\delta}$
$\mathcal{P}^{\varepsilon,\delta}$ vs. $\mathcal{P}^{\varepsilon,0}$	$(\mathbf{u},W)\in\mathcal{A}^{\varepsilon,0}$	$\left(\mathbf{u}(\mathbf{x},\alpha),\sum_{j=0}^{d}\chi_{\{\mathbf{z}=\mathbf{z}^j\}}\mathbf{w}^j(\mathbf{x},\alpha)\right)\in\mathcal{A}^{\varepsilon,\delta}$
$\mathcal{P}^{\varepsilon,\delta}$ vs. $\mathcal{P}^{0,\delta}$	$(U,\mathbf{w})\in\mathcal{A}^{0,\delta}$	$\left(\sum_{i=0}^{m}\chi_{\{\alpha=i\}}\mathbf{u}^i(\mathbf{x},\mathbf{z}),\mathbf{w}(\mathbf{x},\mathbf{z})\right)\in\mathcal{A}^{\varepsilon,\delta}$
$\mathcal{P}^{\varepsilon,0}$ vs. $\mathcal{P}^{0,0}$	$(U,W)\in\mathcal{A}^{0,0}$	$\left(\sum_{i=0}^{m}\chi_{\{\alpha=i\}}\mathbf{u}^i(\mathbf{x}),W(\mathbf{x})\right)\in\mathcal{A}^{\varepsilon,0}$
$\mathcal{P}^{0,\delta}$ vs. $\mathcal{P}^{0,0}$	$(U,W)\in\mathcal{A}^{0,0}$	$\left(U(\mathbf{x}),\sum_{j=0}^{d}\chi_{\{\mathbf{z}=\mathbf{z}^j\}}\mathbf{w}^j(\mathbf{x})\right)\in\mathcal{A}^{0,\delta}$

to be constructed are $\mathcal{P}^{\varepsilon,\delta}$ vs. $\mathcal{P}^{0,\delta}$, $\mathcal{P}^{\varepsilon,0}$ vs. $\mathcal{P}^{0,0}$, and $\mathcal{P}^{0,\delta}$ vs. $\mathcal{P}^{0,0}$. The results for these are shown in Table 2 in rows 3, 4, and 5, respectively. The proof of row 2 is similar to the proof of Theorem 6.2. The proofs of rows 3 and 4 follow from [25, Thm. 6.2].

7. Concluding remarks.

In this paper, we have presented asymptotic optimality results for hierarchical production and advertising planning in a marketing–production system with random capacity and demand. We describe a procedure to construct a control for the given system, derived from the solution to one of the upper-level problems. The upper-level problems happen to be simpler problems obtained by averaging the given stochastic production capacity process and/or averaging the given stochastic demand process. Therefore, by showing that the associated value functions for the lower-level problems converge to the value functions of the upper-level problems, we are able to construct a control for a lower-level problem from the optimal control of the corresponding upper-level problem. It turns out that the controls so constructed are asymptotically optimal as the rates of transition between the capacity states go to infinity and/or the rates of transition between the demand states go to infinity, respectively.

Several open problems remain. Particularly important to us is the extension of these results to marketing–production systems with state constraints such as those with machines in tandem analyzed in Sethi, Zhang, and Zhou [27]. We would also like to obtain the convergence rates of value functions and error estimates of the constructed controls.

REFERENCES

[1] P. L. ABAD, *Approach to decentralized marketing–production planning*, Internat. J. Systems Sci., 13 (1982), pp. 227–235.

[2] R. N. ANTHONY, *Planning and Control of Systems: A Framework for Analysis*, Harvard University Press, Cambridge, MA, 1965.

[3] P. BILLINGSLEY, *Convergence of Probability Measures*, John Wiley, New York, 1968.

[4] G. BITRAN, AND D. TIRUPATI, *Hierarchical Production Planning*, in Logistics of Production and Inventory, S. C. Graves, A. H. G. Rinnooy Kan, and P. H. Zipkin, eds., Series Handbooks

in Operations Research and Management Sciences, Vol. 4, North–Holland, Amsterdam, 1993, pp. 523–568.

[5] P. N. D. Bukh, *A bibliography of hierarchical production planning techniques, methodology, and applications* (1974–1991), Institute of Management, University of Aarhus, Aarhus, Denmark, 1992, working paper.

[6] Committee on the Next Decade in Operations Research, *Operations research: the next decade*, Oper. Res., 36 (1988), pp. 619–637.

[7] M. G. Crandall, L. C. Evans, and P. L. Lions, *Some Properties of Viscosity Solutions of Hamilton–Jacobi Equations,* Trans. Amer. Math. Soc., 282 (1984), pp. 487–501.

[8] M. H. A. Davis, *Markov Models and Optimization*, Chapman and Hall, New York, 1993.

[9] S. N. Ethier and T. G. Kurtz, *Markov Processes: Characterization and Convergence*, John Wiley, New York, 1986.

[10] G. Feichtinger, R. F. Hartl, and S. P. Sethi, *Dynamic optimal control models in advertising: Recent developments*, Management Sci., 40 (1994), pp. 195–226.

[11] W. H. Fleming, chair, *A Report of the Panel on Future Directions in Control Theory: A Mathematical Perspective*, SIAM Reports on Issues in the Mathematical Sciences, Society for Industrial and Applied Mathematics, Philadelphia, 1988.

[12] W. H. Fleming and R. W. Rishel, *Deterministic and Stochastic Optimal Control*, Springer-Verlag, New York, 1975.

[13] W. H. Fleming and H. M. Soner, *Controlled Markov Processes and Viscosity Solutions*, Springer-Verlag, New York, 1992.

[14] S. B. Gershwin, *Hierarchical flow control: A framework for scheduling and planning discrete events in manufacturing systems*, Proc. IEEE, Special Issue on Dynamics of Discrete Event Systems, 77 (1989), pp. 195–209.

[15] ———, *Manufacturing Systems Engineering*, Prentice–Hall, Englewood Cliffs, NJ, 1994.

[16] K. P. Kistner and M. Switalski, *Hierarchical Production Planning: Necessity, Problems and Methods*, Z. Oper. Res., 33 (1989), pp. 199–212.

[17] J. Lehoczky, S. P. Sethi, H. M. Soner, and M. Taksar, *An asymptotic analysis of hierarchical control of manufacturing systems under uncertainty*, Math. Oper. Res., 16 (1991), pp. 596–608.

[18] C. M. Libosvar, *Hierarchies in production management and control: A survey*, Technical report LIDS-P-1734, Laboratory for Information and Decision Systems, Massachusetts Institute of Technology, Cambridge, MA, 1988.

[19] H. C. Meal, *Putting Production Decisions Where They Belong*, Harvard Business Review, 62 (1984), pp. 102–111.

[20] S. P. Sethi, *Dynamic optimal control models in advertising: A survey,* SIAM Rev., 19 (1977), pp. 685–725.

[21] S. P. Sethi, M. Taksar, and Q. Zhang, *Capacity and production decisions in stochastic manufacturing systems: An asymptotic optimal hierarchical approach*, Production and Operations Management, 1 (1992), pp. 367–392.

[22] S. P. Sethi and Q. Zhang, *Asymptotic optimality in hierarchical control of manufacturing systems under uncertainty: State of the art*, Operations Research Proceedings 1990, W. Bühler, G. Feichtinger, R. Hartl, F. Radermacher, and P. Stähly, eds., Springer-Verlag, Berlin, 1992, pp. 249–263.

[23] ———, *Multilevel hierarchical controls in dynamic stochastic marketing–production systems*, in Proc. 31th IEEE Conference of Decision and Control, Tucson, AZ, 1992, pp. 2090–2095.

[24] ———, *Asymptotic optimality of hierarchical controls in stochastic manufacturing systems: A review*, in Operations Research: Methods, Models and Applications, J. E. Aronson and S. Zionts, eds., Quorum Books, Westport, CT, 1994.

[25] ———, *Asymptotic optimal controls in stochastic manufacturing systems with machine failures dependent on production rates*, Stochastics Stochastics Rep., 48 (1994), pp. 97–121.

[26] ———, *Hierarchical Decision Making in Stochastic Manufacturing Systems*, Birkhäuser, Boston, Cambridge, MA, 1994.

[27] S. P. Sethi, Q. Zhang, and X. Y. Zhou, *Hierarchical controls in stochastic manufacturing systems with machines in tandem*, Stochastics Stochastics Rep., 41 (1992), pp. 89–118.

[28] ———, *Hierarchical controls of stochastic manufacturing systems with convex costs*, J. Optim. Theory Appl., 80 (1994), pp. 299–317.

[29] H. A. Simon, *The Science of the Artificial,* 2nd ed., MIT Press, Cambridge, MA, 1981.

[30] A. G. Sogomonian and C. S. Tang, *A modeling framework for coordinating promotion and production decisions within a firm*, Management Sci., 39 (1993), pp. 191–203.

[31] H. M. Soner, *Singular perturbations in manufacturing*, SIAM J. Control Optim., 31 (1993), pp. 132–146.

[32] H. M. Soner, *Optimal control with state-space constraint II*, SIAM J. Control Optim., 24 (1986), pp. 1110–1122.

[33] H. Stadtler, *Hierarchische Produktionsplanung bei losweiser Fertigung*, Physica-Verlag, Heidelberg, 1988.

[34] M. Switalski, *Hierarchische Produktionsplanung*, Physica-Verlag, Heidelberg, 1989.

[35] K. Yosida, *Functional Analysis*, 6th ed., Springer-Verlag, New York, 1980.

[36] X. Y. Zhou and S. P. Sethi, *A sufficient condition for near optimal stochastic controls and its application to manufacturing systems*, Appl. Math. Optim., 29 (1994), pp. 67–92.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 33, No. 2, pp. 554–567, March 1995

© 1995 Society for Industrial and Applied Mathematics

011

ON BANG–BANG CONSTRAINED SOLUTIONS OF A CONTROL SYSTEM*

RAPHAËL CERF[†] AND CARLO MARICONDA[‡]

Abstract. Given $\phi_1, \phi_2 \in L^1([0,T])$ and a function $x \in W^{2,1}([0,T])$ solving the control problem (P) $x''+a_1(t)x'+a_0(t)x \in [\phi_1(t), \phi_2(t)]$ a.e., $x(0) = x_0$, $x(T) = x_1$, $x'(0) = v_0$, $x'(T) = v_1$, there exists a bang–bang solution y to (P) satisfying $y \leq x$; moreover there exists a finite union of intervals E such that $y'' + a_1y' + a_0y = \phi_1\chi_E + \phi_2\chi_{[0,T]\setminus E}$. The reachable set of bang–bang constrained solutions is convex: an application to the calculus of variations.

Key words. bang–bang, linear control system, range of a vector measure, reachable set, calculus of variations

AMS subject classifications. 34H05, 49B10, 93C15

1. Introduction.

We consider the family of bidimensional linear control systems (P) described by a generic second-order equation subject to a scalar control:

$$x''+a_1(t)x'+a_0(t)x \in \Phi(t) = [\phi_1(t), \phi_2(t)],\ (x(0), x'(0), x(T), x'(T)) = (x_0, v_0, x_1, v_1),$$

where $\phi_1 \leq \phi_2 \in L^1([0,T])$ and $a_1, a_0 \in \mathcal{C}([0,T])$, $x_0, v_0, x_1, v_1 \in \mathbb{R}, x \in W^{2,1}([0,T])$.

The function y is said to be a bang–bang solution to (P) if it solves (P) and, moreover,

$$y'' + a_1(t)y' + a_0(t)y \in \text{extr } \Phi(t) = \{\phi_1(t), \phi_2(t)\} \text{ a.e.} \tag{1.1}$$

Existence of bang–bang solutions has been proved, for instance, by Cesari [4, Thm. 16.3]. The purpose of this paper is to prove that, given an arbitrary solution x to (P), there exists a bang–bang solution y such that

$$\forall t \in [0,T] \qquad y(t) \leq x(t) \tag{1.2}$$

and, in addition, $y'' + a_1y' + a_0y$ steers from ϕ_1 to ϕ_2 only a finite number of times.

Motivation of such a problem was to study the reachable set

$$\mathcal{Y}_T^c = \{(y(T), y'(T)) : y \leq c,\ y''+a_1(t)y'+a_0(t)y \in \text{extr } \Phi(t),\ (y(0), y'(0)) = (x_0, v_0)\},$$

where c is an arbitrary function. A consequence of Theorem 3.1 is that $\mathcal{Y}_T^c$ coincides with

$$\mathcal{X}_T^c = \{(y(T), y'(T)) : y \leq c,\ y'' + a_1(t)y' + a_0(t)y \in \Phi(t),\ (y(0), y'(0)) = (x_0, v_0)\}.$$

Notice that $\mathcal{X}_T^c$ is convex, so the above assumption implies that $\mathcal{Y}_T^c$ is convex too. Another motivation arises from nonconvex problems of the calculus of variations (see [1]).

A possible approach in finding bang–bang solutions is to use the Lyapunov Theorem on the range of a vector measure [4, §16.1].

* Received by the editors December 28, 1992; accepted for publication (in revised form) November 9, 1993.

[†] Ecole Normale Supérieure Paris, Département de Mathématiques et d'Informatique, 45 rue d'Ulm, 75005 Paris, France (Raphael.Cerf@ens.fr).

[‡] Università di Padova, Dipartimento di Matematica pura e applicata, via Belzoni 7, 35131 Padova, Italy (mariconda@pdmat1.math.unipd.it). The research of this author was supported by the Consiglio Nazionale delle Ricerche.

Here, the solution of $x'' + a_1(t)x' + a_0(t)x = \rho(t)$, $x(0) = x'(0) = 0$ is given by

$$x(t) = \int_0^t h(t,s)\rho(s)\,ds,$$

where $h \in \mathcal{C}^1([0,T] \times [0,T])$, and for each $s \in [0,T]$ the function $h_s(.) = h(.,s) \in \mathcal{C}^2([0,T])$ is the solution to the associated homogeneous differential equation satisfying the initial conditions $h_s(s) = 0$, $h_s'(s) = 1$. The Lyapunov Theorem yields the existence of a measurable subset E of $[0,T]$ such that

$$\int_0^T h(T,s)\rho(s)\,ds = \int_0^T h(T,s)(\phi_1(s)\chi_E(s) + \phi_2(s)\chi_{[0,T]\setminus E}(s))\,ds, \tag{1.3}$$

$$\int_0^T \frac{\partial h}{\partial t}(T,s)\rho(s)\,ds = \int_0^T \frac{\partial h}{\partial t}(T,s)(\phi_1(s)\chi_E(s) + \phi_2(s)\chi_{[0,T]\setminus E}(s))\,ds. \tag{1.4}$$

Clearly, by differentiating under the integral sign, the function y defined by

$$y(t) = \int_0^t h(t,s)(\phi_1(s)\chi_E(s) + \phi_2(s)\chi_{[0,T]\setminus E}(s))\,ds \tag{1.5}$$

is a bang–bang solution. However, this approach does not give any information on the behaviour of y with respect to x on $[0,T]$.

Here we prove a new Lyapunov-type theorem concerning the range of a two-dimensional vector measure whose densities are such that their quotient is monotone; in this case, the set E can be chosen in the form $[\alpha,\beta]$. Note that this is not true in general; for instance, there are no $\alpha,\beta \in [0,3\pi]$ satisfying

$$\int_\alpha^\beta \sin t\,dt = \int_0^{3\pi} \sin t\chi_{[0,\pi]\cup[2\pi,3\pi]}(t)\,dt, \qquad \int_\alpha^\beta 1\,dt = \int_0^{3\pi} 1\chi_{[0,\pi]\cup[2\pi,3\pi]}(t)\,dt.$$

In our application, the equalities $h(s,s) = 0$ and $\frac{\partial h}{\partial t}(s,s) = 1$ imply that the monotonicity condition is locally fulfilled; this allows us to build a set E satisfying (1.3)–(1.4) as a finite union of intervals and, in the case where $\phi_1 < \rho < \phi_2$ are continuous, to choose E in such a way that neither 0 nor T belong to its closure.

These facts, together with a decomposition of the kernel $h(t,s)$ into a linear combination of linearly independent functions, are the main tools that we use to show that the bang–bang solution y defined by (1.5) satisfies the inequality $y \leq x$.

As an application, we consider the problem of minimizing the integral functionals

$$I(x,u) = \int_0^T f(t,x(t),u(t))\,dt,$$

where $x : [0,T] \to \mathbb{R}^2$ is such that $x(0)$, $x'(0)$, $x(T)$, $x'(T)$ are fixed and u is a control belonging to $U(t,x) \subset R^2$. The classical approach to obtain existence of a minimum is to impose conditions in order to have the lower semicontinuity of I with respect to u (for instance convexity of $u \mapsto f(t,x,u)$).

Recently, in an effort to provide existence criteria other than convexity in u, some sufficient conditions have been given: for problems of the calculus of variations ($x' = u$ in the above setting) and for maps of the form $f(t,x,x') = g(t,x) + h(t,x')$, existence of solutions has been obtained by requiring that the real map $x \mapsto g(t,x)$ be monotone [5] or, for x in $\mathbb{R}^n$, that the same function be concave [2]. Optimal control problems escaping to convexity conditions have been handled in [6].

It has been proved further in [3] that there exists a dense subset $\mathcal{D}$ of $\mathcal{C}(\mathbb{R})$ such that, for g in it, the problem

$$\text{minimize} \int_0^T g(x(t))\,dt + \int_0^T h(x'(t))\,dt \quad : \quad x(0) = x_0,\, x(T) = x_1$$

admits a solution for every lower semicontinuous h satisfying growth conditions.

Our theorem gives a straightforward generalization of the above result.

2. Assumptions and preliminary results. Let $\phi_1, \phi_2 \in L^1[0,T]$, $\phi_1 \le \phi_2$, and put $\Phi(t) = [\phi_1(t), \phi_2(t)] \subset \mathbb{R}$. We are interested in the solutions of the following control problem.

Problem P.

$$a_1, a_0 \in \mathcal{C}([0,T]), \quad x_0, x_1, v_0, v_1 \in \mathbb{R}, \quad x \in W^{2,1}([0,T]),$$

$$\text{(P)} \qquad x'' + a_1(t)x' + a_0(t)x \in \Phi(t) \text{ a.e.,}$$

$$x(0) = x_0, \quad x'(0) = v_0, \quad x(T) = x_1, \quad x'(T) = v_1.$$

By extr Φ we mean the extreme points of Φ, i.e., extr $\Phi(t) = \{\phi_1(t), \phi_2(t)\}$.

DEFINITION 2.1. *A function $y \in W^{2,1}([0,T])$ is said to be a bang-bang solution to* (P) *if y solves* (P) *and, moreover,*

$$y'' + a_1(t)y' + a_0(t)y \in \text{ extr } \Phi(t) \ a.e.$$

The following representation formula of the solutions to (P) will be used later.

PROPOSITION 2.1. *There exists a function $h \in \mathcal{C}^1([0,T] \times [0,T])$ satisfying Property* S *below such that, for each function $\rho \in L^1([0,T])$, the solution of*

$$(\text{P}_\rho) \qquad x'' + a_1(t)x' + a_0(t)x = \rho(t), \quad x(0) = x'(0) = 0$$

is given by the formula

$$(2.1) \qquad x(t) = \int_0^t h(t,s)\rho(s)\,ds\,.$$

Moreover, for each $s \in [0,T]$, the function $h(.,s)$ is of class $\mathcal{C}^2([0,T])$.

PROPERTY S.

(1) There exist $w_1, w_2 \in \mathcal{C}^2([0,T])$, $z_1, z_2 \in \mathcal{C}^1([0,T])$ such that

$$(2.2) \qquad \forall s,t \in [0,T] \qquad h(t,s) = w_1(t)z_1(s) + w_2(t)z_2(s)$$

$$\text{and} \quad W(w_1, w_2, t) = \det \begin{vmatrix} w_1(t) & w_2(t) \\ w_1'(t) & w_2'(t) \end{vmatrix} \neq 0.$$

For each t_0 in $[0,T]$ there exists $\delta > 0$ such that if we set $I_\delta = [t_0 - \delta, t_0 + \delta] \cap [0,T]$ then:

(2) $\forall s,t \in I_\delta$ $\quad h(t,s) > 0$ if $s < t$, $\quad h(t,s) < 0$ if $t < s$ $\quad$ (whence $h(s,s) = 0$);

(3) $\forall s,t \in I_\delta$ $\quad \frac{\partial h}{\partial t}(t,s) > 0$;

(4) $\forall t \in I_\delta$ $\quad s \mapsto h(t,s)/\frac{\partial h}{\partial t}(t,s)$ is decreasing on I_δ.

Proof of Proposition 2.1. For each $s \in [0,T]$, let $h_s(.) = h(.,s) \in \mathcal{C}^2([0,T])$ be the solution to

$$h_s''(t) + a_1(t)h_s'(t) + a_0(t)h_s(t) = 0, \qquad h_s(s) = 0,\, h_s'(s) = 1.$$

Set $z(t) = \int_0^t h(t,s)\rho(s)\,ds$. Differentiation under the integral sign shows that z is a solution to (P_ρ) whence, by uniqueness, $z = x$.

To prove the second part of the claim, let $w_1, w_2 \in \mathcal{C}^2([0,T])$ be two solutions of the differential equation

$$x'' + a_1(t)x' + a_0(t)x = 0 \tag{2.3}$$

such that their Wronskian

$$W(w_1,w_2,t) = \det\begin{vmatrix} w_1(t) & w_2(t) \\ w_1'(t) & w_2'(t)\end{vmatrix}$$

is nonzero for every t. Such functions exist because the set of the solutions of a second-order linear differential equation is a two-dimensional vector space. Since for each $s \in [0,T]$ the function h_s is a solution to (2.3), there exist z_1, z_2 defined on $[0,T]$ such that

$$\forall s,t \in [0,T] \qquad h_s(t) = w_1(t)z_1(s) + w_2(t)z_2(s). \tag{2.4}$$

Conditions on h_s at s and equation (2.4) yield

$$\begin{cases} h_s(s) = 0 = w_1(s)z_1(s) + w_2(s)z_2(s), \\ h_s'(s) = 1 = w_1'(s)z_1(s) + w_2'(s)z_2(s). \end{cases}$$

Since $W(w_1,w_2,s) \neq 0$ for each s, we find

$$z_1(s) = -\frac{w_2(s)}{W(w_1,w_2,s)}, \quad z_2(s) = \frac{w_1(s)}{W(w_1,w_2,s)},$$

so that $z_1, z_2 \in \mathcal{C}^1([0,T])$; hence $h(t,s) = h_s(t)$ belongs to $\mathcal{C}^1([0,T] \times [0,T])$. By construction

$$\forall s \in [0,T] \qquad h(s,s) = 0 \quad \text{and} \quad \frac{\partial h}{\partial t}(s,s) = 1$$

implying

$$\begin{aligned} \forall s \in [0,T] \qquad \frac{d}{ds}h(s,s) = 0 &\Leftrightarrow \forall s \in [0,T] \qquad \frac{\partial h}{\partial t}(s,s) + \frac{\partial h}{\partial s}(s,s) = 0 \\ &\Leftrightarrow \forall s \in [0,T] \qquad \frac{\partial h}{\partial s}(s,s) = -1. \end{aligned}$$

As a consequence,

$$\forall s \in [0,T] \qquad \frac{\partial}{\partial s}\left(\frac{h}{\frac{\partial h}{\partial t}}\right)(s,s) = -1.$$

By continuity for a fixed t_0 in $[0,T]$, there exists $\delta > 0$ such that

$$\forall s,t \in [t_0 - \delta, t_0 + \delta] \cap [0,T] \qquad \frac{\partial h}{\partial t}(t,s) > 0 \quad \text{and} \quad \frac{\partial}{\partial s}\left(\frac{h}{\frac{\partial h}{\partial t}}\right)(t,s) < 0;$$

for this δ (2), (3), and (4) in Property S are satisfied.

Assume, for instance, $\Phi(t) = [0, \phi(t)]$ and let $\rho \in \mathrm{L}^1([0,T])$ be such that $0 \le \rho \le \phi$. For a solution x to (P_ρ) formula (2.1) yields, in particular,

$$x(T) = \int_0^T h(T,s)\rho(s)\,ds\,, \tag{2.5}$$

$$x'(T) = \int_0^T \frac{\partial h}{\partial t}(T,s)\rho(s)\,ds. \tag{2.6}$$

Let us point out that the classical Lyapunov Theorem on the range of a vector measure [4, §16.1] allows us to find a bang–bang solution. In fact, its application yields the existence of a measurable subset E of $[0,T]$ such that

$$\int_0^T h(T,s)\rho(s)\,ds = \int_0^T h(T,s)\phi(s)\chi_E(s)\,ds\,, \tag{2.7}$$

$$\int_0^T \frac{\partial h}{\partial t}(T,s)\rho(s)\,ds = \int_0^T \frac{\partial h}{\partial t}(T,s)\phi(s)\chi_E(s)\,ds\,, \tag{2.8}$$

so that the function $\bar{x}$ defined by

$$\bar{x}(t) = \int_0^t h(t,s)\phi(s)\chi_E(s)\,ds$$

is, by Proposition 2.1, a bang–bang solution to (P) (with $\phi_1 = 0$, $\phi_2 = \phi$, $x_0 = v_0 = 0$). However, for $0 < t < T$, the Lyapunov Theorem does not give any information on the relative positions of $\bar{x}$ and the original solution x.

The purpose of Proposition 2.2 below is to show that if $s \mapsto \left(h/\frac{\partial h}{\partial t}\right)(t,s)$ is monotone on $[0,T]$ then the measurable subset E can be chosen to be an interval $[\alpha,\beta]$ with $0 \le \alpha \le \beta \le T$. Taking into account Property S (4), this will allow us to define in §3 a bang–bang solution y satisfying $y(t) \le x(t)$ for each t.

In what follows $[a,b]$ is an interval of $\mathbb{R}$; ρ and ϕ are two functions belonging to $L^1([a,b])$ satisfying $0 \le \rho \le \phi$. We say that $r \in \mathbb{R}$ is positive (resp. negative) if $r \ge 0$ (resp. $r \le 0$).

We consider the following hypothesis.

Hypothesis H. The functions f, g belong to $L^\infty([a,b])$ and are positive almost everywhere. Moreover there exists a strictly monotone positive function k such that

$$g(t) = k(t)f(t) \text{ a.e.}$$

We have the following Lyapunov-type result.

PROPOSITION 2.2. *Let f, g satisfy Hypothesis* H. *Then there exist $\alpha, \beta \in \mathbb{R}$ such that, if we put $E = [\alpha,\beta]$, we have*

$$\int_a^b \rho(s)f(s)\,ds = \int_\alpha^\beta \phi(s)f(s)\,ds = \int_a^b \phi(s)f(s)\chi_E(s)\,ds\,, \tag{2.9}$$

$$\int_a^b \rho(s)g(s)\,ds = \int_\alpha^\beta \phi(s)g(s)\,ds = \int_a^b \phi(s)g(s)\chi_E(s)\,ds\,. \tag{2.10}$$

Moreover, α and β are unique if ρ, ϕ, f, g are continuous, and $0 < \rho < \phi$, $f > 0$, $g > 0$.

To prove Proposition 2.2, we need the following fundamental lemma.

LEMMA 2.1. *Assume that f, g satisfy Hypothesis* H *and let $\alpha, \beta \in [a,b]$ be such that*

$$\int_\alpha^b \phi(s)f(s)\,ds = \int_a^b \rho(s)f(s)\,ds, \tag{2.11}$$

$$\int_a^\beta \phi(s)f(s)\,ds = \int_a^b \rho(s)f(s)\,ds. \tag{2.12}$$

Then, if k is increasing, we have

$$\int_\alpha^b \phi(s)g(s)\,ds \geq \int_a^b \rho(s)g(s)\,ds, \tag{2.13}$$

$$\int_a^\beta \phi(s)g(s)\,ds \leq \int_a^b \rho(s)g(s)\,ds. \tag{2.14}$$

If k is decreasing on $[a,b]$, inequalities (2.13) *and* (2.14) *are reversed. Moreover, inequalities* (2.13)–(2.14) *are strict if $0 < \rho < \phi$ and $f > 0$, $g > 0$ a.e.*

Proof of Lemma 2.1. Assume for instance that k is increasing. To prove (2.14) let f_ϕ, f_ρ be the monotone functions defined by

$$f_\phi(t) = \int_a^t \phi(s)f(s)\,ds, \qquad f_\rho(t) = \int_a^t \rho(s)f(s)\,ds.$$

The Lebesgue–Stieltjes formula for integration by parts yields

$$\begin{aligned}\int_a^b \rho(s)g(s)\,ds &= \int_a^b \rho(s)k(s)f(s)\,ds \\ &= \int_a^b k(s)\,df_\rho(s) \\ &= k(b)f_\rho(b) - k(a)f_\rho(a) - \int_a^b f_\rho(s)\,dk(s);\end{aligned}$$

analogously we have

$$\int_a^\beta \phi(s)g(s)\,ds = k(\beta)f_\phi(\beta) - k(a)f_\phi(a) - \int_a^\beta f_\phi(s)\,dk(s).$$

Taking into account that $f_\phi(a) = f_\rho(a) = 0$ and that by (2.12) $f_\rho(b) = f_\phi(\beta)$, we are thus led to show that

$$\int_a^b f_\rho(s)\,dk(s) - \int_a^\beta f_\phi(s)\,dk(s) \leq (k(b) - k(\beta))f_\rho(b). \tag{2.15}$$

By our assumptions we have

$$\forall t \in [a,b] \qquad f_\phi(t) \geq f_\rho(t); \tag{2.16}$$

therefore,

$$\int_a^b f_\rho(s)\,dk(s) - \int_a^\beta f_\phi(s)\,dk(s) \leq \int_\beta^b f_\rho(s)\,dk(s). \tag{2.17}$$

Furthermore, since functions f_ρ and k are increasing we have

$$\int_\beta^b f_\rho(s)\,dk(s) \leq (k(b) - k(\beta))f_\rho(b),$$

which, together with (2.17), gives (2.15).

To prove the final part of the lemma, it is enough to remark that if $f > 0$ and $\rho > 0$ then, by (2.12), $\beta \neq a$; if, moreover, $0 < \rho < \phi$ a.e., then inequality (2.16) is strict for every $t > a$ so that (2.17) is strict too (k being increasing). Similar arguments prove (2.13). ☐

Proof of Proposition 2.2.

(i) *Existence.* (a) Assume first $0 < \rho < \phi$ and $f > 0$, $g > 0$ a.e. Let α_1, α_2, β_1, $\beta_2 \in [a, b]$ be such that

$$\int_{\alpha_1}^{b} \phi(s)f(s)\,ds = \int_a^b \rho(s)f(s)\,ds, \tag{2.18}$$

$$\int_{\alpha_2}^{b} \phi(s)g(s)\,ds = \int_a^b \rho(s)g(s)\,ds, \tag{2.19}$$

$$\int_{a}^{\beta_1} \phi(s)f(s)\,ds = \int_a^b \rho(s)f(s)\,ds, \tag{2.20}$$

$$\int_{a}^{\beta_2} \phi(s)g(s)\,ds = \int_a^b \rho(s)g(s)\,ds. \tag{2.21}$$

Assume for instance that k is decreasing on $[a, b]$. In this situation Lemma 2.1 yields

$$\beta_2 \leq \beta_1, \qquad \alpha_2 \leq \alpha_1. \tag{2.22}$$

The function v defined by

$$v(x) = \int_a^x \phi(s)f(s)\,ds$$

is continuous and increasing with values in $[0, v(b)]$: let v^{-1} denote its inverse function. Set

$$m = \int_a^b \rho(s)f(s)\,ds.$$

Since, by (2.18), $v(b) = v(\alpha_1) + m$, then $v(\alpha) + m \in [0, v(b)]$ if and only if $a \leq \alpha \leq \alpha_1$; this allows us to introduce the continuous function ξ_1 defined by the formula

$$\forall \alpha \in [a, \alpha_1] \qquad \xi_1(\alpha) = v^{-1}(v(\alpha) + m).$$

By definition, we have

$$\forall \alpha \in [a, \alpha_1] \qquad \int_\alpha^{\xi_1(\alpha)} \phi(s)f(s)\,ds = v(\xi_1(\alpha)) - v(\alpha) = m = \int_a^b \rho(s)f(s)\,ds \tag{2.23}$$

so that, by (2.20) and (2.22), we deduce

$$\forall \alpha \in [a, \alpha_1] \qquad \xi_1(\alpha) \geq \beta_1 \geq \beta_2. \tag{2.24}$$

Similarly, (2.21) allows us to define a continuous function $\xi_2 : [\beta_2, b] \to \mathbb{R}$ such that we have

$$\forall \beta \geq \beta_2 \qquad \int_{\xi_2(\beta)}^{\beta} \phi(s)g(s)\,ds = \int_a^b \rho(s)g(s)\,ds, \tag{2.25}$$

from which, together with (2.19) and (2.22), we deduce

$$\forall \beta \geq \beta_2 \qquad \xi_2(\beta) \leq \alpha_2 \leq \alpha_1. \tag{2.26}$$

We deduce from (2.24) and (2.26) that the composed application

$$\xi_2 \circ \xi_1 : [a, \alpha_1] \xrightarrow{\xi_1} [\beta_2, b] \xrightarrow{\xi_2} [a, \alpha_1]$$

is defined and continuous from $[a, \alpha_1]$ into itself and, therefore, admits a fixed point $\bar{\alpha}$. Thus, if we set $\bar{\beta} = \xi_1(\bar{\alpha})$ we have $\bar{\alpha} = \xi_2(\bar{\beta})$. Equalities (2.23) and (2.25) with α, β replaced by $\bar{\alpha}, \bar{\beta}$ yield the conclusion.

(b) Let $\rho_n = \rho + \frac{1}{n}$, $\phi_n = \phi + \frac{2}{n}$, $f_n = f + \frac{1}{n}$ so that $0 < \rho_n < \phi_n$ and $f_n > 0$ a.e., and set $g_n = kf_n$ so that the monotonicity of k implies that $g_n > 0$ a.e. and f_n, g_n satisfy H. By (a) there exist α_n, β_n such that

$$\int_a^b \rho_n(s) f_n(s)\, ds = \int_{\alpha_n}^{\beta_n} \phi_n(s) f_n(s)\, ds, \tag{2.27}$$

$$\int_a^b \rho_n(s) g_n(s)\, ds = \int_{\alpha_n}^{\beta_n} \phi_n(s) g_n(s)\, ds. \tag{2.28}$$

By compactness we may assume $\alpha_n \to \alpha$, $\beta_n \to \beta$. The conclusion follows by passing through the limit in (2.27) and (2.28).

(ii) *Uniqueness.* Assume that $0 < \rho < \phi$, $f > 0$, $g > 0$ are continuous and that, for instance, k is decreasing. By (i(a)) the points α, such that there exists β satisfying (2.11) and (2.12), are the fixed points of the composed map $\xi_2 \circ \xi_1$. By definition the functions ξ_1, ξ_2 are differentiable and we have

$$\forall \alpha \in [a, \alpha_1] \qquad \xi_1'(\alpha) = \frac{v'(\alpha)}{v'(\xi_1(\alpha))} = \frac{\phi(\alpha) f(\alpha)}{\phi(\xi_1(\alpha)) f(\xi_1(\alpha))},$$

$$\forall \beta \in [\beta_2, b] \qquad \xi_2'(\beta) = \frac{\phi(\beta) g(\beta)}{\phi(\xi_2(\beta)) g(\xi_2(\beta))}.$$

To prove the claim we notice that if α satisfies $\xi_2 \circ \xi_1(\alpha) = \alpha$ then

$$(\xi_2 \circ \xi_1)'(\alpha) = \xi_2'(\xi_1(\alpha))\xi_1'(\alpha) = \frac{k(\xi_1(\alpha))}{k(\alpha)}. \tag{2.29}$$

By (2.23) we have $\xi_1(\alpha) > \alpha$ so that the strict monotonicity of k implies $k(\xi_1(\alpha)) < k(\alpha)$ and thus $(\xi_2 \circ \xi_1)'(\alpha) < 1$ whenever $\xi_2 \circ \xi_1(\alpha) = \alpha$. Let $S = \{\alpha \in [a, b] : \xi_2 \circ \xi_1(\alpha) = \alpha\}$. Clearly, S is compact and nonempty by (i); moreover, taking (2.29) into account, for each $\alpha \in S$ there exists η such that

$$\begin{aligned} &\forall t \in]\alpha - \eta, \alpha[\qquad \xi_2 \circ \xi_1(t) > t, \\ &\forall t \in]\alpha, \alpha + \eta[\qquad \xi_2 \circ \xi_1(t) < t. \end{aligned} \tag{2.30}$$

As a consequence, the set S has no accumulation points and is therefore finite.

Let $\alpha_1 = \min S$ and assume $S \neq \{\alpha_1\}$; let $\alpha_2 = \min S \setminus \{\alpha_1\}$. Then by (2.30) there exist $t_1 < t_2 \in [\alpha_1, \alpha_2]$ such that $\xi_2 \circ \xi_1(t_1) < t_1$ and $\xi_2 \circ \xi_1(t_2) > t_2$. Therefore there exists $\bar{t} \in [t_1, t_2]$ such that $\xi_2 \circ \xi_1(\bar{t}) = \bar{t}$, a contradiction. □

3. Main result.

THEOREM 3.1. *Let $x \in W^{2,1}([0,T])$ be a solution to* (P). *Then there exists a bang-bang solution y to* (P) *satisfying*

$$\forall t \in [0, T] \qquad y(t) \le x(t).$$

Moreover, there exists a set E which is a finite union of intervals such that

$$y'' + a_1(t)y' + a_0(t)y = \phi_1(t)\chi_E(t) + \phi_2(t)\chi_{[0,T]\setminus E}(t) \text{ a.e.}$$

COROLLARY 1. *Under the above assumption, there exists a bang–bang solution y satisfying*

$$\forall t \in [0,T] \qquad y(t) \geq x(t).$$

Proof of Corollary 1. Let $-\Phi$ be defined by the equality $(-\Phi)(t) = -\Phi(t)$. Clearly, $\tilde{x} = -x$ solves

$$\tilde{x}'' + a_1(t)\tilde{x}' + a_0(t)\tilde{x} \in -\Phi(t) \text{ a.e.}$$

By Theorem 3.1 there exists a bang–bang solution $\tilde{y}$ satisfying the same boundary conditions as $\tilde{x}$ and satisfying

$$\forall t \in [0,T] \qquad \tilde{y}(t) \leq \tilde{x}(t).$$

Then the function y defined by

$$\forall t \in [0,T] \qquad y(t) = -\tilde{y}(t)$$

is a solution to our problem. □

Proof of Theorem 3.1. Let h be the function defined in Proposition 2.1.

(i) We show that it is not restrictive to assume

$$\Phi(t) = [0, \phi(t)] \quad (\phi \in L^1([0,T]),\ \phi > 0 \text{ a.e.}) \quad \text{and} \quad x_0 = v_0 = 0.$$

In fact, let $\Phi(t) = [\phi_1(t), \phi_2(t)]$ and x satisfy

$$x'' + a_1(t)x' + a_0(t)x \in \Phi(t) \text{ a.e.}$$

Then the function $\tilde{x}$ defined by

$$\tilde{x}(t) = x(t) - x'(0)t - x(0)$$

satisfies $\tilde{x}(0) = \tilde{x}'(0) = 0$ and

$$\tilde{x}'' + a_1(t)\tilde{x}' + a_0(t)\tilde{x} \in [\psi_1(t), \psi_2(t)] \text{ a.e.,}$$

where

$$\psi_1(t) = \phi_1(t) - a_0(t)x'(0)t - a_1(t)x'(0) - a_0(t)x(0),$$
$$\psi_2(t) = \phi_2(t) - a_0(t)x'(0)t - a_1(t)x'(0) - a_0(t)x(0).$$

Moreover, by Proposition 2.1, the function $\bar{x}$ defined by

$$\bar{x}(t) = \tilde{x}(t) - \int_0^t h(t,s)\psi_1(s)\,ds$$

satisfies $\bar{x}(0) = 0$, $\bar{x}'(0) = 0$ and

$$\bar{x}'' + a_1(t)\bar{x}' + a_0(t)\bar{x} \in [0, \psi_2(t) - \psi_1(t)] \text{ a.e.}$$

If we assume that Theorem 3.1 holds for such an interval and initial boundary conditions, there exists a function $\bar{y}$ satisfying

$$\bar{y}(0) = \bar{x}(0), \quad \bar{y}'(0) = \bar{x}'(0), \quad \bar{y}(T) = \bar{x}(T), \quad \bar{y}'(T) = \bar{x}'(T),$$

$$\bar{y}'' + a_1(t)\bar{y}' + a_0(t)\bar{y} \in \{0, \psi_2(t) - \psi_1(t)\} \text{ a.e.,}$$

$$\forall t \in [0,T] \qquad \bar{y}(t) \leq \bar{x}(t).$$

It is now easy to check that the function y defined by

$$y(t) = \bar{y}(t) + \int_0^t h(t,s)\psi_1(s)\,ds + x'(0)t + x(0)$$

is a solution to our problem.

(ii) Assume first that the δ of Property (S) can be chosen in such a way that $I_\delta = [0,T]$. In this case, if we set

$$\rho = x'' + a_1 x' + a_0 x$$

then by Proposition 2.1 we can write

$$x(t) = \int_0^t h(t,s)\rho(s)\,ds, \tag{3.1}$$

where h satisfies Property (S(1)) and, in addition,

$$\forall s,t \in [0,T] \qquad h(t,s) > 0 \text{ if } s < t, \quad h(t,s) < 0 \text{ if } t < s, \tag{3.2}$$

$$\forall s,t \in [0,T] \qquad \frac{\partial h}{\partial t}(t,s) > 0, \tag{3.3}$$

$$\forall t \in [0,T] \qquad s \mapsto h(t,s)/\frac{\partial h}{\partial t}(t,s) \text{ is decreasing on } [0,t]. \tag{3.4}$$

In particular, the functions f and g defined on $[0,T]$ by

$$g(s) = h(T,s), \qquad f(s) = \frac{\partial h}{\partial t}(T,s)$$

verify Hypothesis H with $k(.) = h(T,.)/\frac{\partial h}{\partial t}(T,.)$.

By Proposition 2.1, each bang–bang solution y such that $x(0) = x'(0) = 0$ is given by the formula $y(t) = \int_0^t h(t,s)\nu(s)\,ds$ for some measurable function ν with values in $\{0,\phi(t)\}$.

We are thus led to show that there exists such a ν satisfying

$$\int_0^T h(T,s)\rho(s)\,ds = \int_0^T h(T,s)\nu(s)\,ds, \tag{3.5}$$

$$\int_0^T \frac{\partial h}{\partial t}(T,s)\rho(s)\,ds = \int_0^T \frac{\partial h}{\partial t}(T,s)\nu(s)\,ds, \tag{3.6}$$

and for each t in $[0,T]$,

$$\int_0^t h(t,s)\rho(s)\,ds \geq \int_0^t h(t,s)\nu(s)\,ds. \tag{3.7}$$

(a) Assume $0 < \rho < \phi$ a.e.

By Proposition 2.2 there exist $\alpha, \beta \in [0,T]$ such that

$$\int_0^T h(T,s)\rho(s)\,ds = \int_\alpha^\beta h(T,s)\phi(s)\,ds, \tag{3.8}$$

$$\int_0^T \frac{\partial h}{\partial t}(T,s)\rho(s)\,ds = \int_\alpha^\beta \frac{\partial h}{\partial t}(T,s)\phi(s)\,ds. \tag{3.9}$$

It is clear that if we set

$$\nu(s) = \phi(s)\chi_{[\alpha,\beta]}(s) \tag{3.10}$$

then (3.5) and (3.6) are satisfied. In order to prove (3.7) we first show that under our assumptions on ρ and ϕ we have

$$0 < \alpha < \beta < T. \tag{3.11}$$

Notice first that the equalities $(\alpha, \beta) = (0, T)$ or $\alpha = \beta$ cannot hold otherwise by (3.8), $\rho = \phi$ or $\rho = 0$ a.e., a contradiction. Assume, for instance, $\alpha = 0$ and $\beta < T$, the case $\alpha > 0$ and $\beta = T$ being similar. Under this assumption, equalities (3.8) and (3.9) become

$$\int_0^T h(T,s)\rho(s)\,ds = \int_0^\beta h(T,s)\phi(s)\,ds, \tag{3.12}$$

$$\int_0^T \frac{\partial h}{\partial t}(T,s)\rho(s)\,ds = \int_0^\beta \frac{\partial h}{\partial t}(T,s)\phi(s)\,ds. \tag{3.13}$$

Property (3.4) and the assumption $0 < \rho < \phi$ a.e. allow us to apply Lemma 2.1, from which we deduce

$$\int_0^T h(T,s)\rho(s)\,ds < \int_0^\beta h(T,s)\phi(s)\,ds,$$

contradicting (3.12).

Set $y(t) = \int_0^t h(t,s)\nu(s)\,ds$ so that (3.8) and (3.9) become $y(T) = x(T)$ and $y'(T) = x'(T)$.

The purpose of what follows is to show (3.7), i.e., that $y(t) \leq x(t)$ for each t. We consider the cases $t \in [0, \alpha]$, $t \in [\beta, T]$, $t \in [\alpha, \beta]$ separately.

Inequality (3.7) is trivial if $t \leq \alpha$; in fact we have

$$y(t) = 0 \leq \int_0^t h(t,s)\rho(s)\,ds = x(t),$$

the inequality being strict for $t \in]0, \alpha]$. In particular

$$y(\alpha) < x(\alpha). \tag{3.14}$$

Assume $t \in [\beta, T]$.

Since, taking (3.2) into account, $h(t,s) \leq 0$ whenever $s \geq t$, we have

$$\forall t \geq \beta \qquad \int_t^T h(t,s)\rho(s)\,ds \leq 0 = \int_t^T h(t,s)\nu(s)\,ds \tag{3.15}$$

or, equivalently,

$$\forall t \geq \beta \int_0^T h(t,s)\rho(s)\,ds - \int_0^t h(t,s)\rho(s)\,ds \leq \int_0^T h(t,s)\nu(s)\,ds - \int_0^t h(t,s)\nu(s)\,ds. \tag{3.16}$$

Therefore, in order to prove that $y(t) \leq x(t)$ for $t \in [\beta, T]$ it is enough to show that

$$\forall t \in [\beta, T] \qquad \int_0^T h(t,s)\rho(s)\,ds = \int_0^T h(t,s)\nu(s)\,ds. \tag{3.17}$$

For this purpose, we use Property (S(1)). Equalities (3.8) and (3.9) become

$$\begin{cases} w_1(T)\displaystyle\int_0^T z_1(s)(\rho(s)-\nu(s))\,ds + w_2(T)\int_0^T z_2(s)(\rho(s)-\nu(s))\,ds = 0,\\ w_1'(T)\displaystyle\int_0^T z_1(s)(\rho(s)-\nu(s))\,ds + w_2'(T)\int_0^T z_2(s)(\rho(s)-\nu(s))\,ds = 0.\end{cases}$$

The condition on the Wronskian of w_1, w_2 at T implies

$$\int_0^T z_1(s)(\rho(s)-\nu(s))\,ds = 0, \tag{3.18}$$

$$\int_0^T z_2(s)(\rho(s)-\nu(s))\,ds = 0. \tag{3.19}$$

Multiplying (3.18) by $w_1(t)$, (3.19) by $w_2(t)$, and adding the two equations we obtain

$$\int_0^T (w_1(t)z_1(s)+w_2(t)z_2(s))\rho(s)\,ds = \int_0^T (w_1(t)z_1(s)+w_2(t)z_2(s))\nu(s)\,ds,$$

which, together with Property (S(1)), gives (3.17). Moreover, note that since inequality (3.15) is strict for $t \neq T$,

$$y(\beta) < x(\beta). \tag{3.20}$$

At this stage, we only need to prove that (3.7) holds for $t \in [\alpha, \beta]$.

Assume by contradiction that there exists $t \in [\alpha, \beta]$ such that $x(t) = y(t)$. Let

$$\bar{t} = \sup\{t \in [\alpha,\beta] : x(t) = y(t)\}.$$

Then $\alpha < \bar{t} < \beta$ and, by the very definition of $\bar{t}$, $x(\bar{t}) = y(\bar{t})$ so that

$$y'(\bar{t}) - x'(\bar{t}) = \lim_{t\to\bar{t}^+} \frac{y(t)-x(t)}{t-\bar{t}} \leq 0.$$

It follows that

$$\int_\alpha^{\bar{t}} h(\bar{t},s)\phi(s)\,ds = \int_0^{\bar{t}} h(\bar{t},s)\rho(s)\,ds, \tag{3.21}$$

$$\int_\alpha^{\bar{t}} \frac{\partial h}{\partial t}(\bar{t},s)\phi(s)\,ds \leq \int_0^{\bar{t}} \frac{\partial h}{\partial t}(\bar{t},s)\rho(s)\,ds. \tag{3.22}$$

For each $s \in [0,\bar{t}[$ let $f(s) = h(\bar{t},s)$, $g(s) = \frac{\partial h}{\partial t}(\bar{t},s)$, and $k = g/f$ so that by (3.2)–(3.4) the function k is increasing and $f > 0$, $g > 0$. If we replace (a,b) by $(0,\bar{t})$, Lemma 2.1 together with (3.21) implies that

$$\int_\alpha^{\bar{t}} \frac{\partial h}{\partial t}(\bar{t},s)\phi(s)\,ds > \int_0^{\bar{t}} \frac{\partial h}{\partial t}(\bar{t},s)\rho(s)\,ds,$$

thus contradicting (3.22).

(b) Assume, in general, $0 \leq \rho \leq \phi$ a.e. and let ϕ_n, $\rho_n \in L^1([0,T])$ be such that

$$0 < \rho_n < \phi_n \text{ a.e.} \quad \text{and} \quad \rho_n \to \rho,\ \phi_n \to \phi \text{ in } L^1([0,T])$$

(for instance, $\rho_n = \rho + \frac{1}{n}$, $\phi_n = \phi + \frac{2}{n}$).

Corresponding to each n, there exist α_n, $\beta_n \in [0,T]$ such that, if we set $\nu_n = \phi_n \chi_{[\alpha_n,\beta_n]}$, we have

$$\int_0^T h(T,s)\rho_n(s)\,ds = \int_0^T h(T,s)\nu_n(s)\,ds, \tag{3.23}$$

$$\int_0^T \frac{\partial h}{\partial t}(T,s)\rho_n(s)\,ds = \int_0^T \frac{\partial h}{\partial t}(T,s)\nu_n(s)\,ds, \tag{3.24}$$

and, for each t in $[0,T]$,

$$\int_0^t h(t,s)\rho_n(s)\,ds \geq \int_0^t h(t,s)\nu_n(s)\,ds. \tag{3.25}$$

Because the interval $[0,T]$ is compact, we may assume $\alpha_n \longrightarrow \alpha$, $\beta_n \longrightarrow \beta$ for some $\alpha \leq \beta \in [0,T]$.

Clearly $\nu_n = \phi_n \chi_{[\alpha_n,\beta_n]}$ converges to $\phi\chi_{[\alpha,\beta]}$ in $L^1([0,T])$; therefore, if we pass through the limit in (3.23), (3.24), and (3.25) and we set $\nu = \phi\chi_{[\alpha,\beta]}$, we obtain (3.5), (3.6), and (3.7).

(iii) In the general case, using Property (S) and the compactness of $[a,b]$, there exists a subdivision $a_0 = 0 < a_1 < \cdots < a_l < T = a_{l+1}$ of $[0,T]$ such that, if we put $I_j = [a_j, a_{j+1}]$, we have

- $\forall s,t \in I_j \quad h(t,s) > 0$ if $s < t$, $\quad h(t,s) < 0$ if $t < s$;
- $\forall s,t \in I_j \quad \frac{\partial h}{\partial t}(t,s) > 0$;
- $\forall t \in I_j \quad s \mapsto h(t,s)/\frac{\partial h}{\partial t}(t,s)$ is decreasing on I_j.

By (ii), on each interval I_j there exist α_j, β_j such that the solution y_j to the problem

$$y'' + a_1(t)y' + a_0(t)y = \phi_1(t)\chi_{[a_j,\alpha_j]\cup[\beta_j,b_j]}(t) + \phi_2(t)\chi_{[\alpha_j,\beta_j]}(t) \text{ a.e. on } I_j$$

with the initial conditions

$$y_j(a_j) = x(a_j), \quad y_j'(a_j) = x'(a_j)$$

satisfies the equalities

$$y_j(a_{j+1}) = x(a_{j+1}), \quad y_j'(a_{j+1}) = x'(a_{j+1}),$$

and, moreover, $y_j(t) \leq x(t)$ for each $t \in I_j$.

Clearly the function $y \in W^{2,1}([0,T])$ obtained by glueing together the functions y_j is a solution to our problem. □

Remark 3.1. The proof of Theorem 3.1, part (ii(a)) shows in fact that when $0 < \rho < \phi$, we have $y(t) < x(t)$ on $]0,T[$.

Remark 3.2. With the notations introduced in Proposition 2.1, the proof of Theorem 3.1, part (ii) shows that if $T = \delta$ then, given a solution x to (P), there exists a bang–bang solution $y \leq x$ satisfying

$$\begin{aligned} y'' + a_1(t)y' + a_0(t)y &= \min \Phi(t) \text{ on } [0,\alpha] \cup [\beta,T], \\ y'' + a_1(t)y' + a_0(t)y &= \max \Phi(t) \text{ on } [\alpha,\beta]. \end{aligned}$$

Because the number δ depends only on the function h, it can happen that $\delta = +\infty$.

This is the case when a_1 and a_0 are constant and the equation $\lambda^2 + a_1\lambda + a_0 = 0$ admits two real roots λ_1, λ_2. In fact, under this assumption we have either

$$h(t,s) = \frac{1}{\lambda_2 - \lambda_1}(e^{\lambda_2(t-s)} - e^{\lambda_1(t-s)}) \text{ if } \lambda_1 \neq \lambda_2, \text{ or}$$
$$h(t,s) = (t-s)e^{\lambda(t-s)} \text{ if } \lambda_1 = \lambda_2 = \lambda.$$

4. Applications. Our first application concerns the reachable set of bang–bang constrained solutions. Let c be an arbitrary function defined on $[0,T]$ and consider the reachable sets $\mathcal{X}_T^c$ and $\mathcal{Y}_T^c$ associated with (P) defined by

$$\mathcal{X}_T^c = \{(y(T), y'(T)) : y \leq c,\ y'' + a_1(t)y' + a_0(t)y \in \Phi(t),\ (y(0), y'(0)) = (x_0, v_0)\},$$

$$\mathcal{Y}_T^c = \{(y(T), y'(T)) : y \leq c,\ y'' + a_1(t)y' + a_0(t)y \in \operatorname{extr}\Phi(t),\ (y(0), y'(0)) = (x_0, v_0)\}.$$

Then Theorem 3.1 claims $\mathcal{X}_T^c = \mathcal{Y}_T^c$, whence $\mathcal{Y}_T^c$ is convex.

Finally, we give an application to the calculus of variations.

THEOREM 4.1. *Let a_0, $a_1 \in \mathcal{C}([0,T])$, $\phi_1, \phi_2 \in L^1([0,T])$ verify $\phi_1(t) \leq \phi_2(t)$. Let x_0, v_0, x_1, v_1 be 4 fixed real numbers. Then there exists a dense subset $\mathcal{D}$ of $\mathcal{C}(\mathbb{R})$ for the uniform convergence such that for g in $\mathcal{D}$ the problem*

$$\textit{minimize} \left\{ \int_0^T g(x(t))\,dt + \int_0^T h(\rho(t))\,dt \right\}$$

on the subset of $W^{2,1}([0,T]) \times L^1([0,T])$ of those functions (x, ρ) satisfying

$$(x(0), x'(0), x(T), x'(T)) = (x_0, v_0, x_1, v_1),\ x'' + a_1(t)x' + a_0(t)x = \rho(t) \in [\phi_1(t), \phi_2(t)]$$

admits at least one solution for every lower semicontinuous function h satisfying the growth condition $h(u) \geq c\psi(|u|)$, ψ being lower semicontinuous and convex, $\lim_{r \to +\infty} \psi(r)/r = +\infty$.

Proof. With Theorem 3.1 and the preceding application, the proof is a direct adaptation of the proof given in [3]. □

5. Acknowledgments. We are deeply grateful to Professor Arrigo Cellina for suggesting the problem, for his useful advice, and for carefully reading this paper. We thank Professors Jean-Pierre Aubin and Helena Frankowska who gave us the opportunity to meet together. We are indebted to two anonymous referees for their useful remarks that improved the quality and the presentation of the paper.

REFERENCES

[1] M. AMAR AND A. CELLINA, *On passing to the limit for non convex variational problems*, Asymptotic Anal., 9 (1994), pp. 135–148.

[2] A. CELLINA AND G. COLOMBO, *On a classical problem of the calculus of variations without convexity assumptions*, Ann. Inst. H. Poincaré Anal. Non Linéaire, 7 (1990), pp. 97–106.

[3] A. CELLINA AND C. MARICONDA, *The existence question in the calculus of variations: a density result*, Proc. Amer. Math. Soc., 120 (1994), pp. 1145–1150.

[4] L. CESARI, *Optimization—Theory and Applications*, Springer-Verlag, New York, 1983.

[5] P. MARCELLINI, *Alcune osservazioni sull'esistenza del minimo di integrali del calcolo delle variazioni senza ipotesi di convessitá*, Rend. Mat. Appl. (2), 13 (1980), pp. 271–281.

[6] J. P. RAYMOND, *Existence theorems in optimal control problems without convexity assumptions*, J. Optim. Theory Appl., 67 (1990), pp. 109–132.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 33, No. 2, pp. 568–589, March 1995

© 1995 Society for Industrial and Applied Mathematics
012

DEALING WITH INTEGRAL STATE CONSTRAINTS IN BOUNDARY CONTROL PROBLEMS OF QUASILINEAR ELLIPTIC EQUATIONS*

EDUARDO CASAS† AND LUIS A. FERNÁNDEZ‡

Abstract. This paper deals with state-constrained optimal control problems governed by a quasilinear elliptic equation. These constraints are given in an integral form and depend on the state and its gradient. Equality and inequality constraints are simultaneously considered. Existence of a solution is investigated and some optimality conditions are obtained with the aid of Ekeland's variational principle.

Key words. optimal control, boundary control, quasilinear elliptic operators, optimality conditions, integral state constraints

AMS subject classifications. 49K20, 49J20

1. Introduction.

The main objective of this paper is to derive some necessary conditions satisfied by the optimal controls of a system governed by a quasilinear elliptic equation. Two difficulties arise in obtaining these optimality conditions. The first one is due to the fact that the relation between the control and the state is not differentiable in many cases. When this relation is differentiable, its differential exists in the Gâteaux sense, but we do not know if it is continuously Gâteaux differentiable. The authors have studied these questions in [7], [11], and [18]. The method followed to treat the nondifferentiable cases consists of introducing a family of approximate control problems by perturbing the state equation in such a way that the dependence of the state with respect to the control is now differentiable and then passing to the limit.

The second difficulty is motivated by the presence of equality and inequality integral state constraints. The case of inequality integral state constraints has been considered by several authors, among them Barbu and Precupanu [3], Lasiecka [21], and Mackenroth [24] for control problems governed by linear partial differential equations, and Casas and Fernández [10] for Dirichlet problems associated with quasilinear elliptic equations. Problems with equality and inequality integral constraints on the state have been studied by Chryssoverghi [13] and Tröltzsch [33] for a parabolic semilinear equation. In the framework of the quasilinear equations, when dealing with these state constraints we cannot apply the known multiplier rules like the one of Carathéodory and John (see Mangasarian and Fromovitz [25], McShane [26], Halkin [19], and Pourciau [30]), because it is unknown to us if the functionals involved in the control problem are continuously differentiable or even Frechet differentiable. This lack of Frechet differentiability is motivated by the relationship between the control and the state. Here we are concerned with boundary controls; the reader is referred to [9] for the case of distributed controls.

The plan of the paper is as follows. In the next section we formulate the control problem for a particular quasilinear operator that is essentially the α-Laplacian, $\alpha > 1$.

* Received by the editors July 20, 1992; accepted for publication (in revised form) November 10, 1993. This research was partially supported by Dirección General de Investigación Científica y Técnica, Madrid.

† Departamento de Matemática Aplicada y Ciencias de la Computación, Escuela Tecnica Superior de Ingenieros de Caminos, Canales y Puertos, Universidad de Cantabria, 39071 Santander, Spain.

‡ Departamento de Matemáticas, Estadística y Computación, Facultad de Ciencias, Universidad de Cantabria, 39071 Santander, Spain.

The reason for this choice is twofold: on the one hand, the α-Laplacian is a very important operator because it appears in many physical models (steady laminar flow of non-Newtonian fluids, some reaction-diffusion problems, magnetostatics, glaciology, etc.; see Ames [1], Aris [2], and Pelissier [29]) and it has been extensively studied, being the main example among the quasilinear elliptic equations; on the other hand, this choice avoids technical complications and simplifies the exposition. In §3 we will describe a more general kind of quasilinear elliptic operator, containing the α-Laplacian as a particular case, for which it is possible to extend the results stated in this work; see Remarks 5.2 and 6.1 for more precision. A theorem of existence of solution is proved in §4, and some optimality conditions of Fritz John type are obtained in §§5 and 6 corresponding to the strongly elliptic and degenerate operators, respectively. Finally, in §7 we study control problems submitted only. to inequality state constraints and derive the optimality conditions in a qualified form for almost every problem.

2. Setting of the problem. Let Ω be an open-bounded subset of R^n, $n > 1$, with a Lipschitz boundary Γ; see Nečas [28]. Let us consider the following Neumann problem:

$$(2.1)\qquad \begin{cases} -\operatorname{div}\left\{[k+|\nabla y|]^{\alpha-2}\nabla y\right\}+\lambda y=f & \text{in } \Omega, \\ (k+|\nabla y|)^{\alpha-2}\nabla y\cdot\vec{\nu}=u & \text{on } \Gamma, \end{cases}$$

where k is a nonnegative real number, $\lambda > 0$, $\alpha > 1$, and $\vec{\nu}$ is the unit outward normal vector to Γ.

If we assume on the data (f,u) that $f \in L^p(\Omega)$ with $p > n/\alpha$ and $u \in L^\infty(\Gamma)$, the existence of a unique $y_u \in W^{1,\alpha}(\Omega) \cap L^\infty(\Omega)$ solution of (2.1) can be proved; see Theorem 3.1 below.

Now, if we fix an element $f \in L^p(\Omega)$ with $p > n/\alpha$, the optimal control problem is formulated as follows:

$$(P_\delta)\quad \begin{cases} \text{minimize } \; J(u)=\displaystyle\int_\Omega L(x,y_u(x))dx+\int_\Gamma l(x,y_u(x),u(x))d\sigma(x) \\ \text{such that } \; u\in K \text{ and} \\ G_j(u)=\delta_j, \quad 1\le j\le n_e, \\ G_j(u)\le\delta_j, \quad n_e+1\le j\le n_e+n_i, \end{cases}$$

where $\delta = (\delta_j)_{j=1}^{n_e+n_i} \subset R^{n_e+n_i}$, K is a *weakly closed convex bounded subset of $L^\infty(\Gamma)$,

$$G_j(u)=\int_\Omega g_j(x,y_u(x),\nabla y_u(x))dx,$$

and $L : \Omega \times R \longrightarrow R$, $l : \Gamma \times (R \times R) \longrightarrow R$ and $g_j : \Omega \times (R \times R^n) \longrightarrow R$ are Carathéodory functions of class C^1 with respect to the second and third variables and satisfy

$$(2.2)\qquad l(\cdot,0,0)\in L^1(\Gamma),\; L(\cdot,0),g_j(\cdot,0,0)\in L^1(\Omega),$$

$$\left|\frac{\partial l}{\partial u}(x,y,u)\right| \leq h_1(x)\varphi(|y|+|u|), \tag{2.3}$$

$$\left|\frac{\partial l}{\partial y}(x,y,u)\right| \leq h_2(x)\varphi(|y|+|u|) \tag{2.4}$$

for all $x \in \Gamma$, $y, u \in R$ with $h_1 \in L^1(\Gamma)$ and $h_2 \in L^\tau(\Gamma)$,

$$\left|\frac{\partial L}{\partial y}(x,y)\right| \leq h_3(x)\varphi(|y|), \tag{2.5}$$

$$\left|\frac{\partial g_j}{\partial y}(x,y,\eta)\right| \leq \varphi(|y|)(h_3(x) + |\eta|^{\alpha/\rho}), \tag{2.6}$$

$$\left|\frac{\partial g_j}{\partial \eta_i}(x,y,\eta)\right| \leq \varphi(|y|)(h_4(x) + |\eta|^{\alpha/\theta}), \quad i = 1,\dots,n \tag{2.7}$$

for all $x \in \Omega$, $y \in R$, and $\eta \in R^n$, $1 \leq j \leq n_e + n_i$, with $h_3 \in L^\rho(\Omega)$, $h_4 \in L^\theta(\Omega)$, and $\varphi : R_+ \longrightarrow R_+$ being an increasing function. Initially, the constants τ, ρ, and θ are supposed to be strictly greater than one, later their values will be suitably limited.

The difficulties in deriving the optimality conditions satisfied by a solution of problem (P_δ) depend on the values of k and α in the state equation (2.1). In the next section we will see that the relation between the control and the state is differentiable when $k > 0$ and $\alpha \geq 2$, which will allow us to deduce these conditions in §5. The differentiability in the case $\alpha < 2$ and $k > 0$ is an open problem for us; however, we will obtain the optimality conditions for this case in §5 by introducing a family of approximate problems corresponding to the case $\alpha = 2$ and passing to the limit in the optimality systems of these problems. When $k = 0$ the mapping $u \longrightarrow y_u$ is not necessarily differentiable, as we will prove in the next section. To deal with this case we study the limits of the previous optimality systems when $k \longrightarrow 0$.

3. Sensitivity analysis of quasilinear elliptic equations. It is known that, in order to derive the optimality system satisfied by the solutions of (P_δ), an important question to study is the differentiable character of the relation control-state. In this section, we present a general result about this question concerning the Neumann problem associated with a quasilinear operator in divergence form:

$$\begin{cases} Ay = -\mathrm{div}\,(a(x,\nabla y)) + a_0(x,y) = f & \text{in } \Omega, \\ a(x,\nabla y)\cdot\vec{\nu} = u & \text{on } \Gamma, \end{cases} \tag{3.1}$$

where $a(x,\eta) = (a_1(x,\eta),\dots,a_n(x,\eta))$. On the operator coefficients we will assume the following conditions:

$$\begin{cases} a_j(\cdot,\eta) \text{ is a measurable function in } \Omega, \\ a_j(x,\cdot) \text{ belongs to } C^1(R^n), \qquad j = 1,\dots,n, \end{cases} \tag{3.2}$$

$$\begin{cases} a_0(\cdot,s) \text{ is a measurable function in } \Omega, \\ a_0(x,\cdot) \text{ belongs to } C^1(R), \end{cases} \tag{3.3}$$

$$\sum_{i,j=1}^{n} \frac{\partial a_j}{\partial \eta_i}(x,\eta)\xi_i\xi_j \geq \Lambda_1(k+|\eta|)^{\alpha-2}|\xi|^2, \tag{3.4}$$

$$\sum_{i,j=1}^{n} \left|\frac{\partial a_j}{\partial \eta_i}(x,\eta)\right| \leq \Lambda_2(k+|\eta|)^{\alpha-2}, \tag{3.5}$$

$$\Lambda_3 \leq \frac{\partial a_0}{\partial s}(x,s) \leq h_0(|s|), \tag{3.6}$$

$$a_0(x,0) = a_j(x,0) = 0, \qquad j = 1,\dots,n \tag{3.7}$$

for some $k \geq 0$, some $\alpha \in (1,+\infty)$, some strictly positive constants $\Lambda_1, \Lambda_2, \Lambda_3$, some positive increasing function h_0, all $x \in \Omega$, all $s \in R$, and all $\eta, \xi \in R^n$.

Let us remark that the operator given in (2.1) satisfies these assumptions.

In spite of the general growth condition allowed to the term a_0 (see (3.6)), we can still prove the existence and uniqueness of the solution for (3.1), together with the continuous dependence of this solution with respect to the control u.

THEOREM 3.1. *Under hypotheses* (3.2)–(3.7), *and given* $f \in L^p(\Omega)$ *with* $p > n/\alpha$, *for every* $u \in L^\infty(\Gamma)$ *there exists a unique* $y_u \in W^{1,\alpha}(\Omega) \cap L^\infty(\Omega)$ *solution of the Neumann problem* (3.1). *Moreover, if* $\{u_j\}_{j\in N}$ *is bounded in* $L^\infty(\Gamma)$ *there exists a positive constant* k_1 *independent of* j *such that*

$$\|y_{u_j}\|_{W^{1,\alpha}(\Omega)} + \|y_{u_j}\|_{L^\infty(\Omega)} \leq k_1 \quad \forall j \in N. \tag{3.8}$$

Finally, if $\{u_j\}$ *converges* **weakly towards* u *in* $L^\infty(\Gamma)$, *then* $\{y_{u_j}\}$ *converges to* y_u *strongly in* $W^{1,\alpha}(\Omega)$ *and* **weakly in* $L^\infty(\Omega)$.

Proof. For every natural number m, let us define the function

$$a_0^m(x,s) = \begin{cases} a_0(x,s) & \text{if } |s| \leq m, \\ \Lambda_3(s^{\alpha-1} - m^{\alpha-1}) + a_0(x,m) & \text{if } s > m, \\ \Lambda_3(m^{\alpha-1} - (-s)^{\alpha-1}) + a_0(x,-m) & \text{if } s < -m. \end{cases}$$

Using the hypotheses assumed on a_0, it can be easily verified that a_0^m is a Carathéodory function satisfying

$$|a_0^m(x,s)| \leq \Lambda_3(|s|^{\alpha-1} + m^{\alpha-1}) + h_0(m)m, \tag{3.9}$$

$$(a_0^m(x,s) - a_0^m(x,s'))(s-s') > 0 \quad \text{if } s \neq s', \tag{3.10}$$

$$\begin{cases} a_0^m(x,s)s \geq \Lambda_3 s^2 & \text{if } \alpha \geq 2, \\ a_0^m(x,s)s \geq \Lambda_3(|s|^\alpha - k_2) & \text{if } \alpha < 2 \end{cases} \tag{3.11}$$

for some $k_2 \geq 0$. In fact, when $\alpha < 2$,

$$a_0^m(x,s)s \geq \left\{ \begin{array}{ll} \Lambda_3 s^2 & \text{if } |s| \leq m \\ \Lambda_3|s|^\alpha & \text{if } |s| \geq m \end{array} \right\} \geq \Lambda_3\left(|s|^\alpha - \frac{2-\alpha}{\alpha}\right)$$

for all $x \in \Omega$ and all $s \in R$. Now, let us consider the Neumann problem

$$(3.12)\qquad \begin{cases} -\operatorname{div}(a(x,\nabla y)) + a_0^m(x,y) = f & \text{in } \Omega, \\ a(x,\nabla y)\cdot \vec{\nu} = u & \text{on } \Gamma. \end{cases}$$

By virtue of the corresponding hypotheses on a_j for $j = 1,\dots,n$ and the conditions (3.9)–(3.11), for every $m \in N$, we can apply classical results (see, for instance, Lions [22, Thm. 2.8, p. 183]) to derive the existence of a unique $y_m \in W^{1,\alpha}(\Omega)$ solution of the Neumann problem (3.12). For the boundedness of y_m, let us remark that, if $\alpha > n$, it follows from the Sobolev inclusion $W^{1,\alpha}(\Omega) \subset L^\infty(\Omega)$. If $\alpha \le n$, it can be deduced using Stampacchia's procedure [31], and taking into account the continuous inclusions

$$L^p(\Omega) \subset (W^{1,s}(\Omega))' \quad\text{and}\quad L^\infty(\Gamma) \subset W^{-1/r,r}(\Gamma),$$

where $r = s/(s-1)$ with $r \ge \beta = \alpha/(\alpha-1)$ and $r > n/(\alpha-1)$, because of condition $p > n/\alpha$. Furthermore, there exists a positive constant C independent of m such that

$$\|y_m\|_{W^{1,\alpha}(\Omega)} + \|y_m\|_{L^\infty(\Omega)} \le C.$$

Given $m \ge C$, it follows from its definition that $a_0^m(x,y_m) = a_0(x,y_m)$. Therefore, for every $m \ge C$ the function y_m is a solution of (3.1). On the other hand, the uniqueness of the solution of this problem in $W^{1,\alpha}(\Omega) \cap L^\infty(\Omega)$ is an immediate consequence of the strict monotonicity of A. Finally, the uniform bound (3.8) and the continuous dependence can be obtained by standard arguments. □

Now let us begin the analysis of the differentiability question by showing that the mapping $u \longrightarrow y_u$ is not necessarily differentiable if $k = 0$. Let us take $\alpha = 3$, $\Omega = (-1,+1)^n$, $\lambda = 36$, and $f = 0$ in (2.1). Then it is easy to verify that the associated states to the controls $u_0 = 0$ and $u_t = u_0 + tv$, with $t > 0$ and

$$v(x) = \begin{cases} 1 & \text{if } |x_1| = 1, \\ 0 & \text{otherwise}, \end{cases}$$

are $y_{u_0} = 0$ and

$$y_{u_t}(x) = \begin{cases} \left(|x_1| - 1 + \left(\dfrac{t}{9}\right)^{1/4}\right)^3 & \text{if } |x_1| > 1 - \left(\dfrac{t}{9}\right)^{1/4}, \\ 0 & \text{otherwise}. \end{cases}$$

Thus the nonexistence of the Gâteaux differential of the relation control-state at u_0 in the direction v is a consequence of the following fact:

$$\left\|\frac{y_{u_t} - y_{u_0}}{t}\right\|_{W^{1,1}(\Omega)} = 2^n\left(\frac{1}{36} + \frac{1}{\sqrt{27}t^{1/4}}\right) \longrightarrow +\infty \quad\text{when}\quad t \searrow 0.$$

For $k \ne 0$ we are going to establish that the relationship between the control and the state is Gâteaux differentiable if $\alpha \ge 2$. In this study, some weighted Sobolev spaces appear in a natural form.

DEFINITION 3.1. *Given $y \in W^{1,\alpha}(\Omega)$ and $k \neq 0$, let us define the space $H^y(\Omega)$ as the completion of $C^\infty(\bar{\Omega})$ with respect to the norm*

$$\|z\| = \left(\int_\Omega (k + |\nabla y|)^{\alpha-2} |\nabla z|^2 dx + \int_\Omega |z|^2 dx \right)^{1/2}$$

It may be easily verified that $H^y(\Omega)$ is a Hilbert space with the inner product

$$(z_1, z_2) = \int_\Omega (k + |\nabla y|)^{\alpha-2} \nabla z_1 \nabla z_2 dx + \int_\Omega z_1 z_2 dx.$$

Moreover, we have

$$W^{1,\alpha}(\Omega) \subset H^y(\Omega) \subset H^1(\Omega) \text{ if } \alpha \geq 2,$$

$$H^1(\Omega) \subset H^y(\Omega) \subset W^{1,\alpha}(\Omega) \text{ if } \alpha \leq 2$$

with continuous injections.

More general spaces of this type have been studied by Coffman, Duffin, and Mizel [15], Murthy and Stampacchia [27], and Trudinger [34].

THEOREM 3.2. *Let us suppose $k \neq 0$ and $\alpha \geq 2$. Let $F : L^\infty(\Gamma) \longrightarrow H^1(\Omega)$ be the functional defined by $F(u) = y_u$. Then F is Gâteaux differentiable. Moreover, if $DF(u)v = z$, then z belongs to $H^{y_u}(\Omega)$ and is the unique solution in this space of problem*

$$(3.13) \qquad \begin{cases} -\text{div} \left(\dfrac{\partial a}{\partial \eta}(x, \nabla y_u) \nabla z \right) + \dfrac{\partial a_0}{\partial s}(x, y_u) z = 0 & \text{in } \Omega, \\ \\ \dfrac{\partial a}{\partial \eta}(x, \nabla y_u) \nabla z \cdot \vec{\nu} = v & \text{on } \Gamma. \end{cases}$$

The proof of this theorem is basically the same as the proof of [11, Thm. 3.1] with minor technical changes. Let us remark that the uniqueness of solution of (3.13) in $H^{y_u}(\Omega)$ is a direct consequence of the Lax Milgram theorem applied to the bilinear form

$$B(z_1, z_2) = \int_\Omega \nabla z_1(x)^T \frac{\partial a}{\partial \eta}(x, \nabla y_u(x)) \nabla z_2(x) dx + \int_\Omega \frac{\partial a_0}{\partial s}(x, y_u(x)) z_1(x) z_2(x) dx,$$

which is continuous and coercive on $H^{y_u}(\Omega)$.

4. Existence of a solution for the control problem.

Under an assumption of convexity of l with respect to the last variable, we are going to prove the existence of a solution for the problem (P_δ).

THEOREM 4.1. *Let us assume that L, l and g_j, $1 \leq j \leq n_e + n_i$, are Carathéodory functions in $\Omega \times R$, $\Gamma \times (R \times R)$, and $\Omega \times (R \times R^n)$, respectively, satisfying the following: For every $M > 0$ there exist functions $\psi_M^1 \in L^1(\Gamma)$, $\psi_M^2 \in L^1(\Omega)$, and a constant $C_M > 0$ such that*

$$(4.1) \qquad l(x, y, u) \geq \psi_M^1(x) \quad \forall x \in \Gamma,\ |y|,\ |u| \leq M,$$

$$L(x,y) \geq -\psi_M^2(x) \quad \forall x \in \Omega,\ |y| \leq M, \tag{4.2}$$

$$|g_j(x,y,\eta)| \leq \psi_M^2(x) + C_M|\eta|^\alpha \quad \forall x \in \Omega,\ \eta \in R^n,\ |y| \leq M,\ 1 \leq j \leq n_e + n_i. \tag{4.3}$$

Moreover, let us suppose that $l(x,y,\cdot)$ is convex in u for all $(x,y) \in \Gamma \times R$ and that there exists a feasible control u_0. Then the problem (P_δ) has at least one solution.

Proof. Because of the existence of a feasible control for the problem (P_δ), we can take a minimizing sequence $\{u_k\}_{k=1}^\infty \subset K$ such that every $y_k = y_{u_k}$ satisfies the state constraints. Since K is bounded in $L^\infty(\Gamma)$, with the aid of Theorem 3.1, we can deduce the existence of a subsequence, denoted in the same way, a function $\phi \in L^\alpha(\Omega)$, a constant C, and elements u and $y = y_u$ such that

$$u_k \to u \quad \text{in the }{}^\star\text{weak topology of } L^\infty(\Gamma), \tag{4.4}$$

$$y_k \to y \quad \text{in } W^{1,\alpha}(\Omega), \tag{4.5}$$

$$y_k(x) \to y(x) \ \text{ a.e.}[\sigma]\ x \in \Gamma \quad \text{and} \quad y_k(x) \to y(x) \ \text{ a.e. } x \in \Omega, \tag{4.6}$$

$$\nabla y_k(x) \to \nabla y(x) \ \text{ a.e. } x \in \Omega, \tag{4.7}$$

and

$$\|y_k\|_{L^\infty(\Omega)} \leq C \quad \text{and} \quad |\nabla y_k(x)| \leq \phi(x) \ \text{ a.e. } x \in \Omega \quad \forall k. \tag{4.8}$$

Taking into account that K is *weakly closed in $L^\infty(\Gamma)$, we deduce from (4.4) that $u \in K$. Let us take

$$M = \max\{C, \max_{v \in K} \|v\|_{L^\infty(\Gamma)}\}.$$

Then using the dominated convergence theorem, hypothesis (4.3), and relations (4.6)–(4.8), we get that $G_j(u_k) \to G_j(u)$ for each $j = 1, \ldots, n_e + n_i$. Consequently, u is a feasible control for (P_δ).

On the other hand, thanks to the convexity of l with respect to the last variable, hypotheses (4.1)–(4.2), and convergences (4.4) and (4.6), it follows that (see Cesari [12] or Ekeland and Temam [17])

$$\begin{aligned} J(u) &= \int_\Omega L(x,y(x))dx + \int_\Gamma l(x,y(x),u(x))d\sigma(x) \\ &\leq \liminf_{k\to\infty}\left(\int_\Omega L(x,y_k(x))dx + \int_\Gamma l(x,y_k(x),u_k(x))d\sigma(x)\right) = \inf(P_\delta), \end{aligned}$$

which proves that u is a solution of (P_δ). □

Let us remark that hypothesis (4.3) is satisfied under assumptions (2.2)–(2.7) if $\theta \geq \beta = \alpha/(\alpha-1)$.

5. Optimality conditions: $k > 0$. In this section we are going to deduce the optimality conditions for the problem (P_δ) under the assumption $k > 0$. We will distinguish the cases $\alpha \geq 2$ and $\alpha < 2$.

Hereafter, for the sake of simplicity, let us denote

$$M_k(y)(x) = \frac{(\alpha-2)}{|\nabla y(x)|}[k + |\nabla y(x)|]^{\alpha-3}\nabla y(x)\nabla y(x)^T + [k + |\nabla y(x)|]^{\alpha-2}I, \tag{5.1}$$

where I denotes the identity matrix $n \times n$.

5.1. Case $\alpha \geq 2$. As mentioned in the introduction, since it is unknown to us whether the functions involved in the optimization problem are Frechet differentiable, we cannot apply any known rule of Lagrange multipliers because of the presence of equality constraints; see Pourciau [30] for a discussion of this question and Halkin [19] for a strong result not covering our case.

THEOREM 5.1. *Let us suppose $\alpha \geq 2$, $k > 0$, and that assumptions* (2.2)–(2.7) *are satisfied with $\tau \geq (2n-2)/n$, $\rho \geq (2n)/(n+2)$, and $\theta \geq 2$. Let $\bar{u}$ be a solution of (P_δ). Then there exist real numbers $\bar{\mu}_j$, $j = 0, 1, \dots, n_e + n_i$, and elements $\bar{y} \in W^{1,\alpha}(\Omega) \cap L^\infty(\Omega)$ and $\bar{p} \in H^1(\Omega)$ such that*

$$\bar{\mu}_0 \geq 0 \quad \textit{and} \quad \bar{\mu}_j \geq 0,\ n_e + 1 \leq j \leq n_e + n_i, \tag{5.2}$$

$$\bar{\mu}_0 + \sum_{j=1}^{n_e} |\bar{\mu}_j| + \sum_{j=n_e+1}^{n_e+n_i} \bar{\mu}_j > 0, \tag{5.3}$$

$$\left\{ \begin{aligned} -\text{div}\left\{[k + |\nabla\bar{y}|]^{\alpha-2}\nabla\bar{y}\right\} + \lambda\bar{y} = f \quad & \textit{in } \Omega, \\ (k + |\nabla\bar{y}|)^{\alpha-2}\nabla\bar{y}\cdot\vec{\nu} = \bar{u} \quad & \textit{on } \Gamma, \end{aligned} \right. \tag{5.4}$$

$$\left\{ \begin{aligned} & -\text{div}\left\{M_k(\bar{y})\nabla\bar{p}\right\} + \lambda\bar{p} = \bar{\mu}_0\frac{\partial L}{\partial y}(x,\bar{y}) \\ & + \sum_{j=1}^{n_e+n_i} \bar{\mu}_j \left\{\frac{\partial g_j}{\partial y}(x,\bar{y},\nabla\bar{y}) - \text{div}\left(\frac{\partial g_j}{\partial \eta}(x,\bar{y},\nabla\bar{y})\right)\right\} \quad \textit{in } \Omega, \\ & M_k(\bar{y})\nabla\bar{p}\cdot\vec{\nu} = \bar{\mu}_0\frac{\partial l}{\partial y}(x,\bar{y},\bar{u}) \ \textit{on } \Gamma, \end{aligned} \right. \tag{5.5}$$

with M_k given by (5.1),

$$\bar{\mu}_j\left(\int_\Omega g_j(x,\bar{y}(x),\nabla\bar{y}(x))dx - \delta_j\right) = 0, \quad j > n_e, \tag{5.6}$$

$$\int_\Gamma \left(\bar{p} + \bar{\mu}_0\frac{\partial l}{\partial u}(x,\bar{y},\bar{u})\right)(u - \bar{u})d\sigma(x) \geq 0 \quad \forall u \in K. \tag{5.7}$$

Moreover,

$$\begin{aligned} & \int_\Omega \nabla\bar{p}^T M_k(\bar{y})\nabla\bar{p}dx + \int_\Omega \lambda\bar{p}^2 dx \leq \bar{\mu}_0\left(\int_\Omega \frac{\partial L}{\partial y}(x,\bar{y})\bar{p}dx + \int_\Gamma \frac{\partial l}{\partial y}(x,\bar{y},\bar{u})\bar{p}d\sigma(x)\right) \\ & + \sum_{j=1}^{n_e+n_i} \bar{\mu}_j\left(\int_\Omega \frac{\partial g_j}{\partial y}(x,\bar{y},\nabla\bar{y})\bar{p}dx + \int_\Omega \frac{\partial g_j}{\partial \eta}(x,\bar{y},\nabla\bar{y})\nabla\bar{p}dx\right). \end{aligned} \tag{5.8}$$

Before proving this theorem, let us make some remarks. Since the right-hand side of the partial differential equation in (5.5) involves derivatives of $L^2(\Omega)$-functions and the boundary condition is of Neumann type, the solution of this boundary problem must be intended in a variational sense, i.e., an element $\bar{p} \in H^1(\Omega)$ satisfies (5.5) if and only if for every $\psi \in C^\infty(\bar{\Omega})$ the following identity holds:

$$\int_\Omega \nabla\psi^T M_k(\bar{y})\nabla\bar{p}dx + \int_\Omega \lambda\psi\bar{p}dx = \bar{\mu}_0\left(\int_\Omega \frac{\partial L}{\partial y}(x,\bar{y})\psi dx + \int_\Gamma \frac{\partial l}{\partial y}(x,\bar{y},\bar{u})\psi d\sigma(x)\right)$$

$$+\sum_{j=1}^{n_e+n_i}\bar{\mu}_j\left(\int_\Omega \frac{\partial g_j}{\partial y}(x,\bar{y},\nabla\bar{y})\psi dx + \int_\Omega \frac{\partial g_j}{\partial \eta}(x,\bar{y},\nabla\bar{y})\nabla\psi dx\right).$$

Let us point out that the first integral in the above expression is finite thanks to (5.8).

It is well known that the difficulty with the Neumann problems is the interpretation of the boundary condition; see, for instance, Lions and Magenes [23, Vol. 1] for linear problems and Casas and Fernández [6] for quasilinear elliptic problems.

To illustrate this situation and the previous theorem, let us give an example:

$$(P_\delta)\begin{cases} \text{minimize } J(u) = \dfrac{1}{2}\displaystyle\int_\Omega (y_u(x) - z_d(x))^2 dx \\ \text{such that } 0 \le u(x) \le 1 \text{ a.e. } [\sigma]\ x \in \Gamma \text{ and} \\ G_j(u) = \delta_j, \quad 1 \le j \le n, \end{cases}$$

with

$$G_j(u) = \int_\Omega \frac{\partial y_u}{\partial x_j}(x)dx, \quad j = 1,\dots,n$$

and z_d a given element of $L^2(\Omega)$.

If we assume that problem (P_δ) has a feasible control, the existence of at least one solution $\bar{u}$ follows from Theorem 4.1. Now, applying Theorem 5.1, we deduce the existence of some real numbers $\{\bar{\mu}_j\}_{j=0}^n$ and elements $\bar{y} \in W^{1,\alpha}(\Omega) \cap L^\infty(\Omega)$ and $\bar{p} \in H^1(\Omega)$ such that

$$\bar{\mu}_0 \ge 0 \quad\text{and}\quad \bar{\mu}_0 + \sum_{j=1}^n |\bar{\mu}_j| > 0, \tag{5.9}$$

$$\begin{cases} -\operatorname{div}\{[k+|\nabla\bar{y}|]^{\alpha-2}\nabla\bar{y}\} + \lambda\bar{y} = f & \text{in } \Omega, \\ (k+|\nabla\bar{y}|)^{\alpha-2}\nabla\bar{y}\cdot\vec{\nu} = \bar{u} & \text{on } \Gamma, \end{cases} \tag{5.10}$$

$$\begin{cases} -\operatorname{div}\{M_k(\bar{y})\nabla\bar{p}\} + \lambda\bar{p} = \bar{\mu}_0(\bar{y} - z_d) - \displaystyle\sum_{j=1}^n \bar{\mu}_j\frac{\partial}{\partial x_j} & \text{in } \Omega, \\ M_k(\bar{y})\nabla\bar{p}\cdot\vec{\nu} = 0 \quad \text{on } \Gamma, \end{cases} \tag{5.11}$$

$$\int_\Gamma \bar{p}(u - \bar{u})d\sigma(x) \ge 0 \quad \forall\, 0 \le u(x) \le 1 \text{ a.e. } [\sigma]\ x \in \Gamma. \tag{5.12}$$

According to the preceding comments, the solution of (5.11) must be intended in a variational sense as follows: for every $\psi \in C^\infty(\bar\Omega)$

$$\int_\Omega \nabla\psi^T M_k(\bar y)\nabla\bar p dx + \int_\Omega \lambda\psi\bar p dx = \bar\mu_0 \int_\Omega (\bar y - z_d)\psi dx + \sum_{j=1}^n \bar\mu_j \int_\Omega \frac{\partial\psi}{\partial x_j} dx.$$

As it is classical in control theory, from (5.12) we can deduce a behaviour of bang-bang type of $\bar u$:

$$\bar u(x) = \begin{cases} 1 & \text{if } \bar p(x) < 0, \\ 0 & \text{if } \bar p(x) > 0. \end{cases}$$

Proof of Theorem 5.1. We will follow a penalization method, where the penalty objective function is a slight differentiable modification of the one used by Clarke in [14]. For every $\epsilon > 0$, let us consider the problem

$$(P_{\delta,\epsilon}) \begin{cases} \text{minimize } J_\epsilon(u) \\ \text{such that } u \in K, \end{cases}$$

where

$$J_\epsilon(u) = \left\{ [(J(u) - J(\bar u) + \epsilon)^+]^2 + \sum_{j=1}^{n_e} |G_j(u) - \delta_j|^2 + \sum_{j=n_e+1}^{n_e+n_i} [(G_j(u) - \delta_j)^+]^2 \right\}^{1/2}$$

and $c^+ = \max\{c, 0\}$. It is obvious that (K, d) is a complete metric space, with $d(u,v) = \|u - v\|_{L^\infty(\Gamma)}$, $J_\epsilon : K \longrightarrow R$ is a continuous function, and

$$J_\epsilon(\bar u) = \epsilon \leq \epsilon + \inf_{u\in K} J_\epsilon(u).$$

Then we can apply Ekeland's variational principle [16] to deduce the existence of an element $u_\epsilon \in K$ such that

$$d(u_\epsilon, \bar u) = \|u_\epsilon - \bar u\|_{L^\infty(\Gamma)} \leq \sqrt\epsilon \tag{5.13}$$

and u_ϵ is a solution of problem

$$(Q_{\delta,\epsilon}) \begin{cases} \text{minimize } J_\epsilon(u) + \sqrt\epsilon\|u_\epsilon - u\|_{L^\infty(\Gamma)} \\ \text{such that } u \in K. \end{cases}$$

Since J_ϵ is Gâteaux differentiable in K (note that $J_\epsilon(u) > 0 \ \ \forall u \in K$ because $\bar u$ is a solution of (P_δ)) and $\phi(u) = \|u_\epsilon - u\|_{L^\infty(\Gamma)}$ is a convex function, we obtain (for example, see [4, Lem. 2])

$$J'_\epsilon(u_\epsilon)\cdot(u - u_\epsilon) + \sqrt\epsilon\|u - u_\epsilon\|_{L^\infty(\Gamma)} \geq 0 \quad \forall u \in K. \tag{5.14}$$

Now it is easy to verify that

$$J'_\epsilon(u_\epsilon)\cdot(u - u_\epsilon) = \left(\mu_{0\epsilon}J'(u_\epsilon) + \sum_{j=1}^{n_e+n_i} \mu_{j\epsilon}G'_j(u_\epsilon)\right)\cdot(u - u_\epsilon), \tag{5.15}$$

where

$$\begin{cases} \mu_{0\epsilon} = J_\epsilon(u_\epsilon)^{-1}(J(u_\epsilon) - J(\bar{u}) + \epsilon)^+, \\ \mu_{j\epsilon} = J_\epsilon(u_\epsilon)^{-1}(G_j(u_\epsilon) - \delta_j),\ 1 \le j \le n_e, \\ \mu_{j\epsilon} = J_\epsilon(u_\epsilon)^{-1}(G_j(u_\epsilon) - \delta_j)^+,\ j > n_e. \end{cases}$$

So we have that $\mu_{0\epsilon} \ge 0$ and $\mu_{j\epsilon} \ge 0$ for $j > n_e$; moreover,

$$\sum_{j=0}^{n_e+n_i} \mu_{j\epsilon}^2 = 1. \tag{5.16}$$

On the other hand, let p_ϵ be the unique variational solution in $H^{y_\epsilon}(\Omega)$ of the problem

$$\begin{cases} -\text{div}\,\{M_k(y_\epsilon)\nabla p_\epsilon\} + \lambda p_\epsilon = \mu_{0\epsilon}\dfrac{\partial L}{\partial y}(x, y_\epsilon) \\ \displaystyle + \sum_{j=1}^{n_e+n_i} \mu_{j\epsilon}\left\{\frac{\partial g_j}{\partial y}(x, y_\epsilon, \nabla y_\epsilon) - \text{div}\left(\frac{\partial g_j}{\partial \eta}(x, y_\epsilon, \nabla y_\epsilon)\right)\right\} \quad \text{in } \Omega, \\ M_k(y_\epsilon)\nabla p_\epsilon \cdot \vec{\nu} = \mu_{0\epsilon}\dfrac{\partial l}{\partial y}(x, y_\epsilon, u_\epsilon) \quad \text{on } \Gamma, \end{cases} \tag{5.17}$$

where $y_\epsilon \in W^{1,\alpha}(\Omega) \cap L^\infty(\Omega)$ is the state associated with u_ϵ. Let us note that

$$z \longrightarrow \sum_{j=1}^{n_e+n_i} \mu_{j\epsilon}\left(\int_\Omega \frac{\partial g_j}{\partial y}(x, y_\epsilon, \nabla y_\epsilon)z + \frac{\partial g_j}{\partial \eta}(x, y_\epsilon, \nabla y_\epsilon)\nabla z dx\right)$$

$$+ \mu_{0\epsilon}\left(\int_\Omega \frac{\partial L}{\partial y}(x, y_\epsilon)z dx + \int_\Gamma \frac{\partial l}{\partial y}(x, y_\epsilon, u_\epsilon)z d\sigma(x)\right)$$

defines an element belonging to $(H^1(\Omega))' \subset (H^{y_\epsilon}(\Omega))'$ thanks to the choice of τ, ρ and θ, the Sobolev imbedding theorems, and assumptions (2.4)–(2.7).

Let u be an arbitrary element of K and let us denote $z_\epsilon = DF(u_\epsilon) \cdot (u - u_\epsilon)$, where $F(v) = y_v$; see Theorem 3.2. Then, from conditions (2.3)–(2.7), the identity (5.15), and equations (3.13) and (5.17) we obtain

$$\begin{aligned} J'_\epsilon(u_\epsilon) \cdot (u - u_\epsilon) &= \mu_{0\epsilon}\int_\Omega \frac{\partial L}{\partial y}(x, y_\epsilon)z_\epsilon dx \\ &\quad + \mu_{0\epsilon}\int_\Gamma \left\{\frac{\partial l}{\partial y}(x, y_\epsilon, u_\epsilon)z_\epsilon + \frac{\partial l}{\partial u}(x, y_\epsilon, u_\epsilon)(u - u_\epsilon)\right\} d\sigma(x) \\ &\quad + \sum_{j=1}^{n_e+n_i} \mu_{j\epsilon}\int_\Omega \left\{\frac{\partial g_j}{\partial y}(x, y_\epsilon, \nabla y_\epsilon)z_\epsilon + \frac{\partial g_j}{\partial \eta}(x, y_\epsilon, \nabla y_\epsilon)\nabla z_\epsilon\right\} dx \\ &= \int_\Omega \{\nabla z_\epsilon^T M_k(y_\epsilon)\nabla p_\epsilon + \lambda z_\epsilon p_\epsilon\}\, dx + \int_\Gamma \mu_{0\epsilon}\frac{\partial l}{\partial u}(x, y_\epsilon, u_\epsilon)(u - u_\epsilon)d\sigma(x) \\ &= \int_\Gamma \left(p_\epsilon + \mu_{0\epsilon}\frac{\partial l}{\partial u}(x, y_\epsilon, u_\epsilon)\right)(u - u_\epsilon)d\sigma(x). \end{aligned}$$

From (5.13) it follows that $u_\epsilon \to \bar{u}$ in $L^\infty(\Gamma)$; therefore, $y_\epsilon = y_{u_\epsilon} \to \bar{y} = y_{\bar{u}}$ in $W^{1,\alpha}(\Omega)$. Now, thanks to Theorem 3.1 and (5.16), we deduce the existence of a constant $C > 0$, a function $\phi \in L^\alpha(\Omega)$, and a subsequence, still denoted in the same way, verifying

$$\mu_{j\epsilon} \to \bar{\mu}_j, \quad j = 0, 1, \ldots, n_e + n_i, \tag{5.18}$$

$$y_\epsilon(x) \to \bar{y}(x) \ \text{ a.e.}[\sigma]\ x \in \Gamma \quad \text{and} \quad y_\epsilon(x) \to \bar{y}(x) \ \text{ a.e. } x \in \Omega, \tag{5.19}$$

$$\nabla y_\epsilon(x) \to \nabla \bar{y}(x) \ \text{ a.e. } x \in \Omega, \tag{5.20}$$

and

$$\|y_\epsilon\|_{L^\infty(\Omega)} \le C \quad \text{and} \quad |\nabla y_\epsilon(x)| \le \phi(x) \ \text{ a.e. } x \in \Omega \quad \forall \epsilon. \tag{5.21}$$

With the aid of the previous relations, and taking into account the definitions of $\mu_{j\epsilon}$ and identity (5.16), we can pass to the limit and get (5.2)–(5.4) and (5.6). On the other hand, from hypotheses (2.3)–(2.7) and (5.18)–(5.21), we deduce the convergence of the right-hand terms of (5.17) toward the corresponding terms of (5.5) in $(H^1(\Omega))'$. Utilizing the fact that p_ϵ is the solution of (5.17) and integrating by parts we get

$$C_1 \|p_\epsilon\|^2_{H^1(\Omega)} \le \int_\Omega \nabla p_\epsilon^T M_k(y_\epsilon) \nabla p_\epsilon dx + \lambda \int_\Omega p_\epsilon^2 dx \le C_2 \|p_\epsilon\|_{H^1(\Omega)}, \tag{5.22}$$

where C_1 and C_2 are positive constants. This implies that $\{p_\epsilon\}_\epsilon$ is bounded in $H^1(\Omega)$. Then, we can take a new subsequence, denoted again by $\{p_\epsilon\}_\epsilon$, such that

$$p_\epsilon \to \bar{p} \quad \text{weakly in } H^1(\Omega) \tag{5.23}$$

for a certain element $\bar{p} \in H^1(\Omega)$. By virtue of the expression obtained for the derivate $J'_\epsilon(u_\epsilon) \cdot (u - u_\epsilon)$, we can pass to the limit in (5.14) and derive (5.7).

To finish the proof it is enough to establish that $\bar{p}$ satisfies (5.5) and inequality (5.8). First, we point out that $M_k(y_\epsilon)(x)$ and $M_k(\bar{y})(x)$ are symmetric and positive definite matrices for every $x \in \Omega$. Thus, applying the Cholesky method, we deduce the existence of lower triangular matrices U_ϵ and U with strictly positive elements such that

$$M_k(y_\epsilon)(x) = U_\epsilon(x) U_\epsilon^T(x) \quad \text{and} \quad M_k(\bar{y})(x) = U(x) U^T(x).$$

From (5.22) and the boundedness of $\{p_\epsilon\}_\epsilon$ in $H^1(\Omega)$ it follows that $\{U_\epsilon^T \nabla p_\epsilon\}_\epsilon$ is bounded in $\left(L^2(\Omega)\right)^n$. Furthermore, using (5.21), the expressions of the elements of U_ϵ and U as function of those of $M_k(y_\epsilon)$ and $M_k(\bar{y})$, respectively, and taking a new subsequence, if necessary, we obtain

$$\begin{aligned} \|U_\epsilon(x)\|_s &\le \sqrt{\alpha - 1}(k + |\nabla y_\epsilon(x)|)^{(\alpha-2)/2} \\ &\le \sqrt{\alpha - 1}(k + |\phi(x)|)^{(\alpha-2)/2} \ \text{ a.e. } x \in \Omega \quad \forall \epsilon, \end{aligned} \tag{5.24}$$

and

$$U_\epsilon(x) \longrightarrow U(x) \quad \text{a.e. } x \in \Omega \quad \text{when } \epsilon \to 0, \tag{5.25}$$

where $\|.\|_s$ denotes the spectral matrix norm defined by

$$\|U\|_s = \left(\max_i\{|\gamma_i| : \gamma_i \text{ is an eigenvalue of } U^TU\}\right)^{1/2};$$

see, for instance, Isaacson and Keller [20]. Together with the dominated convergence theorem, this implies that

$$U_\epsilon \to U \quad \text{in } (L^2(\Omega))^{n\times n}.$$

Since $\nabla p_\epsilon \to \nabla \bar{p}$ weakly in $(L^2(\Omega))^n$ as $\epsilon \to 0$, we have that

$$U_\epsilon^T \nabla p_\epsilon \longrightarrow U^T\nabla\bar{p} \quad \text{weakly in } (L^2(\Omega))^n. \tag{5.26}$$

Combining (5.26) with the convergence

$$U_\epsilon^T\nabla\psi \longrightarrow U^T\nabla\psi \quad \text{in } (L^2(\Omega))^n$$

for $\psi \in C^\infty(\bar{\Omega})$ arbitrary, it follows that

$$\begin{aligned}\int_\Omega \nabla\psi^T M_k(y_\epsilon)\nabla p_\epsilon dx &= \int_\Omega [U_\epsilon^T\nabla\psi]^T[U_\epsilon^T\nabla p_\epsilon]dx\\ \longrightarrow \int_\Omega [U^T\nabla\psi]^T[U^T\nabla\bar{p}]dx &= \int_\Omega \nabla\psi^T M_k(\bar{y})\nabla\bar{p}dx,\end{aligned}$$

and therefore $\bar{p}$ is a solution of (5.5). Finally, as a consequence of the previous relations and convergences, hypotheses (2.4)–(2.7), the values of τ, ρ, and θ, and the Sobolev imbedding theorem for $H^1(\Omega)$ we conclude that

$$\begin{aligned}&\int_\Omega \nabla\,\bar{p}^T M_k(\bar{y})\nabla\bar{p}dx + \int_\Omega \lambda\bar{p}^2dx\\
&= \|U^T\nabla\bar{p}\|^2_{L^2(\Omega)} + \lambda\|\bar{p}\|^2_{L^2(\Omega)}\\
&\leq \liminf_{\epsilon\to 0}\left(\|U_\epsilon^T\nabla p_\epsilon\|^2_{L^2(\Omega)} + \lambda\|p_\epsilon\|^2_{L^2(\Omega)}\right)\\
&= \liminf_{\epsilon\to 0}\left(\int_\Omega \nabla p_\epsilon^T M_k(y_\epsilon)\nabla p_\epsilon dx + \lambda\int_\Omega p_\epsilon^2dx\right)\\
&= \liminf_{\epsilon\to 0}\left\{\mu_{0\epsilon}\left(\int_\Omega \frac{\partial L}{\partial y}(x,y_\epsilon)p_\epsilon dx + \int_\Gamma \frac{\partial l}{\partial y}(x,y_\epsilon,u_\epsilon)p_\epsilon d\sigma(x)\right)\right.\\
&\qquad\left. + \sum_{j=1}^{n_e+n_i}\mu_{j\epsilon}\left(\int_\Omega \frac{\partial g_j}{\partial y}(x,y_\epsilon,\nabla y_\epsilon)p_\epsilon dx + \int_\Omega \frac{\partial g_j}{\partial \eta}(x,y_\epsilon,\nabla y_\epsilon)\nabla p_\epsilon dx\right)\right\}\\
&= \bar{\mu}_0\left(\int_\Omega \frac{\partial L}{\partial y}(x,\bar{y})\bar{p}dx + \int_\Gamma \frac{\partial l}{\partial y}(x,\bar{y},\bar{u})\bar{p}d\sigma(x)\right)\\
&\qquad + \sum_{j=1}^{n_e+n_i}\bar{\mu}_j\left(\int_\Omega \frac{\partial g_j}{\partial y}(x,\bar{y},\nabla\bar{y})\bar{p}dx + \int_\Omega \frac{\partial g_j}{\partial\eta}(x,\bar{y},\nabla\bar{y})\nabla\bar{p}dx\right),\end{aligned}$$

or equivalently, $\bar{p}$ satisfies (5.8). □

Remark 5.1. In the absence of equality constraints ($n_e = 0$), we can derive the optimality conditions by a direct application of the abstract Lagrange multiplier rule given in [5, Thm. 5.2], where only Gâteaux differentiability of the functions defining the problem is required. Following this approach we derive the existence of a unique adjoint state $\bar{p} \in H^{\bar{y}}(\Omega)$.

5.2. Case $\alpha < 2$. The goal of this section is to prove the following theorem.

THEOREM 5.2. *Let $\bar{u}$ be a solution of (P_δ), $\alpha < 2$ and $k > 0$. Furthermore, let us assume that $l(x,y,\cdot) : R \longrightarrow R$ is a convex function for all $(x,y) \in \Omega \times R$, $\tau \geq (n\alpha - \alpha)/(n\alpha - n)$, $\rho \geq (\alpha n)/(\alpha n + \alpha - n)$, and $\theta \geq \alpha/(\alpha - 1)$. Then there exist real numbers $\bar{\mu}_j$, $j = 0, 1, \dots, n_e + n_i$, and elements $\bar{y} \in W^{1,\alpha}(\Omega) \cap L^\infty(\Omega)$ and $\bar{p} \in W^{1,\alpha}(\Omega)$ verifying* (5.2)–(5.8).

As we have already mentioned, in this case we do not know if the relation control-state is differentiable. To overcome this difficulty the state equation is perturbed in such a way that this relation again becomes differentiable. So, given $\epsilon > 0$, let us consider the following operator

$$A_\epsilon y = -\text{div}\left\{(\epsilon + [k + |\nabla y|]^{\alpha-2})\nabla y\right\} + \lambda y,$$

which satisfies hypotheses (3.2)–(3.7) with an exponent 2. Therefore the Neumann problem

$$\text{(5.27)} \qquad \begin{cases} A_\epsilon y = f & \text{in } \Omega, \\ \left(\epsilon + [k + |\nabla y|]^{\alpha-2}\right) \nabla y \cdot \vec{\nu} = u & \text{on } \Gamma \end{cases}$$

has a unique solution $y_\epsilon(u) \in H^1(\Omega) \cap L^\infty(\Omega)$ for every $u \in L^\infty(\Gamma)$. Moreover, since we are in the conditions of Theorem 3.2, the mapping $u \longrightarrow y_\epsilon(u)$ is Gâteaux differentiable.

Now associated with the previous Neumann problem we introduce the following family of control problems that approximate (P_δ) in a sense to be specified later:

$$(Q_{\delta,\epsilon}) \begin{cases} \text{minimize } \tilde{J}_\epsilon(u) \\ \text{such that } u \in K \text{ and} \\ G_{j\epsilon}(u) = \delta_{j\epsilon}, \quad 1 \leq j \leq n_e, \\ G_{j\epsilon}(u) \leq \delta_{j\epsilon}, \quad n_e + 1 \leq j \leq n_e + n_i, \end{cases}$$

where

$$\tilde{J}_\epsilon(u) = \int_\Omega L(x, y_\epsilon(u)(x))dx + \int_\Gamma l(x, y_\epsilon(u)(x), u(x))d\sigma(x) + \frac{1}{2}\int_\Gamma |u(x) - \bar{u}(x)|^2 d\sigma(x),$$

$$G_{j\epsilon}(u) = \int_\Omega g_j(x, y_\epsilon(u)(x), \nabla y_\epsilon(u)(x))dx,$$

$\delta_{j\epsilon} = G_{j\epsilon}(\bar{u})$ if $1 \leq j \leq n_e$, and $\delta_{j\epsilon} = \max\{\delta_j, G_{j\epsilon}(\bar{u})\}$ if $j > n_e$.

THEOREM 5.3. *Under the assumptions of Theorem 5.2, for every $\epsilon > 0$ the problem $(Q_{\delta,\epsilon})$ has at least one solution u_ϵ. Moreover there exist real numbers $\mu_{j\epsilon}$, $j = 0, 1, \dots, n_e + n_i$, and elements $y_\epsilon \in H^1(\Omega) \cap L^\infty(\Omega)$ and $p_\epsilon \in H^1(\Omega)$ such that*

$$\text{(5.28)} \qquad \mu_{0\epsilon} \geq 0 \quad \textit{and} \quad \mu_{j\epsilon} \geq 0, \; n_e + 1 \leq j \leq n_e + n_i,$$

$$\text{(5.29)} \qquad \mu_{0\epsilon} + \sum_{j=1}^{n_e} |\mu_{j\epsilon}| + \sum_{j=n_e+1}^{n_e+n_i} \mu_{j\epsilon} > 0,$$

$$(5.30)\qquad \begin{cases} -\mathrm{div}\left\{(\epsilon + [k + |\nabla y_\epsilon|]^{\alpha-2})\nabla y_\epsilon\right\} + \lambda y_\epsilon = f & \textit{in } \Omega \\ \left(\epsilon + [k + |\nabla y_\epsilon|]^{\alpha-2}\right) \nabla y_\epsilon \cdot \vec{\nu} = u_\epsilon & \textit{on } \Gamma, \end{cases}$$

$$(5.31)\qquad \begin{cases} -\mathrm{div}\left\{(\epsilon + M_k(y_\epsilon))\nabla p_\epsilon\right\} + \lambda p_\epsilon = \mu_{0\epsilon}\dfrac{\partial L}{\partial y}(x, y_\epsilon) \\ \quad + \displaystyle\sum_{j=1}^{n_e+n_i} \mu_{j\epsilon}\left\{\frac{\partial g_j}{\partial y}(x, y_\epsilon, \nabla y_\epsilon) - \mathrm{div}\left(\frac{\partial g_j}{\partial \eta}(x, y_\epsilon, \nabla y_\epsilon)\right)\right\} \quad \textit{in } \Omega, \\ (\epsilon + M_k(y_\epsilon))\, \nabla p_\epsilon \cdot \vec{\nu} = \mu_{0\epsilon}\dfrac{\partial l}{\partial y}(x, y_\epsilon, u_\epsilon) \quad \textit{on } \Gamma, \end{cases}$$

with M_k *given by* (5.1),

$$(5.32)\qquad \mu_{j\epsilon}\left(\int_\Omega g_j(x, y_\epsilon(x), \nabla y_\epsilon(x))dx - \delta_{j\epsilon}\right) = 0, \quad j > n_e,$$

$$(5.33)\qquad \int_\Gamma \left(p_\epsilon + \mu_{0\epsilon}\left(\frac{\partial l}{\partial u}(x, y_\epsilon, u_\epsilon) + u_\epsilon - \bar{u}\right)\right)(u - u_\epsilon)d\sigma(x) \geq 0 \quad \forall u \in K.$$

Moreover,

$$(5.34)\qquad \begin{aligned} &\epsilon \int_\Omega |\nabla p_\epsilon|^2 dx + \int_\Omega \nabla p_\epsilon^T M_k(y_\epsilon)\nabla p_\epsilon dx + \int_\Omega \lambda p_\epsilon^2 dx \\ &\leq \mu_{0\epsilon}\left(\int_\Omega \frac{\partial L}{\partial y}(x, y_\epsilon)p_\epsilon dx + \int_\Gamma \frac{\partial l}{\partial y}(x, y_\epsilon, u_\epsilon)p_\epsilon d\sigma(x)\right) \\ &+ \sum_{j=1}^{n_e+n_i} \mu_{j\epsilon}\left(\int_\Omega \frac{\partial g_j}{\partial y}(x, y_\epsilon, \nabla y_\epsilon)p_\epsilon dx + \int_\Omega \frac{\partial g_j}{\partial \eta}(x, y_\epsilon, \nabla y_\epsilon)\nabla p_\epsilon dx\right). \end{aligned}$$

Proof. Because $\bar{u}$ is a feasible control for $(Q_{\delta,\epsilon})$, we are in the conditions of Theorem 4.1; thus we obtain the existence of a solution for $(Q_{\delta,\epsilon})$. The optimality conditions are derived as in Theorem 5.1; see Remark 5.2 below. □

The next result shows in what sense (P_δ) is approximated by the problems $(Q_{\delta,\epsilon})$.

THEOREM 5.4. *Let* u_ϵ *and* y_ϵ *be as they were in the previous theorem and let us denote* $\bar{y} = y_{\bar{u}}$. *Then as* $\epsilon \to 0$ *we have the following convergences*:

$$(5.35)\qquad u_\epsilon \to \bar{u} \quad \textit{in } L^t(\Gamma) \quad \forall\, t < +\infty,$$

$$(5.36)\qquad y_\epsilon \to \bar{y} \quad \textit{in } W^{1,\alpha}(\Omega),$$

$$(5.37)\qquad \tilde{J}_\epsilon(u_\epsilon) \to J(\bar{u}).$$

Moreover, $\{y_\epsilon\}_\epsilon$ *is uniformly bounded in* $L^\infty(\Omega)$.

The proof of this theorem uses the following lemma that can be proved as in [11, §4] (see also [18]) taking into account that the condition $\alpha > n/2$ assumed there can be removed because here the controls are bounded and $f \in L^p(\Omega)$ with $p > n/\alpha$.

LEMMA 5.5. *Let us suppose that* $\{v_\epsilon\}_\epsilon \subset L^\infty(\Gamma)$ *and* $v_\epsilon \to v$ *in the* **weak topology of* $L^\infty(\Gamma)$. *Then* $y_\epsilon(v_\epsilon) \to y_v$ *in* $W^{1,\alpha}(\Omega)$ *and in the* **weak topology of* $L^\infty(\Omega)$.

Proof of Theorem 5.4. Taking in the previous lemma $v_\epsilon = \bar{u}$ for all $\epsilon > 0$, we deduce that $y_\epsilon(\bar{u}) \to \bar{y}$ in $W^{1,\alpha}(\Omega)$. Then we can extract a subsequence of $\{y_\epsilon(\bar{u})\}_\epsilon$, denoted in the same way, such that

$$(5.38)\qquad y_\epsilon(\bar{u})(x) \to \bar{y}(x) \ \text{ a.e.}[\sigma]\ x \in \Gamma \quad \text{and} \quad y_\epsilon(\bar{u})(x) \to \bar{y}(x) \ \text{ a.e. } x \in \Omega,$$

$$\nabla y_\epsilon(\bar{u})(x) \to \nabla \bar{y}(x) \text{ a.e. } x \in \Omega, \tag{5.39}$$

and

$$\|y_\epsilon(\bar{u})\|_{L^\infty(\Omega)} \le C_1 \quad \text{and} \quad |\nabla y_\epsilon(\bar{u})(x)| \le \phi_1(x) \text{ a.e. } x \in \Omega \quad \forall\epsilon \tag{5.40}$$

for some positive constant C_1 and some function $\phi_1 \in L^\alpha(\Omega)$. From (5.38)–(5.40) and hypotheses (2.2), (2.6), and (2.7) it follows that $G_{j\epsilon}(\bar{u}) \to G_j(\bar{u})$ when ϵ tends to 0, which implies that $\delta_{j\epsilon} \to \delta_j$, $1 \le j \le n_e + n_i$.

On the other hand, since K is a *weakly closed bounded subset of $L^\infty(\Gamma)$ we can take another subsequence of $\{u_\epsilon\}_\epsilon$, again denoted in the same way, converging to an element $u \in K$ in the *weak topology of $L^\infty(\Gamma)$. Using Lemma 5.5 once again we deduce, taking a new subsequence if necessary,

$$y_\epsilon \to y_u \quad \text{in } W^{1,\alpha}(\Omega), \tag{5.41}$$

$$y_\epsilon(x) \to y_u(x) \text{ a.e. } [\sigma]\ x \in \Gamma \quad \text{and} \quad y_\epsilon(x) \to y_u(x) \text{ a.e. } x \in \Omega, \tag{5.42}$$

$$\nabla y_\epsilon(x) \to \nabla y_u(x) \text{ a.e. } x \in \Omega, \tag{5.43}$$

and

$$\|y_\epsilon\|_{L^\infty(\Omega)} \le C_2 \quad \text{and} \quad |\nabla y_\epsilon(x)| \le \phi_2(x) \text{ a.e. } x \in \Omega \quad \forall\epsilon \tag{5.44}$$

for some positive constant C_2 and some function $\phi_2 \in L^\alpha(\Omega)$. Arguing as above, we obtain that $G_{j\epsilon}(u_\epsilon) \to G_j(u)$ for every j. Hence u is a feasible control for problem (P_δ).

Hypotheses (2.2)–(2.4) and the convexity of l in the last variable allow us to deduce the lower semicontinuity of $\tilde{J}_\epsilon$ in the *weak topology of $L^\infty(\Gamma)$; see Ekeland and Temam [17]. Now, remembering that u_ϵ is a solution of $(Q_{\delta,\epsilon})$ and $\bar{u}$ is a feasible control for this problem, we obtain

$$\begin{aligned}
&\int_\Omega L(x, y_u)dx + \int_\Gamma l(x, y_u, u)d\sigma(x) + \frac{1}{2}\int_\Gamma (u - \bar{u})^2 d\sigma(x) \\
&\quad \le \liminf_{\epsilon\to 0} \tilde{J}_\epsilon(u_\epsilon) \le \limsup_{\epsilon\to 0} \tilde{J}_\epsilon(u_\epsilon) \le \limsup_{\epsilon\to 0} \tilde{J}_\epsilon(\bar{u}) = J(\bar{u}) \le J(u) \\
&\quad = \int_\Omega L(x, y_u)dx + \int_\Gamma l(x, y_u, u)d\sigma(x),
\end{aligned}$$

which implies that $u = \bar{u}$ and convergence (5.37). Finally, from the relations

$$\begin{aligned}
&\frac{1}{2}\limsup_{\epsilon\to 0} \int_\Gamma (u_\epsilon - \bar{u})^2 d\sigma(x) \\
&\quad \le \limsup_{\epsilon\to 0} \left(\tilde{J}_\epsilon(u_\epsilon) - \int_\Omega L(x, y_\epsilon)dx - \int_\Gamma l(x, y_\epsilon, u_\epsilon)d\sigma(x)\right) \\
&\quad = J(\bar{u}) - \liminf_{\epsilon\to 0} \left\{\int_\Omega L(x, y_\epsilon)dx + \int_\Gamma l(x, y_\epsilon, u_\epsilon)d\sigma(x)\right\} \le J(\bar{u}) - J(u) \le 0,
\end{aligned}$$

we deduce the desired convergence of $\{u_\epsilon\}_\epsilon$ to $\bar{u}$. □

Proof of Theorem 5.2. The idea is to pass to the limit in (5.28)–(5.34). It follows exactly as in Theorem 5.1, except for some peculiarities that can be summarized as follows:

- We can suppose without loss of generality that

$$\omega_\epsilon = \mu_{0\epsilon} + \sum_{j=1}^{n_e} |\mu_{j\epsilon}| + \sum_{j=n_e+1}^{n_e+n_i} \mu_{j\epsilon} = 1.$$

In the other case, it is enough to divide the expressions (5.31)–(5.34) by ω_ϵ and rename $\omega_\epsilon^{-1}p_\epsilon$ and $\omega_\epsilon^{-1}\mu_{j\epsilon}$ as p_ϵ and $\mu_{j\epsilon}$, respectively.

- After Theorems 5.3 and 5.4 and Lemma 5.5, the main question to prove is the boundedness of $\{p_\epsilon\}_\epsilon$ in $W^{1,\alpha}(\Omega)$. First, applying Hölder's inequality, with $q = 2/\alpha$ and $q' = 2/(2-\alpha)$, and thanks to (5.36) we obtain that

$$\begin{aligned}
&\|\nabla p_\epsilon\|^2_{L^\alpha(\Omega)^n} + \|p_\epsilon\|^2_{L^\alpha(\Omega)} \\
&\leq C_1 \left(\int_\Omega [k + |\nabla y_\epsilon|]^{\alpha-2} |\nabla p_\epsilon|^2 dx\right) \left(\int_\Omega [k + |\nabla y_\epsilon|]^\alpha\right)^{(2-\alpha)/\alpha} \\
&\quad + C_2 \int_\Omega |p_\epsilon|^2 dx \leq C_3 \left(\int_\Omega [k + |\nabla y_\epsilon|]^{\alpha-2} |\nabla p_\epsilon|^2 dx + \int_\Omega |p_\epsilon|^2 dx\right).
\end{aligned}$$

On the other hand, from hypotheses (2.4)–(2.7), the values of τ, ρ, and θ, the Sobolev imbedding theorem for $W^{1,\alpha}(\Omega)$, and Lemma 5.5, it follows that the right-hand terms of problem (5.31) define a bounded element of $(W^{1,\alpha}(\Omega))'$. Hence, with the aid of (5.34), we obtain

$$\begin{aligned}
&(\alpha - 1) \int_\Omega [k + |\nabla y_\epsilon|]^{\alpha-2} |\nabla p_\epsilon|^2 dx + \lambda \int_\Omega |p_\epsilon|^2 dx \\
&\leq \mu_{0\epsilon} \left(\int_\Omega \frac{\partial L}{\partial y}(x, y_\epsilon) p_\epsilon dx + \int_\Gamma \frac{\partial l}{\partial y}(x, y_\epsilon, u_\epsilon) p_\epsilon d\sigma(x)\right) \\
&\quad + \sum_{j=1}^{n_e+n_i} \mu_{j\epsilon} \left(\int_\Omega \frac{\partial g_j}{\partial y}(x, y_\epsilon, \nabla y_\epsilon) p_\epsilon dx + \int_\Omega \frac{\partial g_j}{\partial \eta}(x, y_\epsilon, \nabla y_\epsilon) \nabla p_\epsilon dx\right) \\
&\leq C_4 \|p_\epsilon\|_{W^{1,\alpha}(\Omega)} \quad \forall \epsilon > 0.
\end{aligned}$$

- The weak convergence (5.26) can now be derived noting that

$$\|U_\epsilon(x)\|_s \leq C_5 \text{ a.e. } x \in \Omega \quad \forall \epsilon$$

(with the notations established in the proof of Theorem 5.1) and therefore

$$U_\epsilon \rightharpoonup U \quad \text{in } (L^t(\Omega))^{n\times n} \quad \forall\, t < +\infty,$$

in particular, for $t = \alpha/(\alpha - 1)$.

- Finally, let us note that the boundedness of $\epsilon^{1/2}\|\nabla p_\epsilon\|_{L^2(\Omega)^n}$ (that follows from (5.34)) and Hölder's inequality imply that

$$\epsilon \int_\Omega \nabla\psi \nabla p_\epsilon dx \to 0 \quad \forall \psi \in C^\infty(\bar{\Omega}). \qquad \square$$

Remark 5.2. The results stated in Theorems 5.1 and 5.2 can be extended to more general quasilinear elliptic equations such as that given in (3.1), under assumptions

(3.2)–(3.7), following essentially the same argumentation; some technical changes appear in relation to the adjoint state equation. In this framework equation (5.5) and inequality (5.8) become

$$\left\{\begin{array}{l} -\operatorname{div}\left(\left[\dfrac{\partial a}{\partial \eta}(x,\nabla\bar{y})\right]^T \nabla\bar{p}\right)+\dfrac{\partial a_0}{\partial s}(x,\bar{y})\bar{p}=\bar{\mu}_0\dfrac{\partial L}{\partial y}(x,\bar{y}) \\ \quad +\displaystyle\sum_{j=1}^{n_e+n_i}\bar{\mu}_j\left\{\dfrac{\partial g_j}{\partial y}(x,\bar{y},\nabla\bar{y})-\operatorname{div}\left(\dfrac{\partial g_j}{\partial \eta}(x,\bar{y},\nabla\bar{y})\right)\right\} \text{ in } \Omega, \\ \qquad \left(\dfrac{\partial a}{\partial \eta}(x,\nabla\bar{y})\right)^T \nabla\bar{p}\cdot\vec{\nu}=\bar{\mu}_0\dfrac{\partial l}{\partial y}(x,\bar{y},\bar{u}) \quad \text{on } \Gamma, \end{array}\right.$$

and

$$\begin{aligned} &\int_\Omega \nabla\bar{p}^T\frac{\partial a}{\partial \eta}(x,\nabla\bar{y})\nabla\bar{p}dx+\int_\Omega\frac{\partial a_0}{\partial s}(x,\bar{y})\bar{p}^2dx \\ &\le \bar{\mu}_0\left(\int_\Omega\frac{\partial L}{\partial y}(x,\bar{y})\bar{p}dx+\int_\Gamma\frac{\partial l}{\partial y}(x,\bar{y},\bar{u})\bar{p}d\sigma(x)\right) \\ &\quad+\sum_{j=1}^{n_e+n_i}\bar{\mu}_j\left(\int_\Omega\frac{\partial g_j}{\partial y}(x,\bar{y},\nabla\bar{y})\bar{p}dx+\int_\Omega\frac{\partial g_j}{\partial \eta}(x,\bar{y},\nabla\bar{y})\nabla\bar{p}dx\right), \end{aligned}$$

respectively.

6. Optimality conditions: $k = 0$. When $k = 0$ and $\alpha > 2$, the operator of the state equation (2.1) becomes degenerate in the sense that its modulus of ellipticity vanishes in the subset $\{x \in \Omega : \nabla y(x) = 0\}$; when $\alpha < 2$, the operator is singular because that modulus blows up in the same subset. Nevertheless, we can still derive some optimality conditions for (P_δ) with the peculiarity that the adjoint state equation is only satisfied in the subset Ω_0 defined by

$$\Omega_0 = \{x \in \Omega : \nabla\bar{y}(x) \neq 0\},$$

where $\bar{y} = y_{\bar{u}}$ (this is an open set because $\nabla\bar{y}$ is continuous in Ω; see [32]). To do this we introduce a family of approximate problems that belong to the case $k > 0$ and we pass to the limit in their optimality systems making $k \longrightarrow 0$.

We will only deal with the case $\alpha < 2$. For $\alpha > 2$, we can prove that the adjoint states corresponding to the approximate problems converge in $H^1_{\text{loc}}(\Omega_0)$, but not on Γ (even when $\Omega_0 = \Omega$). Therefore, we are not able to derive the condition (5.7) and the optimality conditions obtained lose interest.

THEOREM 6.1. *Let $\bar{u}$ be a solution of (P_δ), $\alpha < 2$ and $k = 0$. Let us assume that $\tau \geq (n\alpha - \alpha)/(n\alpha - n)$, $\rho \geq (\alpha n)/(\alpha n + \alpha - n)$, $\theta \geq \alpha/(\alpha - 1)$, and $l(x, y, \cdot) : R \longrightarrow R$ is a convex function for all $(x, y) \in \Omega \times R$. Then there exist real numbers $\bar{\mu}_j$, $j = 0, 1, \ldots, n_e + n_i$, and elements $\bar{y} \in W^{1,\alpha}(\Omega) \cap L^\infty(\Omega)$ and $\bar{p} \in W^{1,\alpha}(\Omega)$ verifying* (5.2)–(5.4) (*with* $k = 0$), (5.6)–(5.7), *and*

$$\begin{aligned} &-\operatorname{div}\{M_0(\bar{y})\nabla\bar{p}\}+\lambda\bar{p}=\bar{\mu}_0\frac{\partial L}{\partial y}(x,\bar{y}) \\ (6.1)\qquad &+\sum_{j=1}^{n_e+n_i}\bar{\mu}_j\left\{\frac{\partial g_j}{\partial y}(x,\bar{y},\nabla\bar{y})-\operatorname{div}\left(\frac{\partial g_j}{\partial \eta}(x,\bar{y},\nabla\bar{y})\right)\right\} \textit{ in } \Omega_0 \end{aligned}$$

with M_0 given by (5.1).

Sketch of the proof. Given $\epsilon > 0$, let us consider the perturbed operator given by

$$A_\epsilon y = -\mathrm{div}\left\{[\epsilon + |\nabla y|]^{\alpha-2}\nabla y\right\} + \lambda y,$$

and the family of approximate problems $(Q_{\delta,\epsilon})$ of §5.2. The operator A_ϵ now verifies hypotheses (3.2)–(3.7) with exponent α and constant $k = \epsilon$. Consequently, given $u \in L^\infty(\Gamma)$, the Neumann problem

$$\begin{cases} A_\epsilon y = f & \text{in } \Omega, \\ (\epsilon + |\nabla y|)^{\alpha-2}\nabla y \cdot \vec{\nu} = u & \text{on } \Gamma, \end{cases}$$

has a unique solution $y_\epsilon(u) \in W^{1,\alpha}(\Omega) \cap L^\infty(\Omega)$.

Now applying Theorem 5.2 to each control problem $(Q_{\delta,\epsilon})$, we derive a result analogous to that of Theorem 5.3 with $y_\epsilon \in W^{1,\alpha}(\Omega) \cap L^\infty(\Omega)$, $p_\epsilon \in W^{1,\alpha}(\Omega)$, and $k = \epsilon$. Of course, the terms $\epsilon\triangle y_\epsilon$, $\epsilon\nabla y_\epsilon$, $\epsilon\triangle p_\epsilon$, $\epsilon\nabla p_\epsilon$, and $\epsilon\int_\Omega |\nabla p_\epsilon|^2 dx$ do not appear. Furthermore, Lemma 5.5 is still valid (see [8, Lem. 4.5] for the Dirichlet problem), and thus Theorem 5.4 is too.

We can suppose again that

$$\omega_\epsilon = \mu_{0\epsilon} + \sum_{j=1}^{n_e} |\mu_{j\epsilon}| + \sum_{j=n_e+1}^{n_e+n_i} \mu_{j\epsilon} = 1,$$

and pass to the limit as $\epsilon \to 0$ without difficulty in (5.28)–(5.30) and (5.32) to obtain (5.2)–(5.4) (with $k = 0$) and (5.6).

The boundedness of $\{p_\epsilon\}_\epsilon$ in $W^{1,\alpha}(\Omega)$ follows exactly as in the case $\alpha < 2$, $k > 0$. Thus, at least for a subsequence, it follows that

$$p_\epsilon \longrightarrow \bar{p} \quad \text{weakly in } W^{1,\alpha}(\Omega)$$

for a certain element $\bar{p} \in W^{1,\alpha}(\Omega)$.

To establish (6.1), we utilize a regularity result due to Tolksdorf [32]. Thanks to this result, $\bar{y}$ and $y_\epsilon(u)$ belong to $C^{1,t}(\Omega)$ for some $0 < t < 1$. Let Ω' be an open set such that $\Omega' \subset \overline{\Omega'} \subset \Omega_0$. Therefore, there exists $C_1 > 0$ such that

$$(6.2) \qquad |\nabla\bar{y}(x)| > C_1 \quad \forall x \in \overline{\Omega'}.$$

Now, using the estimates given in [32, Thm. 1], we can apply the Ascoli–Arzelá theorem to the family $\{\nabla y_\epsilon\}_\epsilon$ for deriving that $\{\nabla y_\epsilon\}_\epsilon$ is precompact in $C(\overline{\Omega'})^n$. Together with (5.36) this implies the existence of a subsequence, denoted in the same way, such that

$$(6.3) \qquad \nabla y_\epsilon \to \nabla\bar{y} \quad \text{in } C(\overline{\Omega'})^n.$$

From (6.2) and (6.3) it follows that there exists $\epsilon' > 0$ such that

$$(6.4) \qquad |\nabla y_\epsilon(x)| > C_1 \quad \forall x \in \overline{\Omega'} \quad \forall \epsilon < \epsilon'.$$

To deduce (6.1) it is now enough to multiply the equation of (5.31) by $\phi \in D(\Omega_0)$, integrate by parts, and pass to the limit as $\epsilon \to 0$ with the aid of estimate (6.4) for $\Omega' \supset \mathrm{sop}\phi$.

Finally, it is immediate to obtain (5.7) passing to the limit in (5.33). □

Remark 6.1. Theorem 6.1 can be extended to a wider class of quasilinear elliptic equations of the type

$$Ay = -\mathrm{div}\,(a(x, |\nabla y|)\nabla y) + a_0(x, y)$$

with $a : \Omega \times (0, +\infty) \to (0, +\infty)$; see [8] for the Dirichlet problem.

7. A qualification condition for $(\mathbf{P}_\delta)$. In the absence of equality constraints (i.e., $n_e = 0$), it is possible to deduce the previous optimality systems in a qualified form (i.e., with $\bar\mu_0 = 1$) for almost every problem (P_δ).

THEOREM 7.1. *Let us suppose that $n_e = 0$ and Δ is a cube of R^{n_i} such that (P_δ) has at least one solution for every $\delta \in \Delta$. Then, Theorems 5.1, 5.2, and 6.1 remain valid for each solution $\bar u$ of (P_δ) with $\bar\mu_0 = 1$ for almost every $\delta \in \Delta$.*

Proof. The proof is a modified version of Clarke's argument [14, §4]. Let us consider the function $\phi : \Delta \to R$ defined by

$$\phi(\delta) = \inf\{J(u) : u \in K,\quad G_j(u) \le \delta_j,\ 1 \le j \le n_i\} = \inf(P_\delta).$$

It is obvious that $\phi(\delta)$ is decreasing as a function of each component of δ separately and, thus, it is differentiable almost everywhere. Now, let us fix $\delta \in \Delta$ such that there exists $D\phi(\delta)$. Therefore, given a $\bar u$ solution of (P_δ), we can argue as in [14, Thm. 2] to derive the existence of strictly positive constants r and γ such that $\bar u$ is a solution of the following problem:

$$(Q_{\delta,r,\gamma}) \begin{cases} \text{minimize } J(u) + r\displaystyle\sum_{j=1}^{n_i}(G_j(u) - \delta_j)^+ \\ \text{such that } \ u \in K_\gamma, \end{cases}$$

where $K_\gamma = K \cap \{u \in L^\infty(\Gamma) : \|u - \bar u\|_{L^\infty(\Gamma)} \le \gamma\}$. The proof will be completed when we obtain the optimality system for $(Q_{\delta,r,\gamma})$.

In the case $\alpha \ge 2$ and $k \ne 0$, this can be done again using Ekeland's variational principle. To do this, let us introduce the family of problems

$$(Q_{\delta,r,\gamma,s}) \begin{cases} \text{minimize } \ J_s(u) \\ \text{such that } \ u \in K_\gamma, \end{cases}$$

where $s > 1$ and

$$J_s(u) = J(u) + r\sum_{j=1}^{n_i}\left\{\frac{1}{s^s} + [(G_j(u) - \delta_j)^+]^s\right\}^{1/s}.$$

Obviously, K_γ is a complete metric space with $d(u,v) = \|u - v\|_{L^\infty(\Gamma)}$, $J_s : K_\gamma \longrightarrow R$ is a continuous function, and

$$J_s(\bar u) = J(\bar u) + \frac{rn_i}{s} \le \frac{rn_i}{s} + \inf(Q_{\delta,r,\gamma,s}).$$

Hence, applying Ekeland's variational principle [16], we deduce the existence of an element $u_s \in K_\gamma$ such that

$$\|u_s - \bar u\|_{L^\infty(\Gamma)} \le \sqrt{\gamma_s} \tag{7.1}$$

with $\gamma_s = rn_i/s$ and u_s is a solution of problem

$$\begin{cases} \text{minimize } \ J_s(u) + \sqrt{\gamma_s}\|u_s - u\|_{L^\infty(\Gamma)} \\ \text{such that } \ u \in K_\gamma. \end{cases}$$

It follows from (7.1) that the constraint $\|u_s - \bar u\|_{L^\infty(\Gamma)} \le \gamma$ is not active for every $s > (rn_i)/\gamma^2$. Thus we conclude the proof exactly as in the proof of Theorem 5.1, noting that J_s is Gâteaux differentiable,

$$J_s'(u_s)\cdot(u - u_s) = \left(J'(u_s) + \sum_{j=1}^{n_i}\mu_{js}G_j'(u_s)\right)\cdot(u - u_s),$$

where

$$\mu_{js} = r\left\{\frac{1}{s^s} + [(G_j(u_s) - \delta_j)^+]^s\right\}^{(1/s)-1} [(G_j(u_s) - \delta_j)^+]^{s-1} \le r$$

for every $s > 1$ and $1 \le j \le n_i$, and passing to the limit as $s \to +\infty$.

If $\alpha < 2$ we approximate $(Q_{\delta,r,\gamma})$ by problems

$$(Q_{\delta,r,\gamma,\epsilon}) \begin{cases} \text{minimize } \tilde{J}_\epsilon(u) + r\sum_{j=1}^{n_i}(G_j(u) - \delta_j)^+ \\ \text{such that } \ u \in K_\gamma, \end{cases}$$

with $\tilde{J}_\epsilon$ defined as in §5.2 for $k > 0$ and §6 for $k = 0$. Now, $(Q_{\delta,r,\gamma,\epsilon})$ falls into the case $\alpha \ge 2$ and $k \ne 0$. Therefore we can use the results established above in this proof to obtain an optimality system for $(Q_{\delta,r,\gamma,\epsilon})$ with Lagrange multipliers $\mu_{0\epsilon} = 1$ and $\mu_{j\epsilon} \le r$ for all $\epsilon > 0$ and $j = 1, \ldots, n_i$. Finally, we pass to the limit as in the proofs of Theorems 5.2 and 6.1 to derive the desired result. □

Remark 7.1. In case $\alpha \ge 2$ and $n_e = 0$, and under the following Slater-type condition: there exists $u_o \in K$ such that

$$G_j(\bar{u}) + DG_j(\bar{u}) \cdot (u_0 - \bar{u}) < \delta_j, \quad j = 1, \ldots, n_i,$$

we can obtain the optimality conditions in a qualified form by applying [5, Thm. 5.2]; see Remark 5.1.

REFERENCES

[1] W. Ames, *Nonlinear Partial Differential Equations in Engineering*, Academic Press, New York, London, 1965.

[2] R. Aris, *The Mathematical Theory of Diffusion and Reaction in Permeable Catalysts*, Vols. 1 and 2, Clarendon Press, Oxford, 1975.

[3] V. Barbu and T. Precupanu, *Convexity and Optimization in Banach Spaces*, Editura Academiei, Sijthoff and Noordhoff, Bucharest, 1978.

[4] J. Bonnans and E. Casas, *Optimal control of semilinear multistate systems with state constraints*, SIAM J. Control Optim., 27 (1989), pp. 446–455.

[5] E. Casas, *Boundary control of semilinear elliptic equations with pointwise state constraints*, SIAM J. Control Optim., 31 (1993), pp. 993–1006.

[6] E. Casas and L. Fernández, *A Green's formula for quasilinear elliptic operators*, J. Math. Anal. Appl., 142 (1989), pp. 62–72.

[7] ———, *Optimal control of quasilinear elliptic equations*, in Control of Partial Differential Equations, Lecture Notes in Control and Information Sciences 114, A. Bermúdez, ed., Springer-Verlag, Berlin, Heidelberg, New York, 1989, pp. 92–99.

[8] ———, *Optimal control of quasilinear elliptic equations with non differentiable coefficients at the origin*, Rev. Mat. Univ. Complut. Madrid, 4 (1991), pp. 227–250.

[9] ———, *Optimality conditions for state-constrained control problems of quasilinear elliptic equations*, in 30th IEEE Conference on Control and Decision, Brighton, England, 1991, pp. 1991–1995.

[10] ———, *State-constrained control problems of quasilinear elliptic equations*, in Optimal Control of Partial Differential Equations, Lecture Notes in Control and Information Sciences 149, K. Hoffmann and W. Krabs, eds., Springer-Verlag, Berlin, Heidelberg, New York, 1991, pp. 11–25.

[11] ———, *Distributed control of systems governed by a general class of quasilinear elliptic equations*, J. Differential Equations, 104 (1993), pp. 20–47.

[12] L. Cesari, *Optimization. Theory and Applications*, Springer-Verlag, New York, Heidelberg, Berlin, 1983.

[13] I. Chryssoverghi, *Nonconvex optimal control of nonlinear monotone parabolic systems*, Systems Control Lett., 8 (1986), pp. 55–62.

[14] F. Clarke, *A new approach to Lagrange multipliers*, Math. Oper. Res., 1 (1976), pp. 165–174.

[15] C. Coffman, V. Duffin, and V. Mizel, *Positivity of weak solutions of non-uniformly elliptic equations*, Ann. Mat. Pura Appl., 104 (1975), pp. 209–238.

[16] I. Ekeland, *Nonconvex minimization problems*, Bull. Amer. Math. Soc., 1 (1979), pp. 76–91.

[17] I. Ekeland and R. Temam, *Analyse Convexe et Problèmes Variationnels*, Dunod/Gauthier-Villars, Paris, 1974.

[18] L. Fernández, *Control Optimo de Sistemas Gobernados por Ecuaciones Elípticas Cuasilineales*, Ph.D. thesis, Universidad de Cantabria, Spain, 1990.

[19] H. Halkin, *Implicit functions and optimization problems without continuous differentiability of the data*, SIAM J. Control, 12 (1974), pp. 229–236.

[20] E. Isaacson and H. Keller, *Analysis of Numerical Methods*, John Wiley, New York, 1966.

[21] I. Lasiecka, *State constrained control problems for parabolic systems: Regularity of optimal solutions*, Appl. Math. Optim., 6 (1980), pp. 1–29.

[22] J. Lions, *Quelques Méthodes de Résolution des Problèmes aux Limites non Linéaires*, Dunod, Paris, 1969.

[23] J. Lions and E. Magenes, *Problèmes aux Limites non Homogènes*, Dunod, Paris, 1968.

[24] U. Mackenroth, *On parabolic distributed optimal control problems with restrictions on the gradient*, Appl. Math. Optim., 10 (1983), pp. 69–95.

[25] O. Mangasarian and S. Fromovitz, *The Fritz John necessary optimality conditions in the presence of equality and inequality constraints*, J. Math. Anal. Appl., 17 (1967), pp. 37–47.

[26] E. McShane, *The Lagrange multiplier rule*, Amer. Math. Monthly, 81 (1973), pp. 922–925.

[27] M. Murthy and G. Stampacchia, *Boundary value problems for some degenerate elliptic operators*, Ann. Mat. Pura Appl., 80 (1968), pp. 1–122.

[28] J. Nečas, *Les Méthodes Directes en Théorie des Equations Elliptiques*, Editeurs Academia, Prague, 1967.

[29] M. Pelissier, *Sur quelques problémes non linéaires en glaciologie*, Ph.D. thesis, Université d'Orsay, 1975.

[30] B. Pourciau, *Modern multiplier rules*, Amer. Math. Monthly, 87 (1980), pp. 433–452.

[31] G. Stampacchia, *Le problème de Dirichlet pour les équations elliptiques du second ordre à coefficients discontinus*, Ann. Inst. Fourier (Grenoble), 15 (1965), pp. 189–258.

[32] P. Tolksdorf, *Regularity for a more general class of quasi-linear elliptic equations*, J. Differential Equations, 51 (1984), pp. 126–150.

[33] F. Tröltzsch, *Optimality Conditions for Parabolic Control Problems and Applications*, Teubner-Texte, Leipzig, 1984.

[34] N. Trudinger, *Linear elliptic equations with measurable coefficients*, Ann. Scuola Norm. Sup. Pisa, 27 (1973), pp. 265–308.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 33, No. 2, pp. 590–624, March 1995

© 1995 Society for Industrial and Applied Mathematics
013

THE STOCHASTIC MAXIMUM PRINCIPLE FOR LINEAR, CONVEX OPTIMAL CONTROL WITH RANDOM COEFFICIENTS*

ABEL CADENILLAS† AND IOANNIS KARATZAS‡

This paper is dedicated to Professor Herbert Robbins on the occasion of his 80*th birthday.*

Abstract. This paper considers a stochastic control problem with linear dynamics, convex cost criterion, and convex state constraint, in which the control enters both the drift and diffusion coefficients. These coefficients are allowed to be random, and no L^p-bounds are imposed on the control. An explicit solution for the adjoint equation and a global stochastic maximum principle are obtained for this model. This is evidently the first version of the stochastic maximum principle that covers the consumption-investment problem. The mathematical tools are those of stochastic calculus and convex analysis.

When it is assumed, as in other versions of the stochastic maximum principle, that the admissible controls are square-integrable, not only a necessary but also a sufficient condition for optimality is obtained. It is then shown that this particular case of the general model may be applied to solve a variety of problems in stochastic control, including the linear-regulator, predicted-miss, and Beneš problems.

Key words. stochastic maximum principle, convex analysis, stochastic control, backwards stochastic differential equations, adjoint equation, consumption-investment problem

AMS subject classifications. 93E20, 60H30, 90A09, 60G44, 90A16, 49N10

1. The stochastic maximum principle.

1.1. Introduction.

We consider the stochastic differential equation

$$dX_t = f(t, X_t, u_t)dt + g(t, X_t, u_t)dW_t, \tag{1}$$

$$X_0 = x, \tag{2}$$

where W is a Brownian motion on a probability space $(\Omega, \mathcal{F}, P, (\mathcal{F}_t))$, u is a suitable control process adapted to $(\mathcal{F}_t)$, and X is the trajectory of the system controlled by u. We require that X take values in a fixed set V. Our problem is to choose u in such a way as to minimize a functional of the type

$$E\left[\int_0^T L(t, X_t, u_t)dt + \Psi(X_T)\right]. \tag{3}$$

A control process that solves this problem is called *optimal.* In §1, we impose conditions on f, g, L, Ψ, and V to obtain a stochastic maximum principle for our model. Roughly speaking, this principle asserts the following: we assume that $\hat{u}$ is an optimal control, define the *Hamiltonian*

$$H(t, p, q, x, u) = -L(t, x, u) + p \bullet f(t, x, u) + q \bullet g(t, x, u), \tag{4}$$

and consider the solution of the *adjoint equation*

$$dp_t = -H_x(t, p_t, q_t, \hat{X}_t, \hat{u}_t)dt + q_t dW_t, \tag{5}$$

$$p_T = -\Psi_x(\hat{X}_T) \tag{6}$$

* Received by the editors November 25, 1992; accepted for publication (in revised form) November 11, 1993. The results of this paper have been drawn from the first author's doctoral dissertation at Columbia University. This research was supported in part by U.S. Army Research Office grant DAAL 03-91-G-0170.

† Isaac Newton Institute for Mathematical Sciences, University of Cambridge, 20 Clarkson Road, Cambridge CB3 OEH, United Kingdom.

‡ Department of Statistics, Columbia University, New York, New York 10027.

for the pair of $(\mathcal{F}_t)$-adapted processes (p, q), where $\hat{X}$ is the trajectory of the system controlled by $\hat{u}$. The stochastic maximum principle states roughly that, under certain conditions on f, g, L, Ψ, and V,

$$\max_u H(t, p_t, q_t, \hat{X}_t, u) = H(t, p_t, q_t, \hat{X}_t, \hat{u}_t), \tag{7}$$

i.e., *that a necessary condition for the optimality of a control is to maximize the Hamiltonian.*

The initial work on the stochastic maximum principle was done by Kushner [31], [32]. Then in the 1970s, Haussmann developed a powerful version of the stochastic maximum principle (see [21] for an account of his theory and references to his previous work [15]–[20] on this subject), and applied it to solve some important problems in stochastic control [18]. The main limitation of Haussmann's theory is that the control does not affect the diffusion coefficient.

Versions of the stochastic maximum principle in which the diffusion coefficient is controlled were developed by Bensoussan [4], Elliott [12], and Peng [34] in a form weaker than (7), and by Arkin and Saksonov [1], Bismut [5], [6], [7], and Saksonov [38] in the strong form (7). The main limitation of these approaches is that they impose L^p-bounds on the controls. Other limitations are that they assume the trajectory to be unconstrained, and the running cost, terminal cost, and/or their derivatives to have polynomial growth. Furthermore, only Arkin and Saksonov [1], Bismut [5], [6], [7], and Saksonov [38] consider the case of random coefficients.

The present paper treats linear systems with random coefficients and with the control affecting the diffusion term. Its two main contributions are the development of a stochastic maximum principle in its strongest form (7) *without imposing L^p-bounds on the controls*, and the explicit solution of the adjoint equation for the resulting model. As pointed out by Whittle [41, p. 183], the solution of the adjoint equation is a difficult problem, and it is one of the two main difficulties that face the stochastic maximum principle (the other being that the principle is not simple when the control affects the diffusion coefficient). An additional contribution of our paper is that it also allows for state-constraints and does not impose polynomial growth conditions on the cost functions or their derivatives. Furthermore, we derive necessary as well as sufficient conditions for optimality, and show that these coincide in the presence of L^2-bounds on the class of admissible controls.

In §2, we apply our version of the stochastic maximum principle to the consumption-investment problem. This important problem could not be covered by any of the previous versions of the stochastic maximum principle. The reasons were multiple: in this problem the optimal control is not square-integrable, the trajectory of the system is constrained, and the running cost, terminal cost, and their derivatives do not necessarily obey global polynomial growth conditions. (In this problem, the typical running cost and/or terminal cost is the negative of the logarithmic function.)

In §3, we make the additional assumption that the controls are square-integrable, and notice that, in this particular case of our general model, a *necessary and sufficient* condition for the optimality of a control is the maximization of the Hamiltonian. We also show that this particular version of our general model may be applied to solve completely other stochastic control problems as well, including the linear-regulator, predicted-miss, and Beneš problems.

After this paper was submitted for publication, Cadenillas and Haussmann [9] generalized some of the results of §1 and §3 to obtain a stochastic maximum principle for a problem in which the control has both absolutely continuous and singular components.

1.2. Notations and assumptions. We suppose that a d-dimensional Brownian motion W is defined on a complete probability space $(\Omega, \mathcal{F}, (\mathcal{F}_t), P)$. Here, $(\mathcal{F}_t)$ is the P-augmentation of the natural filtration $(\mathcal{F}_t^W)$ defined by

$$\mathcal{F}_t^W = \sigma(W(s) : 0 \leq s \leq t) \quad \forall t \in [0, \infty).$$

Let T be a fixed strictly positive real number, U a closed convex subset of $\Re^k$, and let us consider the functions

$$\begin{aligned}
A &: [0,T] \times \Omega \longmapsto \mathcal{L}(\Re^n; \Re^n), \\
B &: [0,T] \times \Omega \longmapsto \mathcal{L}(\Re^k; \Re^n), \\
C &: [0,T] \times \Omega \longmapsto \Re^n, \\
D &: [0,T] \times \Omega \longmapsto \mathcal{L}(\Re^n; \mathcal{L}(\Re^d; \Re^n)), \\
E &: [0,T] \times \Omega \longmapsto \mathcal{L}(\Re^k; \mathcal{L}(\Re^d; \Re^n)), \\
F &: [0,T] \times \Omega \longmapsto \mathcal{L}(\Re^d; \Re^n).
\end{aligned}$$

Here $\mathcal{L}(V; W)$ denotes the space of linear transformations of a vector space V into a vector space W. We shall assume that $A, B, C, D, E,$ and F are all progressively measurable with respect to $(\mathcal{F}_t)$, and bounded uniformly in $(t, \omega) \in [0, \infty) \times \Omega$. Let us now consider the linear stochastic differential equation

$$dX_t = f(t, X_t, u_t)dt + g(t, X_t, u_t)dW_t, \tag{8}$$

$$X_0 = x, \tag{9}$$

where

$$f(t, x, u) = A_t x + B_t u + C_t, \tag{10}$$

$$g(t, x, u) = D_t x + E_t u + F_t. \tag{11}$$

In order that the above stochastic differential equation make sense, we need that

$$P\left\{ \int_0^T |f(t, X_t, u_t)| dt < \infty \right\} = 1 \tag{12}$$

and

$$P\left\{ \int_0^T |g(t, X_t, u_t)|^2 dt < \infty \right\} = 1. \tag{13}$$

A sufficient condition for this to happen is

$$P\left\{ \int_0^T |B_t u_t| dt < \infty, \quad \int_0^T |E_t u_t|^2 dt < \infty \right\} = 1. \tag{14}$$

Corresponding to any $(\mathcal{F}_t)$-adapted control process $u : [0,T] \times \Omega \mapsto U \subset \Re^k$ that is $\mathcal{B}([0,T]) \otimes \mathcal{F}/\mathcal{B}(\Re^k)$-measurable, and satisfies (14), we denote by X^u the solution of the linear stochastic integral equation

$$X_t = x + \int_0^t (A_s X_s + B_s u_s + C_s) ds + \int_0^t (D_s X_s + E_s u_s + F_s) dW_s, \quad 0 \leq t \leq T, \tag{15}$$

and call it the *trajectory of the linear system* (8)–(9) *controlled by* u. Evidently, from the linearity of this system and for any control processes u, v as above, we have for any $\alpha \in [0,1]$

$$X^{\alpha u+(1-\alpha)v} = \alpha X^u + (1-\alpha)X^v, \quad \text{a.s.} \tag{16}$$

We shall restrict the class of admissible control processes u in the following manner.

DEFINITION 1.1 (admissibility). *Let V be a fixed, nonempty, convex subset of $\Re^n$. For any $x \in V$, we shall denote by $\mathcal{U}(x,V)$ the class of* admissible control processes *$u : [0,T] \times \Omega \mapsto U$ that are measurable, adapted to $(\mathcal{F}_t)$, satisfy condition* (14), *and are such that the corresponding trajectory X^u of* (15) *satisfies*

$$X_t^u \in V \quad \forall t \in [0,T] \quad \text{a.s.}$$

Whenever x and V are fixed, we shall denote $\mathcal{U}(x,V)$ by $\mathcal{U}$, without danger of confusion.

Quite obviously from this definition and (16), we have the following remark.

controls is convex.

Let us now consider the measurable functions $\Psi : \Omega \mapsto C^1(V;\Re)$ and $L : [0,T] \times \Omega \mapsto C^{1,1}(V \times U;\Re)$. We assume that Ψ is $\mathcal{F}_T$-measurable, that L is $(\mathcal{F}_t)$-progressively measurable, and that for each $(t,\omega) \in [0,T] \times \Omega$, $L(t,\cdot,\cdot) \in C^{1,1}(V \times U;\Re)$ and $\Psi(\cdot) \in C^1(V;\Re)$ are convex functions. This means that $\forall (t,\omega) \in [0,T] \times \Omega, x \in V, y \in V, u \in U, v \in U, \alpha \in [0,1]$:

$$L(t, \alpha x + (1-\alpha)y, \alpha u + (1-\alpha)v) \leq \alpha L(t,x,u) + (1-\alpha)L(t,y,v),$$

and

$$\Psi(\alpha x + (1-\alpha)y) \leq \alpha\Psi(x) + (1-\alpha)\Psi(y).$$

We say that, for each $(t,\omega) \in [0,T] \times \Omega$, $L(t,\cdot,\cdot)$ is strictly convex or $\Psi(\cdot)$ is strictly convex if the first or the second inequalities, respectively, are strict. We are interested in the functional $J : \mathcal{U} \mapsto \Re$ defined by

$$J(u) = E\left[\int_0^T L(t, X_t, u_t)dt + \Psi(X_T)\right]. \tag{17}$$

The following property is then obvious.

PROPOSITION 1.1. *The functional J is convex. Furthermore, if, for each $(t,\omega) \in [0,T] \times \Omega$, $L(t,\cdot,\cdot)$ or $\Psi(\cdot)$ are strictly convex, then J is strictly convex.*

1.3. The problem. In this section we want to address the following stochastic control problem:

$$\inf_{u \in \mathcal{U}} J(u), \tag{18}$$

where $\mathcal{U}$ and J have been defined in §1.2. That is, we want to select the control $u \in \mathcal{U}$ that minimizes the criterion J.

We are imposing minimal conditions on the admissible controls u in $\mathcal{U}$. Indeed, all the versions of the stochastic maximum principle in which the control enters into the diffusion coefficient (including the ones in [1], [4]–[6], [12], [34], [38]) require at least that the controls be square-integrable. Assuming this technical condition makes the problem much easier, as we shall see in §3.

Furthermore, our objective function is more general than the one considered in Saksonov's linear problem, the linear-regulator problem, the predicted-miss problem, or the consumption-investment problem. Indeed, if L were identically equal to zero, and Ψ were linear and deterministic, then we would be in the linear case studied in [38]. On the other hand, with $V = \Re^n$,

$$L(t,x,u) = x^* M(t) x + u^* N(t) u,$$

and

$$\Psi(x) = x^* \tilde{N} x,$$

where M, N are $(\mathcal{F}_t)$-progressively measurable, $\tilde{N}$ is $\mathcal{F}_T$-measurable, and, for each $(t,\omega) \in [0,T] \times \Omega$, the $n \times n$ matrices $M(t)$ and $\tilde{N}$ are symmetric, nonnegative-definite, and the $k \times k$ matrix $N(t)$ is symmetric and positive-definite, we recover the linear-regulator problem studied, for instance, in [6], [14], and [18]. Instead, if $V = \Re^n$, L were identically equal to zero, and

$$\Psi(x) = (v \bullet x)^2 \quad \text{for some } v \in \Re^n,$$

then our objective function would be the same as the one in the predicted-miss problem studied in [3] and [18]. Finally, in the particular case of $V = (0,\infty)$,

$$L(t,x,u) = L(t,x,(\pi,c)) = -U_1(t,c); \qquad u = (\pi,c) \in \Re^m \times [0,\infty)$$

and

$$\Psi(x) = -U_2(x),$$

where $U_1(t,\cdot)$ and U_2 are strictly concave and strictly increasing functions, we recover the consumption-investment model studied in [24], [27]; we shall study that model in §2.

1.4. The adjoint processes. The purpose of the stochastic maximum principle is to find a necessary condition for the optimality of a control. Thus, let us suppose in this section that $\hat{u}$ is a candidate for optimal control for problem (18) (i.e., that it achieves the infimum there), and denote by $\hat{X}$ the trajectory of the system controlled by $\hat{u}$.

DEFINITION 1.2 (adjoint equation). *The adjoint equation is the backward stochastic differential equation*

$$dp_t = \left[L_x(t, \hat{X}_t, \hat{u}_t) - A_t^* p_t - \sum_{j=1}^{d} D_t^{(j*)} q_t^{(j)} \right] dt + q_t dW_t, \tag{19}$$

$$p_T = -\Psi_x(\hat{X}_T). \tag{20}$$

We want to find a pair of measurable, adapted processes $p : [0,T] \times \Omega \longmapsto \Re^n$, $q : [0,T] \times \Omega \longmapsto \mathcal{L}(\Re^d; \Re^n)$ that solve the adjoint equation. They will be called adjoint processes. In our notation, for each $j \in \{1, 2, \ldots, d\}$, $D_t^{(j)}$ and $q_t^{(j)}$ are $n \times n$ and $n \times 1$ matrices, respectively.

The adaptivity of the processes (p, q) is a very strong requirement. Indeed, without this requirement, it would be very easy to solve the system (19)–(20): we could take q identically equally to zero, and solve the resulting ordinary differential equation with terminal condition. Of course, that *trivial solution* would not be adapted!

Assumption 1.1. In this section, we shall need to assume that

$$E|\Psi_x(\hat{X}_T)|^2 < \infty, \tag{21}$$

and

$$E\int_0^T |L_x(t, \hat{X}_t, \hat{u}_t)|^2 dt < \infty. \tag{22}$$

set of vector-valued processes $p : [0, T] \times \Omega \longmapsto \Re^n$ that are measurable, adapted, and satisfy

$$E\int_0^T |p(t)|^2 dt < \infty. \tag{23}$$

Similarly, $M^2(0, T; \mathcal{L}(\Re^d; \Re^n))$ will denote the set of matrix-valued processes $q : [0, T] \times \Omega \longmapsto \mathcal{L}(\Re^d; \Re^n)$ that are measurable, adapted, and satisfy

$$E\int_0^T |q(t)|^2 dt < \infty. \tag{24}$$

The following existence and uniqueness result is proved in Theorem 3.1 of [33].

THEOREM 1.1. *If assumptions* (21)–(22) *hold, then there exists a unique pair*

$$(p, q) \in M^2(0, T; \Re^n) \times M^2(0, T; \mathcal{L}(\Re^d; \Re^n))$$

that solves the adjoint equation (19)–(20). *Furthermore, the process p satisfies*

$$E\left[\sup_{0 \le t \le T} |p(t)|^2\right] < \infty. \tag{25}$$

From now on, we shall call the pair (p, q) the *adjoint processes.*

1.4.1. Explicit solutions.

Now that we have a result about existence and uniqueness of the adjoint processes, we would like to compute them. Let $\Phi : [0, T] \times \Omega \mapsto \mathcal{L}(\Re^n; \Re^n)$ be the solution of the matrix stochastic integral equation

$$\Phi(t) = I + \int_0^t A(s)\Phi(s)ds + \sum_{j=1}^{d} \int_0^t D^{(j)}(s)\Phi(s)dW_s^{(j)}. \tag{26}$$

From §5.6.D of [25], we know that the stochastic integral equation (26) has a unique, strong solution. According to Exercise 5.6.22 of [25, pp. 362–363], Φ has an inverse, which satisfies

$$\Phi^{-1}(t) = I + \int_0^t \Phi^{-1}(s)\left[\sum_{j=1}^{d}(D^{(j)}(s))^2 - A(s)\right]ds - \sum_{j=1}^{d} \int_0^t \Phi^{-1}(s)D^{(j)}(s)dW_s^j. \tag{27}$$

From Corollary 2.5.12 of [30, p. 86], we also know that $\forall m \in \aleph$:

$$E\left[\sup_{0\le t\le T}|\Phi(t)|^m\right]+E\left[\sup_{0\le t\le T}|\Phi^{-1}(t)|^m\right]<\infty. \tag{28}$$

Equation (27) may also be written as

$$\begin{aligned}\Phi^*(t)^{-1}=I&+\int_0^t\left[\sum_{j=1}^d(D^{(j)}(s)^*)^2-A^*(s)\right]\Phi^*(s)^{-1}ds\\&-\sum_{j=1}^d\int_0^t D^{(j)}(s)^*\Phi^*(s)^{-1}dW_s^{(j)}.\end{aligned}\tag{29}$$

Let $\theta:[0,T]\times\Omega\longmapsto\mathcal{L}(\Re^n;\mathcal{L}(\Re^d;\Re^n))$ be any measurable, adapted process such that

$$P\left\{\sum_{j=1}^d\int_0^T|\theta_s^{(j)}|^2ds<\infty\right\}=1, \tag{30}$$

and let us denote by $Z_\theta:[0,T]\times\Omega\longmapsto\mathcal{L}(\Re^n;\Re^n)$ the solution of the matrix stochastic integral equation

$$Z_\theta(t)=I-\sum_{j=1}^d\int_0^t\theta^{(j)}(s)Z_\theta(s)dW^{(j)}(s). \tag{31}$$

From §5.6.D of [25], we know that the matrix stochastic integral equation (31) has a unique, strong solution.

Also, let γ be a given vector in $\Re^n$ and consider the adapted process $\Phi^*(t)^{-1}Z_\theta(t)\gamma + \tilde{p}(t)$, where $\tilde{p}:[0,T]\times\Omega\longmapsto\Re^n$ is the solution of the integral equation

$$\tilde{p}(t)=\int_0^t\{L_x(s,\hat{X}_s,\hat{u}_s)-A^*(s)\tilde{p}(s)\}ds. \tag{32}$$

We note that for any fixed $\omega\in\Omega$, (32) is an ordinary integral equation.

Applying the integration by parts formula (see, for instance, [25, p. 155]) to each component of the matrix $\Phi^*(t)^{-1}Z_\theta(t)$ we get from (29), (31), and (32):

$$\begin{aligned}&d(\Phi^*(t)^{-1}Z_\theta(t)\gamma+\tilde{p}(t))\\&=\Phi^*(t)^{-1}dZ_\theta(t)\gamma+d\Phi^*(t)^{-1}Z_\theta(t)\gamma+d<\Phi^*(t)^{-1},Z_\theta(t)>\gamma+d\tilde{p}(t)\\&=\Phi^*(t)^{-1}\left\{-\sum_{j=1}^d\theta^{(j)}(t)Z_\theta(t)dW^{(j)}(t)\right\}\gamma\\&\quad+\left\{\left[\sum_{j=1}^d(D^{(j)}(t)^*)^2-A^*(t)\right]\Phi^*(t)^{-1}dt-\sum_{j=1}^d D^{(j)}(t)^*\Phi^*(t)^{-1}dW^{(j)}(t)\right\}Z_\theta(t)\gamma\\&\quad+\sum_{j=1}^d D^{(j)}(t)^*\Phi^*(t)^{-1}\theta^{(j)}(t)Z_\theta(t)dt\bullet\gamma+\{L_x(t,\hat{X}_t,\hat{u}_t)-A^*(t)\tilde{p}(t)\}dt\end{aligned}$$

$$= \left\{ L_x(t,\hat{X}_t,\hat{u}_t) - A^*(t)[\Phi^*(t)^{-1}Z_\theta(t)\gamma + \tilde{p}(t)] \right.$$
$$\left. + \sum_{j=1}^{d} D^{(j)}(t)^*[D^{(j)}(t)^*\Phi^*(t)^{-1}Z_\theta(t) + \Phi^*(t)^{-1}\theta^j(t)Z_\theta(t)]\gamma \right\} dt$$
$$- \sum_{j=1}^{d}[\Phi^*(t)^{-1}\theta^{(j)}(t)Z_\theta(t) + D^{(j)}(t)^*\Phi^*(t)^{-1}Z_\theta(t)]\gamma dW^{(j)}(t).$$

From the above equation, we see that the measurable, adapted processes $p : [0,T] \times \Omega \mapsto \Re^n$ and $q : [0,T] \times \Omega \mapsto \mathcal{L}(\Re^d; \Re^n)$ defined by

$$p(t) := \Phi^*(t)^{-1}Z_\theta(t)\gamma + \tilde{p}(t) \tag{33}$$

and

$$q(t)w := -\sum_{j=1}^{d}[\Phi^*(t)^{-1}\theta^{(j)}(t) + D^{(j)}(t)^*\Phi^*(t)^{-1}]Z_\theta(t)\gamma w_j \quad \forall w \in \Re^d, \tag{34}$$

satisfy equation (19). In order that they also satisfy the terminal condition (20), we have to choose θ and γ so that we have, almost surely,

$$\Phi^*(T)^{-1}Z_\theta(T)\gamma + \tilde{p}(T) = -\Psi_x(\hat{X}_T), \tag{35}$$

or equivalently

$$Z_\theta(T)\gamma = -\Phi^*(T)[\Psi_x(\hat{X}_T) + \tilde{p}(T)]. \tag{36}$$

Taking $t = 0$ in equation (33), we see that γ has a very simple interpretation as the initial condition for the adjoint process p: $\gamma = p(0)$.

To check that such θ and γ indeed exist under appropriate conditions, let us consider the random vector Q defined by

$$Q := -\Phi^*(T)[\Psi_x(\hat{X}_T) + \tilde{p}(T)]. \tag{37}$$

If $P\{Q = 0\} = 1$, we may take $\gamma \equiv 0$ to satisfy equation (36). Thus, if

$$P\{Q = 0\} = 1, \tag{38}$$

a solution of the adjoint equation is given by equations (33)–(34). Let us now consider the nontrivial case in which $P\{Q = 0\} < 1$. According to Assumption 1.1, inequality (28), and Hölder's inequality, we have $E[|Q|] < \infty$. Thus, there exists a progressively measurable process $Y : [0,T] \times \Omega \longmapsto \mathcal{L}(\Re^d; \Re^n)$, with

$$P\left\{ \sum_{j=1}^{d} \int_0^T |Y_s^{(j)}|^2 ds < \infty \right\} = 1, \tag{39}$$

such that we have, almost surely,

$$Q(t) := E[Q|\mathcal{F}_t] = E[Q] + \sum_{j=1}^{d} \int_0^t Y_s^{(j)} dW_s^{(j)} \quad \forall\ 0 \le t \le T. \tag{40}$$

Assuming that

$$P\{(E[Q])^*Q > 0\} = 1 \quad \text{or} \quad P\{(E[Q])^*Q < 0\} = 1, \tag{41}$$

the above equation may be written as

$$Q(t) = E[Q] + \sum_{j=1}^{d} \int_0^t \tilde{Y}_s^{(j)} Q(s) dW_s^{(j)}, \tag{42}$$

where $\tilde{Y} : [0,T] \times \Omega \longmapsto \mathcal{L}(\Re^n; \mathcal{L}(\Re^d; \Re^n))$ is the progressively measurable process defined by

$$\tilde{Y}_s^{(j)} := \frac{Y_s^{(j)}\{\text{sgn}[(EQ)^*Q(s)]\}(EQ)^*}{|(EQ)^*Q(s)|} \quad \forall\ 1 \le j \le d,\ 0 \le s \le T. \tag{43}$$

Let us consider the events $A := \{\omega \in \Omega : (E[Q])^*Q(\omega) > 0\}, B := \{\omega \in \Omega : (E[Q])^*Q(\omega) < 0\}$ so that, according to assumption (41), we have either $P\{A\} = 1$ or $P\{B\} = 1$. Let us suppose that $P\{A\} = 1$. We claim that, under this assumption,

$$(E[Q])^*Q(s) = E[(E[Q])^*Q|\mathcal{F}_s] > 0 \quad \forall\, 0 \le s \le T$$

holds almost surely, so that we have

$$K = K(\omega) := \inf_{0 \le s \le T} (E[Q])^*Q(s)(\omega) > 0$$

for $P-$ a.e. $\omega \in \Omega$. To see this, let $A' := \{\omega \in A : K(\omega) = 0\}$, and suppose that $\sigma(\omega) := \inf\{s \in [0,T] : (E[Q])^*Q(s)(\omega) = 0\} \in [0,T)$. Since $\{((E[Q])^*Q(t), \mathcal{F}_t) : 0 \le t \le T\}$ is a continuous nonnegative martingale, we conclude (see, for instance, [25, Problem 1.3.29]) that $(E[Q])^*Q(\omega) = 0$ holds for $P-$ a.e. $\omega \in A'$. This contradicts $A' \subset A$, unless $P\{A'\} \equiv 0$. Thus, if $P\{A\} = 1$ then

$$|\tilde{Y}_s^{(j)}(\omega)|^2 = \frac{|Y_s^{(j)}(\omega)(E[Q])^*|^2}{|(E[Q])^*Q(s)(\omega)|^2} \le \frac{|E[Q]|^2}{K^2(\omega)} |Y_s^{(j)}(\omega)|^2 < \infty,$$

$0 \le s \le T, j \in \{1, 2, \ldots, d\}$ for $P-$ a.e. $\omega \in \Omega$. The same result (but with a different random variable K) is valid if we assume $P\{B\} = 1$. Therefore, from equations (39) and (43), we obtain

$$P\left\{\sum_{j=1}^{d} \int_0^T |\tilde{Y}_s^{(j)}|^2 ds < \infty\right\} = P\left\{\sum_{j=1}^{d} \int_0^T \frac{|Y_s^{(j)}(EQ)^*|^2}{|(EQ)^*Q(s)|^2} ds < \infty\right\} = 1. \tag{44}$$

On the other hand, equation (31) may be written as

$$Z_\theta(t)\gamma = \gamma - \sum_{j=1}^{d} \int_0^t \theta^{(j)}(s) Z_\theta(s) \gamma dW^{(j)}(s). \tag{45}$$

If we choose

$$\gamma = E[Q], \tag{46}$$

$$\theta = -\tilde{Y}, \tag{47}$$

we see from (42) and (45) that $Z_\theta(t)\gamma = Q(t)$, $\forall\, 0 \le t \le T$. From this and (40), (37), we obtain

$$Z_\theta(T)\gamma = Q(T) = Q = -\Phi^*(T)[\Psi_x(\hat{X}_T) + \tilde{p}(T)],$$

so the vector γ and the vector-valued process θ defined by equations (46)–(47) satisfy the terminal condition (35), or equivalently (36).

Therefore, we have proved the following result.

THEOREM 1.2 (explicit solution of the adjoint equation). *Under Assumption* 1.1, *if either* (38) *or* (41) *holds, then a solution of the adjoint equation is given by equations* (33)–(34).

The explicit solution of the adjoint equation (19)–(20) given by formulae (33)–(34) is one of the contributions of this paper. To apply these formulae, we must check that we have either (38) or (41). We now list some simple sufficient conditions for this to happen.

COROLLARY 1.1. *If one of the following nine possibilities*[1] *holds, then a solution of the adjoint equation is given by equations* (33)–(34):

1. $L_x \equiv 0$, $\Psi_x \equiv 0$;
2. $D \equiv 0$, $A \ge 0$, $L_x \equiv 0$, $\Psi_x > 0$;
3. $D \equiv 0$, $A \ge 0$, $L_x \equiv 0$, $\Psi_x < 0$;
4. $D \equiv 0$, $A \ge 0$, $L_x > 0$, $\Psi_x \equiv 0$;
5. $D \equiv 0$, $A \ge 0$, $L_x < 0$, $\Psi_x \equiv 0$;
6. $n = 1$, $L_x > 0$, $\Psi_x \ge 0$;
7. $n = 1$, $L_x < 0$, $\Psi_x \le 0$;
8. $n = 1$, $L_x \ge 0$, $\Psi_x > 0$;
9. $n = 1$, $L_x \le 0$, $\Psi_x < 0$.

The last condition of the above corollary will be satisfied trivially in the model of §2.

It should be noted that we cannot apply Theorem 1.1 to settle the question of uniqueness for the explicit solution (33)–(34). The problem is that we cannot guarantee that (p, q) defined by (33)–(34) is in $M^2(0,T;\Re^n) \times M^2(0,T;\mathcal{L}(\Re^d;\Re^n))$. Nevertheless, we can note the following remark.

Remark 1.2. If θ given by equation (47) is uniformly bounded in $(t,\omega) \in [0,T] \times \Omega$, then (p, q) defined by equations (33)–(34) is the unique solution of the adjoint equation in the space $M^2(0,T;\Re^n) \times M^2(0,T;\mathcal{L}(\Re^d;\Re^n))$ of Theorem 1.1.

In fact, if θ is uniformly bounded in $(t,\omega) \in [0,T] \times \Omega$ then, according to Corollary 2.5.12 of [30, p. 86], $\forall m \in \aleph$,

$$E\left[\sup_{0\le t\le T} |Z_\theta(t)|^m\right] < \infty. \tag{48}$$

This, combined with equations (28), (22), and (32), guarantees that (p, q) defined by equations (33)–(34) is in $M^2(0,T;\Re^n) \times M^2(0,T;\mathcal{L}(\Re^d;\Re^n))$; and we know from Theorem 1.1 that the solution of the adjoint equation in this space is unique.

Professor Shige Peng remarks (private communication) that it would be of interest to find general conditions guaranteeing the boundedness of the processes $Y, \tilde{Y}$, and θ (in L^∞ or L^p, $p > 2$), preferably *without* invoking Malliavin derivatives or the Clark–Ocone formula of [29].

[1] $A \ge 0$ means that every component of the matrix A is nonnegative. Similarly, if α is a vector, $\alpha > 0$ means that each one of its components is positive, and $\alpha < 0$ means that each one of its components is negative.

Remark: The one-dimensional case. We can be much more explicit if we suppose that the state of the system is one-dimensional, i.e., that $n = 1$. Then, Q is a real-valued random variable,

$$\gamma = E[Q] \in \Re,$$

and

$$\theta_s = -\tilde{Y}_s = -\frac{1}{Q_s} Y_s \in \Re^d.$$

According to §5.6.C of [25],

$$\begin{aligned} \Phi(t) = {} & \exp\left\{ \int_0^t A(s)ds \right\} \\ & \exp\left\{ \sum_{j=1}^d \int_0^t D^{(j)}(s)dW^{(j)}(s) - \frac{1}{2}\sum_{j=1}^d \int_0^t |D^{(j)}(s)|^2 ds \right\} \end{aligned} \tag{49}$$

and

$$Z_\theta(t) = \exp\left\{ -\sum_{j=1}^d \int_0^t \theta^{(j)}(s)dW^{(j)}(s) - \frac{1}{2}\sum_{j=1}^d \int_0^t |\theta^{(j)}(s)|^2 ds \right\}, \tag{50}$$

so

$$\begin{aligned} p(t) = {} & \exp\left\{ -\int_0^t A(s)ds \right\} \\ & \exp\left\{ -\sum_{j=1}^d \int_0^t D^{(j)}(s)dW^{(j)}(s) + \frac{1}{2}\sum_{j=1}^d \int_0^t |D^{(j)}(s)|^2 ds \right\} \\ & \exp\left\{ -\sum_{j=1}^d \int_0^t \theta^{(j)}(s)dW^{(j)}(s) - \frac{1}{2}\sum_{j=1}^d \int_0^t |\theta^{(j)}(s)|^2 ds \right\} \gamma \\ & + \exp\left\{ -\int_0^t A(s)ds \right\} \int_0^t \exp\left\{ \int_0^s A(v)dv \right\} L_x(s, \hat{X}_s, \hat{u}_s)ds \end{aligned} \tag{51}$$

and

$$\begin{aligned} q(t) = {} & -\Phi(t)^{-1}(\theta(t) + D(t))Z_\theta(t)\gamma \\ = {} & -\exp\left\{ -\int_0^t A(s)ds \right\} \\ & \exp\left\{ -\sum_{j=1}^d \int_0^t D^{(j)}(s)dW^{(j)}(s) + \frac{1}{2}\sum_{j=1}^d \int_0^t |D^{(j)}(s)|^2 ds \right\} \\ & \exp\left\{ -\sum_{j=1}^d \int_0^t \theta^{(j)}(s)dW^{(j)}(s) - \frac{1}{2}\sum_{j=1}^d \int_0^t |\theta^{(j)}(s)|^2 ds \right\} [\theta(t) + D(t)]\gamma. \end{aligned} \tag{52}$$

We shall apply the last two formulae to solve the consumption-investment problem in §2. In that model, $D \equiv 0$ and $L_x \equiv 0$.

Remark: A second representation. There exists yet another representation for the adjoint processes. By Itô's formula,

$$\begin{aligned} d(\Phi^*(t)p(t)) &= \Phi^*(t)dp(t) + d\Phi^*(t)p(t) + d < \Phi^*, p > (t) \\ &= \Phi^*(t)\left\{\left[L_x(t,\hat{X}_t,\hat{u}_t) - A_t^* p_t - \sum_{j=1}^{d} D_t^{(j*)} q_t^{(j)}\right] dt + q_t dW_t\right\} \\ &\quad + \left\{\Phi^*(t)A^*(t)dt + \sum_{j=1}^{d} \Phi^*(t)D^{(j*)}(t)dW_t^{(j)}\right\} p(t) \\ &\quad + \sum_{j=1}^{d} \Phi^*(t)D^{(j*)}(t)q^{(j)}(t)dt \\ &= \Phi^*(t)L_x(t,\hat{X}_t,\hat{u}_t)dt + \Phi^*(t)q(t)dW(t) \\ &\quad + \Phi^*(t)\sum_{j=1}^{d} D^{(j*)}(t)p(t)dW^{(j)}(t). \end{aligned} \tag{53}$$

Thus, the process $M : [0,T] \times \Omega \mapsto \Re^n$ defined by

$$M(t) := \Phi^*(t)p(t) - \int_0^t \Phi^*(s)L_x(s,\hat{X}_s,\hat{u}_s)ds, \quad 0 \le t \le T \tag{54}$$

is a local martingale. According to inequality (28), Theorem 1.1, Assumption 1.1, and Hölder's inequality, both terms on the right-hand side are integrable. Thus, $E[\sup_{0 \le t \le T} |M(t)|] < \infty$. This shows that M is a local martingale of class DL, hence a martingale. Thus,

$$\begin{aligned} M(t) &= \Phi^*(t)p(t) - \int_0^t \Phi^*(s)L_x(s,\hat{X}_s,\hat{u}_s)ds \\ &= E\left[\Phi^*(T)p(T) - \int_0^T \Phi^*(s)L_x(s,\hat{X}_s,\hat{u}_s)ds|\mathcal{F}_t\right] \\ &= E\left[-\Phi^*(T)\Psi_x(\hat{X}_T) - \int_0^T \Phi^*(s)L_x(s,\hat{X}_s,\hat{u}_s)ds|\mathcal{F}_t\right], \end{aligned} \tag{55}$$

or equivalently $\forall 0 \le t \le T$,

$$\Phi^*(t)p(t) = -E\left[\Phi^*(T)\Psi_x(\hat{X}_T) + \int_t^T \Phi^*(s)L_x(s,\hat{X}_s,\hat{u}_s)ds|\mathcal{F}_t\right]. \tag{56}$$

This gives a formula to compute p, but we still need a formula to compute q. Suppose that for some $\delta > 0$, we have

$$E|\Psi_x(\hat{X}_T)|^{2+\delta} < \infty \tag{57}$$

and

$$E\int_0^T |L_x(t,\hat{X}_t,\hat{u}_t)|^{2+\delta}dt < \infty. \tag{58}$$

Then equation (55), together with the additional conditions (57)–(58), inequality (28), and Hölder's and Minkowski's inequalities, implies that $E[|M(t)|^2] < \infty$ for every $0 \leq t \leq T$. Applying the theorem on the representation of Brownian, square-integrable martingales as stochastic integrals (see, for instance, [25, p. 182]), we see that there exists an adapted process $\alpha = \{(\alpha_t, \mathcal{F}_t) : 0 \leq t \leq T\}$ such that $\forall 1 \leq j \leq d$, $E[\int_0^T |\alpha_t^{(j)}|^2 dt] < \infty$, and

$$M(t) = \Phi^*(0)p(0) + \sum_{j=1}^{d} \int_0^t \alpha_s^{(j)} dW_s^{(j)}, \quad 0 \leq t \leq T. \tag{59}$$

From the uniqueness of the representation (53), we obtain

$$\Phi^*(t)q(t) + \Phi^*(t)D^*(t)p(t) = \alpha(t). \tag{60}$$

Now, we summarize our results.

THEOREM 1.3. *If the additional conditions* (57)–(58) *hold, then the adjoint processes are given, for any* $0 \leq t \leq T$, *by*

$$p(t) = -\Phi^*(t)^{-1} E\left[\Phi^*(T)\Psi_x(\hat{X}_T) + \int_t^T \Phi^*(s)L_x(s, \hat{X}_s, \hat{u}_s)ds | \mathcal{F}_t\right], \tag{61}$$

$$q(t) = \Phi^*(t)^{-1}\{\alpha(t) - \Phi^*(t)D^*(t)p(t)\}, \tag{62}$$

where $\alpha(\cdot)$ *is the integrand in the stochastic integral representation* (59).

The explicit solution of the adjoint equation given by Theorem 1.3 is due to Arkin and Saksonov [1], [38] under stronger conditions. Theorem 1.3 has two limitations: The first is that to find p we have to compute a conditional expectation, and this is not simple in applications. The second limitation is that the expression (62) for q is not explicit; it requires the determination of the integrand process α in (59).

Other references on the adjoint equation include [13], [17], and [35]. On the other hand, Detemple [10] has studied another class of stochastic equations with terminal conditions and provided some economic interpretations.

1.5. The stochastic maximum principle. From Remark 1.1 and Proposition 1.1, we know that $\mathcal{U}$ is convex and $J : \mathcal{U} \mapsto \Re$ is a convex functional defined on $\mathcal{U}$. Thus, we see that *our stochastic control problem is a particular case of the general problem of minimizing a convex functional.* It is then natural to investigate what information the theory of convex analysis can provide about problem (18); we shall do this in the remainder of this section. All the necessary results from convex analysis can be found, for instance, in Chapters I and II of [11].

DEFINITION 1.3. *The Hamiltonian is the function* $H : [0,T] \times \Re^n \times \mathcal{L}(\Re^d; \Re^n) \times V \times U \longmapsto \Re$ *defined by*

$$\begin{aligned} H(t,p,q,x,u) &= -L(t,x,u) + p \bullet f(t,x,u) + q \bullet g(t,x,u) \\ &= -L(t,x,u) + p \bullet (A_t x + B_t u + C_t) + q \bullet (D_t x + E_t u + F_t), \end{aligned} \tag{63}$$

where (p, q) *are the adjoint variables.*

We see that the backward stochastic differential equation (19)–(20) satisfied by the adjoint processes may be written as (5)–(6).

Let us consider the functionals $J_1 : \mathcal{U} \mapsto \Re$ and $J_2 : \mathcal{U} \mapsto \Re$ defined by

$$J_1(u) = E\int_0^T L(t, X_t^u, u_t)dt, \tag{64}$$

$$J_2(u) = E[\Psi(X_T^u)]. \tag{65}$$

Obviously,

$$J = J_1 + J_2.$$

Furthermore, since $L(t, \cdot, \cdot)$ and $\Psi(\cdot)$ are convex, J_1 and J_2 are both convex. On the other hand, for every $u \in \mathcal{U}$, let X^u be the trajectory of the system controlled by u, and Z^u be the solution of the linear stochastic integral equation

$$Z_t = \int_0^t (A_sZ_s + B_su_s)ds + \int_0^t (D_sZ_s + E_su_s)dW_s, \tag{66}$$

so that $u \mapsto Z^u$ is linear. We note that for every $u, v \in \mathcal{U}$,

$$Z^u - Z^v = X^u - X^v. \tag{67}$$

We need to impose the following assumption about the derivatives of L and Ψ.

Assumption 1.2. Let u and v be admissible controls, with corresponding trajectories X^u and X^v. We shall assume that there exists a random variable $\tilde{Y} : \Omega \longmapsto \Re$ and a measurable process $Y : [0,T] \times \Omega \longmapsto \Re$ such that $E|\tilde{Y}| < \infty$, $E\int_0^T |Y_t|dt < \infty$, and

$$\tilde{Y} \geq Z_T^u \bullet \Psi_x(X_T^v + \rho Z_T^u),$$
$$Y_t \geq Z_t^u \bullet L_x(t, X_t^v + \rho Z_t^u, v_t + \rho u_t) + u_t \bullet L_u(t, X_t^v + \rho Z_t^u, v_t + \rho u_t)$$

for arbitrary $u, v \in \mathcal{U}, \rho \in [0,1]$ that satisfy $X_t^v + \rho Z_t^u \in V$ and $v_t + \rho u_t \in U$.

LEMMA 1.1. *J_1 and J_2 are Gâteaux-differentiable with differentials J_1' and J_2' given by*

$$\langle J_1'(v), u\rangle = E\int_0^T \{Z_t^u \bullet L_x(t, X_t^v, v_t) + u_t \bullet L_u(t, X_t^v, v_t)\}dt \tag{68}$$

$$\langle J_2'(v), u\rangle = E[Z_T^u \bullet \Psi_x(X_T^v)]. \tag{69}$$

Hence $J = J_1 + J_2$ is Gâteaux-differentiable with differential given by

$$\begin{aligned}&\langle J'(v), u\rangle \\ &= E\left[\int_0^T \{Z_t^u \bullet L_x(t, X_t^v, v_t) + u_t \bullet L_u(t, X_t^v, v_t)\}dt + Z_T^u \bullet \Psi_x(X_T^v)\right].\end{aligned} \tag{70}$$

Proof. We start by studying the differentiability of J_2. We want to analyse the limit as $\lambda \downarrow 0$ of

$$\frac{J_2(v + \lambda u) - J_2(v)}{\lambda} = \frac{E[\Psi(X_T^v + \lambda Z_T^u)] - E[\Psi(X_T^v)]}{\lambda}.$$

Since $\Psi : \Omega \mapsto C^1(V; \Re)$ is $\mathcal{F}_T$-measurable, for every λ in the interval $(0,1]$, there exists a $\mathcal{F}_T$-measurable random variable $\theta = \theta(\omega)$ in the interval $[0,1]$ such that

$$\frac{1}{\lambda}[\Psi(X_T^v + \lambda Z_T^u) - \Psi(X_T^v)] = Z_T^u \bullet \Psi_x(X_T^v + \theta\lambda Z_T^u).$$

Thus, for any given $\lambda \in (0,1]$, there exists an $\mathcal{F}_T$-random variable θ with values in $[0,1]$ such that

$$\frac{J_2(v+\lambda u)-J_2(v)}{\lambda} = E[Z_T^u \bullet \Psi_x(X_T^v + \theta\lambda Z_T^u)].$$

Since Ψ and J_2 are convex, the left-hand sides of the above two expressions are nondecreasing functions of λ. Thus, .when $\lambda \downarrow 0$, these expressions possess limits. Since, for each $\omega \in \Omega$, $\Psi(\cdot)$ is a differentiable convex function, $\Psi_x(\cdot)$ is monotone (see [11, Prop. 1.5.5]), i.e.,

$$(x-y)\bullet(\Psi_x(x)-\Psi_x(y)) \geq 0 \quad \forall x,y \in \Re^n.$$

In particular, for any given $0 \leq \alpha < \beta \leq 1$, we can take $x = X_T^v + \beta Z_T^u, y = X_T^v + \alpha Z_T^u$ and obtain

$$Z_T^u \bullet \Psi_x(X_T^v + \beta Z_T^u) \geq Z_T^u \bullet \Psi_x(X_T^v + \alpha Z_T^u).$$

If we consider any sequence (λ_k) such that $\lambda_k \downarrow 0$ as $k \uparrow \infty$, we see then

$$\tilde{Y} \geq Z_T^u \bullet \Psi_x(X_T^v + \theta_k\lambda_k Z_T^u) \downarrow Z_T^u \bullet \Psi_x(X_T^v), \quad \text{a.s.}$$

Applying the monotone convergence theorem, we get

$$\begin{aligned}\lim_{\lambda\downarrow 0}\frac{J_2(v+\lambda u)-J_2(v)}{\lambda} &= \lim_{k\uparrow\infty}\frac{J_2(v+\lambda_k u)-J_2(v)}{\lambda_k}\\ &= \lim_{k\uparrow\infty} E[Z_T^u \bullet \Psi_x(X_T^v + \theta_k\lambda_k Z_T^u)]\\ &= E[Z_T^u \bullet \Psi_x(X_T^v)].\end{aligned}$$

Thus, J_2 is Gâteaux-differentiable with differential J_2' given by

$$\langle J_2'(v), u\rangle = E[Z_T^u \bullet \Psi_x(X_T^v)].$$

Similarly, J_1 is Gâteaux-differentiable with differential J_1' given by (68), and (70) follows. □

Let us now consider the semimartingales p and X^u given by equations (19) and (8). Applying the formula of integration by parts, we get, in conjunction with (8), (19),

$$\begin{aligned} p_t \bullet X_t^u &= p_0 \bullet X_0^u + \int_0^t p_s \bullet dX_s^u + \int_0^t X_s^u \bullet dp_s + \langle p_t, X_t^u\rangle\\ &= p_0 \bullet X_0^u + \int_0^t \Bigg\{ p_s \bullet f(s,X_s^u,u_s) + X_s^u \bullet \Bigg[L_x(s,\hat{X}_s,\hat{u}_s)\\ &\qquad -A_s^* p_s - \sum_{j=1}^d D_s^{(j*)} q_s^{(j)}\Bigg] + q_s \bullet g(s,X_s^u,u_s)\Bigg\} ds\\ &\quad + \int_0^t \{p_s \bullet g(s,X_s^u,u_s) + X_s^u \bullet q_s\} dW_s. \end{aligned} \tag{71}$$

The above equation may be written as

$$R_t^u = p_0 \bullet x + \int_0^t \{p_s \bullet C_s + q_s \bullet F_s\} ds + S_t^u, \tag{72}$$

where we denote $\forall u \in \mathcal{U}, t \in [0,T]$,

$$S_t^u := \int_0^t \{p_s \bullet g(s, X_s^u, u_s) + X_s^u \bullet q_s\} dW_s, \tag{73}$$

$$R_t^u := p_t \bullet X_t^u - \int_0^t \{X_s^u \bullet (L_x(s, \hat{X}_s, \hat{u}_s)) + p_s \bullet B_s u_s + q_s \bullet E_s u_s\} ds. \tag{74}$$

If, for every admissible control u, S^u happens to be not only a local martingale, but also a martingale, then $\forall t \in [0,T]$,

$$E[R_t^u] = E\left[p_0 \bullet x + \int_0^t \{p_s \bullet C_s + q_s \bullet F_s\} ds\right] = E[R_t^{\hat{u}}]. \tag{75}$$

But, in general, S^u is not necessarily a martingale, so we have to consider the following four cases. We shall see later that in the first case we get a necessary condition for optimality of a given control $\hat{u}$, while in the second case we get a sufficient one.

Case 1.1. $\forall u \in \mathcal{U}$,

$$E[R_T^{\hat{u}}] \leq E[R_T^u], \tag{76}$$

or equivalently,

$$\begin{aligned} &E\left[p_T \bullet \hat{X}_T - \int_0^T \{\hat{X}_s \bullet (L_x(s, \hat{X}_s, \hat{u}_s)) + p_s \bullet B_s \hat{u}_s + q_s \bullet E_s \hat{u}_s\} ds\right] \\ &\quad \leq E\left[p_T \bullet X_T^u - \int_0^T \{X_s^u \bullet (L_x(s, \hat{X}_s, \hat{u}_s)) + p_s \bullet B_s u_s + q_s \bullet E_s u_s\} ds\right]. \end{aligned} \tag{77}$$

Case 1.2. $\forall u \in \mathcal{U}$,

$$E[R_T^{\hat{u}}] \geq E[R_T^u], \tag{78}$$

or equivalently,

$$\begin{aligned} &E\left[p_T \bullet \hat{X}_T - \int_0^T \{\hat{X}_s \bullet (L_x(s, \hat{X}_s, \hat{u}_s)) + p_s \bullet B_s \hat{u}_s + q_s \bullet E_s \hat{u}_s\} ds\right] \\ &\quad \geq E\left[p_T \bullet X_T^u - \int_0^T \{X_s^u \bullet (L_x(s, \hat{X}_s, \hat{u}_s)) + p_s \bullet B_s u_s + q_s \bullet E_s u_s\} ds\right]. \end{aligned} \tag{79}$$

Case 1.3. $\forall u \in \mathcal{U}$,

$$E[R_T^{\hat{u}}] = E[R_T^u], \tag{80}$$

or equivalently,

$$\begin{aligned} &E\left[p_T \bullet \hat{X}_T - \int_0^T \{\hat{X}_s \bullet (L_x(s, \hat{X}_s, \hat{u}_s)) + p_s \bullet B_s \hat{u}_s + q_s \bullet E_s \hat{u}_s\} ds\right] \\ &\quad = E\left[p_T \bullet X_T^u - \int_0^T \{X_s^u \bullet (L_x(s, \hat{X}_s, \hat{u}_s)) + p_s \bullet B_s u_s + q_s \bullet E_s u_s\} ds\right]. \end{aligned} \tag{81}$$

Case 1.4. $\exists u, v \in \mathcal{U}$ such that

$$E[R_T^u] < E[R_T^{\hat{u}}] < E[R_T^v]. \tag{82}$$

When we study the consumption-investment problem, we shall check that inequality (79) is satisfied. On the other hand, in the case of the square-integrable controls, we shall see that equation (81) is trivially satisfied. An example of Case 1.4 is given in Appendix B of [8].

Now, we are ready to state and prove the most important result of this paper: our version of the stochastic maximum principle (SMP).

Let us consider the function $\tilde{H} : [0,T] \times \Omega \times U \longmapsto \Re$ defined by

$$\tilde{H}(t,\omega,u) := L(t,\hat{X}_t(\omega),u) - p_t(\omega) \bullet B_t(\omega)u - q_t(\omega) \bullet E_t(\omega)u, \tag{83}$$

and note that $\tilde{H}(t,\omega,\cdot)$ is convex.

PROPOSITION 1.2 (the SMP in integral form). *If Case* 1.1 *holds, then a necessary condition for a control $\hat{u}$ to be optimal for the problem*

$$\min_{u\in\mathcal{U}} E\left[\int_0^T L(t,X_t^u,u_t)dt + \Psi(X_T^u)\right] \tag{84}$$

is that $\forall u \in \mathcal{U}$,

$$E\int_0^T \{\tilde{H}_u(t,\omega,\hat{u}_t(\omega)) \bullet (u_t(\omega) - \hat{u}_t(\omega))\}dt \;\; \geq \;\; 0. \tag{85}$$

On the other hand, if Case 1.2 *holds, then inequality* (85) *is a sufficient condition for a control $\hat{u}$ to be optimal for problem* (84).

Proof. The optimal control problem consists of minimizing $J(u)$ over $u \in \mathcal{U}$, where J is a Gâteaux-differentiable convex functional with derivative given by equation (70). Hence, according to Proposition 2.2.1 of [11, pp. 36–37], a necessary and sufficient condition for $\hat{u}$ to be optimal for problem (84) is

$$\langle J'(\hat{u}), u - \hat{u}\rangle \;\; \geq \;\; 0 \quad \forall u \in \mathcal{U}.$$

Thus, according to (70) and (20), $\hat{u}$ is an optimal control if and only if $\forall u \in \mathcal{U}$,

$$\begin{aligned} & E\left[\int_0^T \{(X_t^u - \hat{X}_t) \bullet L_x(t,\hat{X}_t,\hat{u}_t) + (u_t - \hat{u}_t) \bullet L_u(t,\hat{X}_t,\hat{u}_t)\}dt \right. \\ & \qquad \left. + (X_T^u - \hat{X}_T) \bullet \Psi_x(\hat{X}_T)\right] \\ & = E\left[\int_0^T \{(X_t^u - \hat{X}_t) \bullet L_x(t,\hat{X}_t,\hat{u}_t) + (u_t - \hat{u}_t) \bullet L_u(t,\hat{X}_t,\hat{u}_t)\}dt \right. \\ & \qquad \left. + (\hat{X}_T - X_T^u) \bullet p_T\right] \\ & \geq 0. \end{aligned} \tag{86}$$

In Case 1.1 (see inequality (77)), we see that $\forall u \in \mathcal{U}$,

$$\begin{aligned}
&E\int_0^T \{L_u(t,\hat{X}_t,\hat{u}_t)\bullet(u_t-\hat{u}_t)+p_t\bullet B_t(\hat{u}_t-u_t)+q_t\bullet E_t(\hat{u}_t-u_t)\}dt\\
&=E\left[\int_0^T \{L_u(t,\hat{X}_t,\hat{u}_t)\bullet(u_t-\hat{u}_t)+L_x(t,\hat{X}_t,\hat{u}_t)\bullet(X_t^u-\hat{X}_t)\}dt\right.\\
&\qquad+\int_0^T\{p_t\bullet B_t\hat{u}_t+\hat{X}_t\bullet L_x(t,\hat{X}_t,\hat{u}_t)+q_t\bullet E_t\hat{u}_t\}dt\\
&\qquad\left.-\int_0^T\{p_t\bullet B_tu_t+X_t^u\bullet L_x(t,\hat{X}_t,\hat{u}_t)+q_t\bullet E_tu_t\}dt\right]\\
&\geq E\left[\int_0^T\{(X_t^u-\hat{X}_t)\bullet L_x(t,\hat{X}_t,\hat{u}_t)+(u_t-\hat{u}_t)\bullet L_u(t,\hat{X}_t,\hat{u}_t)\}dt\right.\\
&\qquad\left.+(\hat{X}_T-X_T^u)\bullet p_T\right].
\end{aligned}$$

Thus, in Case 1.1 and in conjunction with (86), a necessary condition for a control $\hat{u}$ to be optimal is that $\forall u \in \mathcal{U}$,

$$(87)\quad E\int_0^T\{L_u(t,\hat{X}_t,\hat{u}_t)\bullet(u_t-\hat{u}_t)+p_t\bullet B_t(\hat{u}_t-u_t)+q_t\bullet E_t(\hat{u}_t-u_t)\}dt\geq 0,$$

which is equivalent to (85).

On the other hand, in Case 1.2, $\forall u \in \mathcal{U}$,

$$\begin{aligned}
&E\int_0^T \{L_u(t,\hat{X}_t,\hat{u}_t)\bullet(u_t-\hat{u}_t)+p_t\bullet B_t(\hat{u}_t-u_t)+q_t\bullet E_t(\hat{u}_t-u_t)\}dt\\
&=E\left[\int_0^T \{L_u(t,\hat{X}_t,\hat{u}_t)\bullet(u_t-\hat{u}_t)+L_x(t,\hat{X}_t,\hat{u}_t)\bullet(X_t^u-\hat{X}_t)\}dt\right.\\
&\qquad+\int_0^T\{p_t\bullet B_t\hat{u}_t+\hat{X}_t\bullet L_x(t,\hat{X}_t,\hat{u}_t)+q_t\bullet E_t\hat{u}_t\}dt\\
&\qquad\left.-\int_0^T\{p_t\bullet B_tu_t+X_t^u\bullet L_x(t,\hat{X}_t,\hat{u}_t)+q_t\bullet E_tu_t\}dt\right]\\
&\leq E\left[\int_0^T\{(X_t^u-\hat{X}_t)\bullet L_x(t,\hat{X}_t,\hat{u}_t)+(u_t-\hat{u}_t)\bullet L_u(t,\hat{X}_t,\hat{u}_t)\}dt\right.\\
&\qquad\left.+(\hat{X}_T-X_T^u)\bullet p_T\right].
\end{aligned}$$

Thus, in Case 1.2, a sufficient condition for a control $\hat{u}$ to be optimal is that (87), or equivalently (85), holds for all $u \in \mathcal{U}$. □

The meaning of inequality (85), or equivalently (87), is very *intriguing*. It is a necessary condition for optimality in Case 1.1, and a sufficient condition for optimality in Case 1.2. Now, we want to rewrite this inequality in a suitable form for applications. This is the objective of the next two theorems.

We should note that for a control $\tilde{v} : [0,T] \times \Omega \longmapsto U$, the statements

$$\min_{u \in U} \tilde{H}(t,\omega,u) = \tilde{H}(t,\omega,\tilde{v}_t(\omega)) \tag{88}$$

and

$$\max_{u \in U} H(t,p_t(\omega),q_t(\omega),\hat{X}_t(\omega),u) = H(t,p_t(\omega),q_t(\omega),\hat{X}_t(\omega),\tilde{v}_t(\omega)) \tag{89}$$

are equivalent (recall the notation of (63), (83)). Furthermore, the theorem of measurable selection (see [39, Thm. 3, p. 220]) guarantees that such $\tilde{v}$ is adapted; nevertheless, it does not guarantee that $X^{\tilde{v}} \in V$.

THEOREM 1.4 (the SMP with state-constraints). *Let us assume that $L(t,x,\cdot) : U \mapsto \Re$ is strictly convex, $\forall (t,\omega,x) \in [0,T] \times \Omega \times V$, and there exists a control $\tilde{v}$ such that*

$$\max_{u \in U} H(t,p_t,q_t,\hat{X}_t,u) = H(t,p_t,q_t,\hat{X}_t,\tilde{v}_t), \qquad X_t^{\tilde{v}} \in V, \tag{90}$$

$Leb \otimes P-$ a.e. on $[0,T] \times \Omega$. If the admissible control $\hat{u}$ satisfies Case 1.1, *then a necessary condition for $\hat{u}$ to be optimal for problem* (84) *is*

$$\hat{u} = \tilde{v}, \quad Leb \otimes P - a.e. \ on \ [0,T] \times \Omega. \tag{91}$$

On the other hand, if the admissible control $\hat{u}$ satisfies Case 1.2 *instead of Case* 1.1, *then equation* (91) *is a sufficient condition for the optimality of $\hat{u}$.*

Proof. Let us first consider the case in which $\hat{u}$ is optimal and satisfies Case 1.1. According to Proposition 1.2, we need to prove that inequality (85) implies that, with $\tilde{v}$ as in the hypothesis of this Theorem, we have (91). Since $\tilde{H}(t,\omega,\cdot)$ is convex and (88) is equivalent to (89), Proposition 2.2.1 of [11, pp. 36–37] shows that $\tilde{v}$ of (90) satisfies

$$\tilde{H}_u(t,\omega,u) \bullet (u - \tilde{v}_t(\omega)) \geq 0 \quad \forall u \in U \tag{92}$$

for $Leb \otimes P-$a.e. $(t,\omega) \in [0,T] \times \Omega$ (with equality if and only if $u = \tilde{v}_t(\omega)$, thanks to the strict convexity of $\tilde{H}(t,\omega,\cdot)$). To prove (91), consider the product set

$$\begin{aligned} A &:= \{(t,\omega) \in [0,T] \times \Omega : \hat{u}_t(\omega) \neq \tilde{v}_t(\omega)\} \\ &= \{(t,\omega) \in [0,T] \times \Omega : \tilde{H}_u(t,\omega,\hat{u}_t(\omega)) \bullet (\hat{u}_t(\omega) - \tilde{v}_t(\omega)) > 0\}, \end{aligned}$$

and suppose that $(Leb \otimes P)\{A\} > 0$. Then

$$E \int_0^T \{\tilde{H}_u(t,\omega,\hat{u}_t(\omega)) \bullet (\tilde{v}_t(\omega) - \hat{u}_t(\omega))\} dt \ < \ 0,$$

contradicting inequality (85). Thus, $(Leb \otimes P)\{A\} = 0$, or equivalently (91), holds.

Now, we consider Case 1.2, and assume that $\hat{u}$ satisfies equation (91). According to Proposition 1.2, we need to prove that inequality (85) holds. If $\hat{u}$ satisfies (91), then for $Leb \otimes P-$ a.e. $(t,\omega) \in [0,T] \times \Omega$,

$$\max_{u \in U} H(t,p_t,q_t,\hat{X}_t,u) = H(t,p_t,q_t,\hat{X}_t,\hat{u}_t)$$

or, equivalently,

$$\min_{u \in U} \tilde{H}(t,\omega,u) = \tilde{H}(t,\omega,\hat{u}_t(\omega)).$$

According to Proposition 2.2.1 of [11, pp. 36–37],

$$\tilde{H}_u(t,\omega,\hat{u}_t(\omega)) \bullet (u - \hat{u}_t(\omega)) \geq 0 \quad \forall u \in U, \tag{93}$$

for $Leb \otimes P-$ a.e. $(t,\omega) \in [0,T] \times \Omega$. Therefore, inequality (85) is valid, and from the second part of Proposition 1.2 the control $\hat{u}$ is optimal for problem (84). □

The admissible controls that maximize the Hamiltonian (i.e., satisfy equation (91)) are called *extremal controls.* Thus, the first part of Theorem 1.4 states that, under Case 1.1, every optimal control is extremal; and the second part states that, under Case 1.2, every extremal control is optimal. We shall apply the second part in the next section to solve the consumption-investment problem.

COROLLARY 1.2 (Case 1.3 and state-constraints). *If Case* 1.3 *holds and, for each* $(t,\omega,x) \in [0,T] \times \Omega \times V$, *the function* $L(t,x,\cdot)$ *is strictly convex, then a necessary and sufficient condition for a control* $\hat{u}$ *to be optimal for problem* (84) *is that*

$$\hat{u} = \tilde{v}, \qquad Leb \otimes P - a.e. \text{ on } [0,T] \times \Omega, \tag{94}$$

where $\tilde{v}$ *is a control that satisfies*

$$\max_{u \in U} H(t,p_t,q_t,\hat{X}_t,u) = H(t,p_t,q_t,\hat{X}_t,\tilde{v}_t), \quad X_t^{\tilde{v}} \in V, \tag{95}$$

$Leb \otimes P-$ *a.e. on* $[0,T] \times \Omega$.

In the next theorem we shall see that, in the case of unconstrained state, it is not necessary to assume *strict* convexity of $L(t,x,\cdot)$.

THEOREM 1.5 (the SMP without state-constraints). *Let us suppose that* $V = \Re^n$. *If Case* 1.1 *holds, then a necessary condition for a control* $\hat{u}$ *to be optimal for problem* (84) *is that,* $Leb \otimes P-$ *a.e. on* $[0,T] \times \Omega$,

$$\max_{u \in U} H(t,p_t,q_t,\hat{X}_t,u) = H(t,p_t,q_t,\hat{X}_t,\hat{u}_t). \tag{96}$$

On the other hand, if Case 1.2 *holds, then equation* (96) *is a sufficient condition for a control* $\hat{u}$ *to be optimal for problem* (84).

Proof. Let us first consider Case 1.1, and assume that $\hat{u}$ is an optimal control for problem (84). Again as before, in order to prove (96), it suffices to show that

$$\tilde{H}_u(t,\omega,\hat{u}_t(\omega)) \bullet (u - \hat{u}_t(\omega)) \geq 0 \quad \forall u \in U \tag{97}$$

for $Leb \otimes P-$ a.e. $(t,\omega) \in [0,T] \times \Omega$. To do this, define for any given $v \in U$:

$$B^v := \{(t,\omega) \in [0,T] \times \Omega : \tilde{H}_u(t,\omega,\hat{u}_t(\omega)) \bullet (v - \hat{u}_t(\omega)) < 0\}.$$

Obviously, for each $t \in [0,T]$, $B_t^v \in \mathcal{F}_t$. Let us consider the control $\check{v} : [0,T] \times \Omega \longmapsto U$ in $\mathcal{U}$, defined by

$$\check{v}(t,\omega) := \begin{cases} v & \text{if } (t,\omega) \in B^v, \\ \hat{u}(t,\omega) & \text{otherwise.} \end{cases}$$

Then $\check{v}$ is adapted and we have

$$E \int_0^T \{\tilde{H}_u(t,\omega,\hat{u}_t(\omega)) \bullet (\check{v}_t(\omega) - \hat{u}_t(\omega))\} dt \; < \; 0,$$

contradicting (85), unless $(Leb \otimes \Pr)\{B^v\} = 0$ for every $v \in U$. Thus, for any given $v \in U$:

$$\tilde{H}_u(t, \omega, \hat{u}_t(\omega)) \bullet (v - \hat{u}_t(\omega)) \geq 0$$

for $Leb \otimes P-$ a.e. $(t, \omega) \in [0,T] \times \Omega$. The stronger result (97) follows from the separability of U, and (96) then follows from Proposition 2.2.1 of [11, pp. 36–37].

Now, we consider Case 1.2. One shows exactly as before that (96) implies (93), and thus the optimality of $\hat{u}$. □

COROLLARY 1.3 (Case 1.3 and no state-constraints). *If Case* 1.3 *holds, and* $V = \Re^n$, *then a necessary and sufficient condition for a control* $\hat{u}$ *to be a solution for problem* (84) *is that*

$$(98) \quad \max_{u \in U} H(t, p_t, q_t, \hat{X}_t, u) = H(t, p_t, q_t, \hat{X}_t, \hat{u}_t), \quad Leb \otimes P - a.e. \text{ on } [0,T] \times \Omega.$$

Our methodology of applying the theory of convex analysis to study the stochastic control problem of this section can provide additional results. To obtain them it is necessary to notice that the problem

$$(99) \qquad \gamma = \min_{u \in \mathcal{U}} J(u)$$

is equivalent to

$$(100) \qquad \gamma = \min_{u \in \mathcal{V}} \tilde{J}(u),$$

where $\mathcal{V} = \{v : [0,T] \times \Omega \longmapsto \Re^k\}$ and $\tilde{J} : \mathcal{V} \longmapsto \bar{\Re}$ is the function defined by

$$\tilde{J}(v) = \begin{cases} J(v) & \text{if } v \in \mathcal{U}, \\ \infty & \text{otherwise.} \end{cases}$$

Obviously, $\tilde{J}$ is also a convex function.

LEMMA 1.2. *The set of optimal controls for problem* (99) *is convex.*

Proof. The set of optimal controls for problem (99) is equal to the set of optimal controls for problem (100). But this latter set of optimal controls is $\{v \in \mathcal{V} : \tilde{J}(v) \leq \gamma\}$, which is convex. This proves the lemma. □

THEOREM 1.6 (uniqueness). *If, for each* $(t, \omega) \in [0,T] \times \Omega$, $L(t, \cdot, \cdot)$ *or* $\Psi(\cdot)$ *are strictly convex, then problem* (99) *has at most one optimal control.*

Proof. Let $u^1 \neq u^2$ be two optimal controls for problem (99). According to Lemma 1.2, $\frac{1}{2}u^1 + \frac{1}{2}u^2$ is also an optimal control.

On the other hand, according to Proposition 1.1, J is strictly convex. Hence, $J(\frac{1}{2}u^1 + \frac{1}{2}u^2) < \frac{1}{2}J(u^1) + \frac{1}{2}J(u^2) = \gamma$, a contradiction. Therefore, $u^1 = u^2$. □

2. The consumption-investment problem.

2.1. The financial market model. We consider a financial market in which $m + 1$ securities (financial assets) are traded continuously. One of them is a pure discount *bond*, with price $P_0(t)$ at time t governed by the equation

$$(101) \qquad dP_0(t) = P_0(t) r(t) dt.$$

There are also m risky assets called *stocks* with prices-per-share $P_i(t)$ at time t governed by the linear stochastic differential equations

$$(102) \qquad dP_i(t) = P_i(t) \left[b_i(t)dt + \sum_{j=1}^{d} \sigma_{ij}(t) dW_j(t) \right], \quad i \in \{1, 2, \ldots, m\}.$$

These equations are driven by a d-dimensional Brownian motion $W = (W_1, \dots, W_d)^*$, whose components model the d independent sources of uncertainty that influence this market.

The probabilistic setting is as follows: the Brownian motion W is defined on the complete probability space $(\Omega, \mathcal{F}, P)$, and we denote by $(\mathcal{F}_t)$ the P-augmentation of the natural filtration $(\mathcal{F}_t^W)$. The coefficients of the model (i.e., the interest rate r, the appreciation rate vector $b = (b_i)_{m\times 1}$, and the volatility matrix $\sigma = (\sigma_{ij})_{m\times d}$) are random processes, progressively measurable with respect to $(\mathcal{F}_t)$ and bounded uniformly in $(t,\omega) \in [0,\infty) \times \Omega$. We suppose that $d \geq m$, and that σ has full row rank. It is also assumed that r and the relative risk process $\hat{\theta} : [0,\infty) \times \Omega \longmapsto \Re^d$ defined by

$$\hat{\theta}(t) = \sigma^*(t)(\sigma(t)\sigma^*(t))^{-1}[b(t) - r(t)1] \tag{103}$$

are bounded.

For any measurable adapted process $\theta : [0,\infty) \times \Omega \longmapsto \Re^d$, which is uniformly bounded in $(t,\omega) \in [0,T] \times \Omega$, we define the exponential martingale Z_θ by $\forall t \in [0,\infty)$,

$$Z_\theta(t) := \exp\left\{ -\sum_{j=1}^{d} \int_0^t \theta^{(j)}(s)dW^{(j)}(s) - \frac{1}{2}\sum_{j=1}^{d}\int_0^t |\theta^{(j)}(s)|^2 ds \right\}. \tag{104}$$

This is the solution of the equation

$$Z_\theta(t) = 1 - \int_0^t Z_\theta(s)\theta^*(s)dW(s). \tag{105}$$

We also define the processes β and ζ_θ by

$$\beta(t) := \exp\left\{ -\int_0^t r(s)ds \right\}, \quad \zeta_\theta(t) := \beta(t)Z_\theta(t). \tag{106}$$

All economic activity is supposed to take place on the finite time-horizon $[0,T]$. For a small investor, a portfolio rule π is a process whose components π_i represent the amount of money invested in the corresponding stock $i \in \{1,2,\dots,m\}$.

NOTATION 2.1. *We denote by $\mathcal{P}$ the set of all processes $\pi : [0,T] \times \Omega \mapsto \Re^m$ that are progressively measurable with respect to $(\mathcal{F}_t)$ and satisfy*

$$P\left\{ \int_0^T |\pi(t)|^2 dt < \infty \right\} = 1. \tag{107}$$

The elements of $\mathcal{P}$ are called portfolio processes.

The consumption rate rule c is the rate at which the small investor withdraws funds for consumption.

NOTATION 2.2. *We denote by $\mathcal{C}$ the set of all processes $c : [0,T] \times \Omega \mapsto [0,\infty)$ that are progressively measurable with respect to $(\mathcal{F}_t)$ and satisfy*

$$P\left\{ \int_0^T c(t)dt < \infty \right\} = 1. \tag{108}$$

The elements of $\mathcal{C}$ are called consumption rate processes.

The wealth process $X = X^{x,\pi,c}$ corresponding to initial capital $x > 0$, portfolio rule π, and consumption rate c then satisfies the equation

$$dX(t) = [r(t)X(t) - c(t)]dt + \pi^*(t)[b(t) - r(t)\mathbf{1}]dt + \pi^*(t)\sigma(t)dW(t) \tag{109}$$

with initial condition

$$X(0) = x. \tag{110}$$

From now on, we are going to consider only the following portfolios and consumption rate processes.

DEFINITION 2.1 (admissibility). *A pair $(\pi, c) \in \mathcal{P} \times \mathcal{C}$ of portfolio and consumption rate process is called admissible for the initial capital $x > 0$ if the corresponding wealth process X given by equations* (109)–(110) *satisfies*

$$P\{X(t) \geq 0 \;\; \forall 0 \leq t \leq T\} = 1. \tag{111}$$

We denote by $\mathcal{A}(x)$ the class of such pairs.

DEFINITION 2.2 (utility function). *A utility function is a function $U \in C^1((0,\infty); \Re)$ that is strictly increasing, strictly concave, and has a derivative $U' : (0,\infty) \mapsto (0,\infty)$ that satisfies $\lim_{c\to\infty} U'(c) = 0$. We shall denote by I the inverse of the strictly decreasing function U'.*

From now on, we are going to consider two fixed utility functions: $U_1(t,\cdot)$ and U_2. That is, for every $t \in [0,T]$, $U_1(t,\cdot)$ is a utility function, and $U_1'(t,\cdot)$ has inverse $I_1(t,\cdot)$.

PROBLEM 2.1. *The optimization problem faced by the small investor is to find a pair $(\pi, c) \in \mathcal{A}(x)$ which achieves the maximum in*

$$\mathcal{V}(x) := \max_{(\pi,c)\in\mathcal{A}(x)} E\left[\int_0^T U_1(t, c(t))dt + U_2(X(T))\right]. \tag{112}$$

The function $\mathcal{V} : (0,\infty) \mapsto \Re$ is called the value function of problem (112).

2.2. Relation between the stochastic maximum principle and the consumption-investment problem. In the notation of §1, let us take $u \equiv (\pi, c)$, $U \equiv \Re^m \times [0,\infty)$, $V \equiv [0,\infty)$, $\mathcal{U} \equiv \mathcal{A}(x)$, $f(t,x,u) \equiv r(t)x - c + \pi^*[b(t) - r(t)\mathbf{1}]$, $g(t,x,u) \equiv \pi^*\sigma(t)$, $L(t,x,u) \equiv -U_1(t,c)$, and $\Psi(x) \equiv -U_2(x)$; we see then that all the assumptions made in §1.2 hold for the model described in §2.1. To apply the general theory of §1, it only remains to check that Assumption 1.2 holds in the financial market model. But since both $U_1(t,\cdot)$ and U_2 are utility functions, $U_1'(t,\cdot)$ and U_2' are both positive and decreasing; in particular, Assumption 1.2 holds for the model described in §2.1. Therefore, *we are in a position to apply the stochastic maximum principle developed in the previous section to solve the consumption-investment problem.*

The Hamiltonian for this problem is

$$H(t,p,q,x,(\pi,c)) = U_1(t,c) + p(r(t)x - c + \pi^*[b(t) - r(t)\mathbf{1}]) + q^*\sigma^*(t)\pi, \tag{113}$$

and the adjoint equation takes the form

$$dp(t) = -r(t)p(t)dt + q(t)dW(t), \tag{114}$$

$$p(T) = U_2'(\hat{X}(T)) \tag{115}$$

for a suitable pair of processes (p, q).

Now we want to solve the adjoint equation. But in order to apply Theorem 1.2, we have to check first that

$$E[\{U_2'(\hat{X}(T))\}^2] < \infty. \tag{116}$$

Since we do not have enough information to check this condition right now, we shall just assume it. We shall justify our approach in the comment that follows equation (126). According to equations (51)–(52), an adapted solution to equation (114) is given by the processes

$$p(t) = \gamma\zeta_\theta(t), \tag{117}$$

$$q(t) = -\gamma\zeta_\theta(t)\theta(t) \tag{118}$$

in the notation of (106), where $\gamma = p(0)$. To satisfy the terminal condition (115), we need to find θ and γ such that

$$\gamma\zeta_\theta(T) = U_2'(\hat{X}(T)). \tag{119}$$

Since U_2 is strictly increasing, the right-hand side is positive. Hence, we require that γ be positive. From equation (119) we obtain

$$\hat{X}(T) = I_2(\gamma\zeta_\theta(T)). \tag{120}$$

From equations (117)–(118), and the fact that γ is a positive constant, condition (79) takes the form

(121)
$$\begin{aligned}E&\left[\zeta_\theta(T)X(T) + \int_0^T \{\zeta_\theta(t)(c(t) - \pi^*(t)[b(t) - r(t)1] + \pi^*(t)\sigma(t)\theta(t))\}dt\right]\\ &\leq E\left[\zeta_\theta(T)\hat{X}(T) + \int_0^T \{\zeta_\theta(t)(\hat{c}(t) - \hat{\pi}^*(t)[b(t) - r(t)1] + \hat{\pi}^*(t)\sigma(t)\theta(t))\}dt\right].\end{aligned}$$

The second part of Theorem 1.4 then leads to the following theorem.

THEOREM 2.1. *Suppose that* $(\hat{\pi}, \hat{c}) \in \mathcal{A}(x)$ *satisfies inequality* (121) *for every* $(\pi, c) \in \mathcal{A}(x)$, *and*

$$(p(t), q(t)) = (\gamma\zeta_\theta(t), -\gamma\zeta_\theta(t)\theta(t))$$

is a solution of the adjoint equation (114)–(115). *If*

(122)
$$\begin{aligned}&\max_{\pi\in\Re^m, c\geq 0}\{U_1(t, c) + \gamma\zeta_\theta(t)(r(t)\hat{X}(t) - c + \pi^*[b(t) - r(t)1]) - \gamma\zeta_\theta(t)\pi^*\sigma(t)\theta(t)\}\\ &\quad = U_1(t, \hat{c}(t)) + \gamma\zeta_\theta(t)(r(t)\hat{X}(t) - \hat{c}(t) + \hat{\pi}^*(t)[b(t) - r(t)1]) - \gamma\zeta_\theta(t)\hat{\pi}^*(t)\sigma(t)\theta(t),\end{aligned}$$

Leb$\otimes$*P*– *a.e., then* $(\hat{\pi}, \hat{c})$ *is optimal for problem* (112). *Here,* $\hat{X}$ *is the wealth process associated with the portfolio process* $\hat{\pi}$ *and consumption rate process* $\hat{c}$.

2.3. The candidates for optimal control. In this subsection, we shall guess the form of the optimal control from Theorem 2.1. The idea is first to find the controls that satisfy condition (122), the so-called *extremal controls*, and then those which also satisfy condition (121).

To find the controls that satisfy condition (122), we differentiate that expression, obtaining

$$U_1'(t,\hat{c}(t)) - \gamma\zeta_\theta(t) = 0, \tag{123}$$

$$\gamma\zeta_\theta(t)[b(t)-r(t)1] - \gamma\zeta_\theta(t)\sigma(t)\theta(t) = 0. \tag{124}$$

From equation (123) we get

$$\hat{c}(t) = I_1(t,\gamma\zeta_\theta(t)), \tag{125}$$

and from equation (124) we get

$$\theta(t) = \sigma^*(t)(\sigma(t)\sigma^*(t))^{-1}[b(t)-r(t)1] =: \hat{\theta}(t). \tag{126}$$

This completes the characterization of the extremal controls; it also shows, in conjunction with (119) and the boundedness of $\hat{\theta}$, that they satisfy (116).

Remark 2.1. The controls that satisfy condition (122) are those controls $(\hat{\pi},\hat{c}) \in \mathcal{A}(x)$ that satisfy equations (120) and (125), with $\theta = \hat{\theta}$ as in (126) and with $\gamma > 0$.

The controls which, in addition, satisfy condition (121) are those controls $(\hat{\pi},\hat{c})$ such that $\forall(\pi,c) \in \mathcal{A}(x)$,

(127)
$$E\left[\zeta(T)X(T) + \int_0^T \zeta(t)c(t)dt\right] \le E\left[\zeta(T)\hat{X}(T) + \int_0^T \zeta(t)\hat{c}(t)dt\right]$$
$$= E\left[\zeta(T)I_2(\gamma\zeta(T)) + \int_0^T \zeta(t)I_1(t,\gamma\zeta(t))dt\right].$$

Here, we denote $\zeta = \zeta_{\hat{\theta}}$, where $\hat{\theta}$ is given by equation (126).

We have easily found the controls that satisfy condition (122), but it remains to find more explicitly the controls that, in addition, satisfy (121), or equivalently (127). To that end, it is convenient to note the following.

LEMMA 2.1. *For every $(\pi,c) \in \mathcal{A}(x)$, the process N defined by*

$$N(t) = \zeta(t)X(t) + \int_0^t \zeta(s)c(s)ds, \quad 0 \le t \le T, \tag{128}$$

is a continuous, nonnegative local martingale, hence a supermartingale. In particular,

$$E\left[\zeta(T)X(T) + \int_0^T \zeta(s)c(s)ds\right] \le x. \tag{129}$$

Proof. According to equations (72)–(74), and (117)–(118),

$$\gamma\zeta(t)X(t) = \gamma x + \int_0^t \{\gamma\zeta(s)(-c(s) + \pi^*(s)[b(s)-r(s)1])$$
$$-\gamma\zeta(s)\theta^*(s)\sigma^*(s)\pi(s)\}ds + S_t^u.$$

From equation (124), we obtain $\zeta(t)X(t) = x - \int_0^t \zeta(s)c(s)ds + \gamma^{-1}S_t^u$, where S^u is a local martingale. □

In the remainder of this subsection, we are going to apply Remark 2.1 and Lemma 2.1 to select (heuristically) a candidate for optimal control, and in the next subsection we are going to prove that our selected candidate is indeed optimal.

From Remark 2.1 and Lemma 2.1, we *conjecture* that the optimal controls $(\hat{\pi}, \hat{c})$ satisfy

$$E\left[\zeta(T)\hat{X}(T) + \int_0^T \zeta(s)\hat{c}(s)ds\right] = x. \tag{130}$$

Let us now assume that

$$E\left[\zeta(T)I_2(y\zeta(T)) + \int_0^T \zeta(s)I_1(s, y\zeta(s))ds\right] < \infty \quad \forall y \in (0,\infty), \tag{131}$$

and define the function $\mathcal{X} : (0,\infty) \longmapsto (0,\infty)$ by

$$\mathcal{X}(y) := E\left[\zeta(T)I_2(y\zeta(T)) + \int_0^T \zeta(s)I_1(s, y\zeta(s))ds\right]. \tag{132}$$

Since $\mathcal{X}$ is strictly decreasing and surjective, it has an inverse $\mathcal{Y} \equiv \mathcal{X}^{-1} : (0,\infty) \longmapsto (0,\infty)$ which is also strictly decreasing. If conjecture (130) is valid, then

$$\hat{\gamma} = \mathcal{Y}(x), \tag{133}$$

and the process $\hat{N}$ defined by

$$\hat{N}(t) := \zeta(t)\hat{X}(t) + \int_0^t \zeta(s)\hat{c}(s)ds$$

will be not only a supermartingale, but also a martingale. In fact, $E[\hat{N}(0)] = x = E[\hat{N}(T)]$. Thus, $\hat{N}(t) = E[\hat{N}(T)|\mathcal{F}_t]$, or equivalently $\forall 0 \le t \le T$,

$$\begin{aligned} \zeta(t)\hat{X}(t) + \textstyle\int_0^t \zeta(s)\hat{c}(s)ds &= E\left[\zeta(T)\hat{X}(T) + \textstyle\int_0^T \zeta(s)\hat{c}(s)ds|\mathcal{F}_t\right] \\ &= E\left[\zeta(T)I_2(\hat{\gamma}\zeta(T)) + \textstyle\int_0^T \zeta(s)I_1(s, \hat{\gamma}\zeta(s))ds|\mathcal{F}_t\right]. \end{aligned} \tag{134}$$

In addition, according to the representation theorem of Brownian martingales as stochastic integrals (see, for instance, [25, §3.4.D]), there exists a progressively measurable process $\psi : [0,T] \times \Omega \mapsto \Re^d$ with

$$P\left\{\int_0^T |\psi(s)|^2 ds < \infty\right\} = 1 \tag{135}$$

and

$$E\left[\zeta(T)I_2(\hat{\gamma}\zeta(T)) + \int_0^T \zeta(s)I_1(s, \hat{\gamma}\zeta(s))ds|\mathcal{F}_t\right] = x + \int_0^t \psi^*(s)dW_s. \tag{136}$$

On the other hand, according to equations (72)–(74),

$$\zeta(t)\hat{X}(t) + \int_0^t \zeta(s)\hat{c}(s)ds = x + \int_0^t \zeta(s)[\sigma^*(s)\hat{\pi}(s) - \hat{X}(s)\hat{\theta}(s)]^* dW(s). \tag{137}$$

Thus, comparing equations (134)–(136) with (137), we see that if conjecture (130) is valid, then the process $\hat{\pi}$ will satisfy

$$\zeta(s)[\sigma^*(s)\hat{\pi}(s) - \hat{X}(s)\hat{\theta}(s)] = \psi(s).$$

2.4. The complete market. In the complete market case ($m = d$), $\sigma(t)$ is a square matrix with inverse $\sigma^{-1}(t)$. This case has the distinctive feature that *every contingent claim is attainable* (see, e.g., [24] or [25, pp. 376–378]).

As suggested heuristically in the previous subsection, let us define the processes $\hat{\pi} : [0,T] \times \Omega \mapsto \Re^m$ and $\hat{c} : [0,T] \times \Omega \mapsto [0,\infty)$ by

$$\hat{\pi}(t) := (\sigma^*(t))^{-1} \left\{ \frac{\psi(t)}{\zeta(t)} + \hat{X}(t)\hat{\theta}(t) \right\}, \tag{138}$$

$$\hat{c}(t) := I_1(t, \hat{\gamma}\zeta(t)), \tag{139}$$

where $\hat{\gamma}$ is defined by (133), ψ is the progressively measurable process defined by (135)–(136), $\hat{\theta}$ is defined by (126), and $\zeta = \zeta_{\hat{\theta}}$.

We see that $\hat{\pi}$ and $\hat{c}$ are progressively measurable and satisfy (107)–(108). Furthermore, according to equations (72)–(74), (138), and (136), the control $(\hat{\pi}, \hat{c})$ determines a trajectory of the system $\hat{X}$ which satisfies

$$\begin{aligned} \zeta(t)\hat{X}(t) + \int_0^t \zeta(s)\hat{c}(s)ds &= x + \int_0^t \zeta(s)[\sigma^*(s)\hat{\pi}(s) - \hat{X}(s)\hat{\theta}(s)]^* dW(s) \\ &= x + \int_0^t \psi^*(s)dW(s) \\ &= E\left[\zeta(T)I_2(\hat{\gamma}\zeta(T)) + \int_0^T \zeta(s)I_1(s, \hat{\gamma}\zeta(s))ds | \mathcal{F}_t\right], \end{aligned} \tag{140}$$

so $\forall t \in [0,T]$,

$$\hat{X}(t) = \frac{1}{\zeta(t)} E\left[\zeta(T)I_2(\mathcal{Y}(x)\zeta(T)) + \int_t^T \zeta(s)I_1(s, \mathcal{Y}(x)\zeta(s))ds | \mathcal{F}_t\right]. \tag{141}$$

Thus, the wealth process determined by $(\hat{\pi}, \hat{c})$ of (138)–(139) never takes negative values, which shows that $(\hat{\pi}, \hat{c})$ is indeed admissible (that is, $(\hat{\pi}, \hat{c}) \in \mathcal{A}(x)$). Furthermore, from equations (140), (132)–(133) we see that

$$E\left[\zeta(T)\hat{X}(T) + \int_0^T \zeta(s)\hat{c}(s)ds\right] = \mathcal{X}(\hat{\gamma}) = \mathcal{X}(\mathcal{Y}(x)) = x.$$

Thus, according to Remark 2.1 and Lemma 2.1, condition (121) is satisfied. From equations (125)–(126), we see that condition (122) is also satisfied by the control $(\hat{\pi}, \hat{c})$ defined by (138)–(139).

According to Theorem 2.1, we have proved the following theorem.

THEOREM 2.2. *Suppose that assumption* (131) *holds, and that we have a complete market* (*i.e.,* $m = d$ *and* $\sigma(t)$ *invertible*). *Then the optimal consumption rate process and the optimal portfolio process are given by equations* (138)–(139). *This consumption-investment policy determines a wealth process* $\hat{X}$ *given by*

$$\hat{X}(t) = \frac{1}{\zeta(t)} E\left[\zeta(T) I_2(\mathcal{Y}(x)\zeta(T)) + \int_t^T \zeta(s) I_1(s, \mathcal{Y}(x)\zeta(s)) ds | \mathcal{F}_t\right] \quad \forall 0 \leq t \leq T. \tag{142}$$

We should also notice that the process described in Theorem 2.2 is the *unique solution* of Problem 2.1 (see equation (112)). In fact, since $U_1(t, \cdot)$ and U_2 are strictly concave functions, we may apply Theorem 1.6.

As in [8], we can obtain more results about the consumption-investment problem. Since those results are not a direct application of Theorem 1.4, we prefer to stop here, and refer the interested reader to [8]. Related optimization problems in financial markets appear in [24]–[28] and [40].

3. The square-integrable controls.

3.1. Introduction. In this section we reduce the class of admissible controls and obtain a stochastic maximum principle which is not only a necessary condition for optimality, but also a sufficient one. We also get some results concerning the existence and uniqueness of optimal controls.

Although all the previous versions of the stochastic maximum principle in which the control enters into the diffusion coefficient require at least that the controls be square-integrable, only one of them (Bismut [6]) supplies both necessary and sufficient conditions for optimality. Since that paper considers only the linear-regulator problem, we extend that work by considering general convex cost functions and allowing state constraints. Other references in which the stochastic maximum principle is also a sufficient condition for optimality are [19] and [21, Chap. 10], but those two references allow neither the diffusion coefficient to be controlled nor random coefficients.

3.2. The problem. We are going to consider a problem similar to the one stated in §1. The difference is that now the set of controls has more *structure*, and the terminal and running costs satisfy an additional condition.

DEFINITION 3.1. $\mathcal{U}^*$ *is the set of controls* $u \in \mathcal{U}$ *that satisfy*

$$\|u\|^2 := E \int_0^T |u_t|^2 dt < \infty. \tag{143}$$

As in the general theory of §1, the trajectory of the system controlled by $u \in \mathcal{U}^*$ satisfies equations (8)–(9). We want to minimize the functional $J : \mathcal{U}^* \longmapsto \Re$ defined by equation (17). That is, we want to select the control $\hat{u} \in \mathcal{U}^*$ that solves the problem

$$\min_{u \in \mathcal{U}^*} J(u). \tag{144}$$

We impose the same assumptions about the coefficients of the system as in §1, but now we take advantage of the restriction stated by equation (143). From Corollary 2.5.10 of [30, p. 85], we know that for every u in $\mathcal{U}^*$

$$E\left[\sup_{0 \leq t \leq T} |X_t^u|^2\right] < \infty. \tag{145}$$

Hence, $\forall u \in \mathcal{U}^*$,

$$E\left[\left(\int_0^T |f(t,X_t^u,u_t)|dt\right)^2\right] = E\left[\left(\int_0^T |A_tX_t^u + B_tu_t + C_t|dt\right)^2\right] \tag{146}$$
$$\leq E\int_0^T |A_tX_t^u + B_tu_t + C_t|^2dt < \infty,$$

and

$$E\int_0^T |g(t,X_t^u,u_t)|^2dt = E\int_0^T |D_tX_t^u + E_tu_t + F_t|^2dt < \infty. \tag{147}$$

We should note that the version of the stochastic maximum principle that we are going to develop in this section cannot be applied to the consumption-investment problem. In fact, conditions (143), (145)–(147) are *not* satisfied by the model of §2.

3.3. Assumptions. When we consider $\mathcal{U}^*$ instead of $\mathcal{U}$ as our set of admissible controls we have, in particular, that equations (146) and (147) hold. Now, we are going to impose an additional assumption.

Assumption 3.1. In addition to Assumptions 1.1 and 1.2 of §1 we shall assume in this section that for every $u, v \in \mathcal{U}^*, \rho \in [0,1]$,

$$E\int_0^T |L_x(t, X_t^v + \rho X_t^u, v_t + \rho u_t)|^2dt < \infty. \tag{148}$$

According to Hölder's inequality and inequalities (143) and (145), a sufficient condition for Assumption 3.1 to hold is that there exists a constant $K \in (0,\infty)$ such that for every $(t,x,u) \in [0,T] \times V \times U$,

$$|\Psi_x(x)|^2 + |L_x(t,x,u)|^2 + |L_u(t,x,u)|^2 \leq K(1+|x|^2+|u|^2). \tag{149}$$

3.4. The adjoint equation. We have already obtained some results about the adjoint equation in §1.4. The following proposition is a consequence of Theorem 1.1.

PROPOSITION 3.1. *Under the assumptions of this section, there is a unique pair of adapted processes (p,q) that satisfies equations* (19)–(20), *as well as conditions* (24)–(25).

3.5. The stochastic maximum principle. The purpose of the stochastic maximum principle is usually to state a necessary condition for the optimality of a given control $\hat{u} \in \mathcal{U}$. Nevertheless, the version of the stochastic maximum principle that we are going to build provides not only a necessary but also a sufficient condition for optimality. Let us suppose that $\hat{u}$ is a fixed control, a candidate to be optimal for the problem (144), and $\hat{X}$ is the corresponding state process. We shall start by studying the process S defined in §1.5.

LEMMA 3.1. *Under the assumptions of this section, for every $u \in \mathcal{U}^*$ the process S^u defined by*

$$S_t^u := \int_0^t \{p_s \bullet g(s,X_s^u,u_s) + X_s^u \bullet q_s\}dW_s, \quad 0 \leq t \leq T \tag{150}$$

is a martingale.

Proof. Obviously, the process S^u is a local martingale, but now we want to prove that S^u is indeed a martingale. For that purpose it is good enough to check that S^u is of class D[0,T], or even the weaker condition $E[\sup_{0\leq t\leq T}|S^u(t)|] < \infty$. According to the Burkholder–Davis–Gundy inequality (see, for instance, [25, Thm. 3.3.28]), to check this weaker condition it is enough to verify that

$$E\left[\left(\int_0^T \{|p_s \bullet g(s, X_s^u, u_s)|^2 + |X_s^u \bullet q_s|^2\} ds\right)^{\frac{1}{2}}\right] < \infty.$$

But this follows from Hölder's inequality, together with inequalities (24)–(25), (145), and (147). Hence, $E[\sup_{0\leq t\leq T}|S^u(t)|] < \infty$, which implies that S^u is indeed a martingale. □

COROLLARY 3.1. $\forall u \in \mathcal{U}^*$,

$$E[R_T^{\hat{u}}] = E[R_T^u]. \tag{151}$$

Proof. This is an immediate consequence of equation (75). □

THEOREM 3.1 (state-constraints). *Suppose that for each $(t, \omega, x) \in [0,T] \times \Omega \times V$, $L(t, x, \cdot)$ is strictly convex. Then, under the assumptions of this section, $\hat{u}$ is a solution of the optimal control problem*

$$\min_{u\in\mathcal{U}^*} E\left[\int_0^T L(t, X_t^u, u_t)dt + \Psi(X_T^u)\right] \tag{152}$$

if and only if

$$\hat{u} = \tilde{v}, \quad Leb \otimes P - a.e. \text{ on } [0,T] \times \Omega, \tag{153}$$

where $\tilde{v}$ is a control that satisfies

$$\max_{u\in U} H(t, p_t, q_t, \hat{X}_t, u) = H(t, p_t, q_t, \hat{X}_t, \tilde{v}_t), \quad X^{\tilde{v}} \in V, \tag{154}$$

$Leb \otimes P - a.e.$ on $[0,T] \times \Omega$.

Proof. This is an immediate consequence of Corollaries 1.2 and 3.1. □

THEOREM 3.2 (no state-constraints). *Suppose that $V = \Re^n$. Then $\hat{u}$ is a solution of the optimal control problem of* (152) *if and only if*

$$\max_{u\in U} H(t, p_t, q_t, \hat{X}_t, u) = H(t, p_t, q_t, \hat{X}_t, \hat{u}_t), \quad Leb \otimes P - a.e. \text{ on } [0,T] \times \Omega. \tag{155}$$

Proof. This is an immediate consequence of Corollaries 1.3 and 3.1. □

An interesting remark about this version of the stochastic maximum principle is that it provides not only a necessary, but also a sufficient condition for optimality. That is, *a control is optimal if and only if it maximizes the Hamiltonian.*

Our method of applying the theory of convex analysis to the stochastic control problem of this paper can also provide the following result about the existence and uniqueness of the optimal control.

THEOREM 3.3. *Let us suppose that there exists a convex function $k : \Re \longmapsto \Re$ such that $\lim_{x\to\infty} k(x) = \infty$ and $\forall (t, \omega, x, u) \in [0,T] \times \Omega \times V \times U$,*

$$L(t, x, u) \geq k(|u|^2).$$

Let us also suppose that the function $\Psi(\cdot)$ is bounded from below by a constant $\eta \in \Re$. Then problem (144) *has at least one solution. It has a unique solution if, for each $(t,\omega) \in [0,T] \times \Omega$, $L(t,\cdot,\cdot)$ or $\Psi(\cdot)$ are strictly convex.*

Proof. Applying Jensen's inequality, $\forall u \in \mathcal{U}^*$,

$$\begin{aligned} J(u) &= E\left[\int_0^T L(t, X_t^u, u_t)dt + \Psi(X_T^u)\right] \\ &\geq E\left[\int_0^T k(|u_t|^2)dt + \eta\right] \geq k(\|u\|^2) + \eta. \end{aligned}$$

Thus J is coercive over $\mathcal{U}^*$, that is, $\lim_{\|u\|\to\infty} J(u) = \infty$. Applying Proposition 2.1.2 of [11, p. 35], we get our result. □

We shall finish this paper by showing that the results of this section can be applied to solve three problems that arise in engineering: the linear-regulator problem, the predicted-miss problem, and the Beneš problem.

3.6. Example 1: The linear regulator.

3.6.1. The model. In the linear-regulator model, $\mathcal{U}$ is the set of controls $u : [0,T] \times \Omega \longmapsto \Re^k$ that are measurable, $(\mathcal{F}_t)$-adapted, and satisfy

$$\|u\|^2 = E\left[\int_0^T |u_t|^2 dt\right] < \infty. \tag{156}$$

Each $u \in \mathcal{U}$ determines a trajectory of the system that is the solution of the linear stochastic differential equation

$$dX_t^u = (A(t)X_t^u + B(t)u_t + C(t))dt + (D(t)X_t^u + E(t)u_t + F(t))dW_t, \tag{157}$$

$$X_0^u = x. \tag{158}$$

We assume that A, B, C, D, E, and F are all progressively measurable with respect to $(\mathcal{F}_t)$, and bounded uniformly in $(t,\omega) \in [0,T] \times \Omega$. We want to minimize the criterion $J : \mathcal{U} \longmapsto \Re$ defined by

$$J(u) = E\left[\int_0^T (X_t^{u*} M(t) X_t^u + u_t^* N(t) u_t)dt + X_T^{u*} \tilde{N} X_T^u\right], \tag{159}$$

where $M : [0,T] \times \Omega \mapsto \mathcal{L}(\Re^n; \Re^n)$ and $N : [0,T] \times \Omega \mapsto \mathcal{L}(\Re^k; \Re^k)$ are $(\mathcal{F}_t)$-progressively measurable, $\tilde{N} : \Omega \mapsto \mathcal{L}(\Re^n; \Re^n)$ is $\mathcal{F}_T$-measurable, and for each $(t,\omega) \in [0,T] \times \Omega$, the $n \times n$ matrices $M(t)$ and $\tilde{N}$ are symmetric, nonnegative definite, and the $k \times k$ matrix $N(t)$ is symmetric and positive definite. We shall also assume that M, N, and $\tilde{N}$ are bounded.

3.6.2. Assumptions. Following the notation of §1, we note that $U \equiv \Re^k$, $V \equiv \Re^n$, $f(t,x,u) \equiv A(t)x + B(t)u + C(t)$, $g(t,x,u) \equiv D(t)x + E(t)u + F(t)$, $L(t,x,u) \equiv x^* M(t) x + u^* N(t) u$, and $\Psi(x) \equiv x^* \tilde{N} x$.

Since we have a stochastic control problem with linear dynamics, convex cost-criterion, and unconstrained state, we may apply the general theory of §1 to solve this problem. In addition, since we are assuming (156), inequalities (145)–(147) are also valid for the linear regulator. Nevertheless, to apply the theory of §3.5, we have to check that Assumption 3.1 also holds in this model. But this is straightforward, applying the sufficient condition (149). *Therefore, we may apply the stochastic maximum principle, as stated in Theorem* 3.2. The details of the application of Theorem 3.2 to solve the linear-regulator problem can be found in Chapter 4 of [8]. The adjoint processes are found to be

$$p(t) = -P(t)\hat{X}(t) - R(t), \tag{160}$$

$$q(t) = -Q(t)\hat{X}(t) - P(t)[D(t)\hat{X}(t) + E(t)\hat{u}(t) + F(t)] - S(t), \tag{161}$$

and the optimal control given by

$$\begin{aligned}\hat{u}(t) = -[N(t) + E^*(t)P(t)E(t)]^{-1}\{&[B^*(t)P(t)\\ &+E^*(t)Q(t) + E^*(t)P(t)D(t)]\hat{X}(t) + B^*(t)R(t)\\ &+E^*(t)(P(t)F(t) + S(t))\},\end{aligned} \tag{162}$$

where $P : [0,T] \times \Omega \longmapsto \mathcal{L}(\Re^n; \Re^n)$, $Q : [0,T] \times \Omega \longmapsto \mathcal{L}(\Re^d; \mathcal{L}(\Re^n; \Re^n))$, are a pair of measurable, adapted processes that solve the backward matrix stochastic differential equation

$$\begin{aligned}dP(t) = &- \{P(t)A(t) + A^*(t)P(t) + D^*(t)P(t)D(t) + D^*(t)Q(t) + Q(t)D(t)\\ &-[D^*(t)P(t)E(t) + P(t)B(t) + Q(t)E(t)][N(t) + E^*(t)P(t)E(t)]^{-1}\\ &[E^*(t)P(t)D(t) + B^*(t)P(t) + E^*(t)Q(t)] + 2M(t)\}\, dt + Q(t)dW(t),\\ P(T) = &\ 2\tilde{N}.\end{aligned}$$

We notice that if we assume that the coefficients of the system are deterministic and take $Q \equiv 0$, we recover the familiar Riccati equation. On the other hand, $R : [0,T] \times \Omega \longmapsto \Re^n$ and $S : [0,T] \times \Omega \longmapsto \mathcal{L}(\Re^d; \Re^n)$ are a pair of measurable, adapted processes that solve the backward matrix stochastic differential equation

$$\begin{aligned}dR(t) = &\{[P(t)B(t) + D^*(t)P(t)E(t) + Q(t)E(t)]\ [N(t) + E^*(t)P(t)E(t)]^{-1}B^*(t)\\ &-A^*(t)\}\ R(t)dt\\ &+\{[(P(t)B(t) + D^*(t)P(t)E(t) + Q(t)E(t))\ (N(t) + E^*(t)P(t)E(t))^{-1}\\ &\quad E^*(t) - D^*(t)]\ [P(t)F(t) + S(t)] - P(t)C(t) - Q(t)F(t)\}\, dt + S(t)dW(t),\\ R(T) = &\ 0.\end{aligned}$$

3.7. Example 2: The predicted-miss problem.

3.7.1. The model. In this model, $\mathcal{U}$ is the set of controls $u : [0,T] \times \Omega \longmapsto [-1,1]^k$ that are measurable and $(\mathcal{F}_t)$-adapted. Each $u \in \mathcal{U}$ determines a trajectory of the system that is the solution of the linear stochastic differential equation

$$dX_t^u = (A(t)X_t^u + B(t)u_t)dt + F(t)dW_t, \tag{163}$$

$$X_0^u = x. \tag{164}$$

We assume that A, B, and F are deterministic, bounded, and measurable.

We want to minimize the criterion $J : \mathcal{U} \longmapsto \Re$ defined by

$$J(u) = E[k(v \bullet X_T^u)], \tag{165}$$

where $v \in \Re^n$ is fixed, and $k : \Re \longmapsto \Re$ is given by $k(y) = y^2$. This means that the controller is trying to drive the final state to the hyperplane $H = \{x \in \Re^n : v \bullet x = 0\}$ using bounded controls.

3.7.2. Assumptions. Following the notation of §1, we note that $U \equiv [-1,1]^k$, $V \equiv \Re^n$, $f(t,x,u) \equiv A(t)x + B(t)u$, $g(t,x,u) \equiv F(t)$, $L(t,x,u) \equiv 0$, and $\Psi(x) \equiv k(v \bullet x)$. We can also note that the trajectory of the system satisfies a linear stochastic differential equation, and that the criterion that we want to minimize is a convex functional. The trajectory of the system is unconstrained. Thus, we may apply the general theory of §1. It is not hard to check that the assumptions made in this section are all satisfied. *Hence, we may apply Theorem* 3.2 *to solve the predicted-miss problem.* The details of the application of Theorem 3.2 to solve this problem can be found in Chapter 5 of [8]. The first component of the adjoint process and the optimal control are found to be

$$p(t) = -2s(t)E[v \bullet \hat{X}_T | \mathcal{F}_t] \tag{166}$$

and

$$\hat{u}(t) = -\text{sgn}\{B^*(t)s(t)\}\text{sgn}\{s(t) \bullet \hat{X}(t)\}, \tag{167}$$

respectively, where $s(t) = \Phi^*(t)^{-1}\Phi^*(T)v$, and $\Phi : [0,T] \mapsto \mathcal{L}(\Re^n; \Re^n)$ is the deterministic solution of $\Phi(t) = I + \int_0^t A(s)\Phi(s)ds$.

3.8. Example 3: The Beneš problem.

3.8.1. The model. In this model, $\mathcal{U}$ is the set of measurable, $(\mathcal{F}_t)$-adapted processes $u : [0,T] \times \Omega \longmapsto U$, where $U := \{v \in \Re^k : |v| \leq 1\}$. Each $u \in \mathcal{U}$ determines a trajectory of the system that satisfies the linear stochastic differential equation

$$dX_t^u = ([H(t) + \gamma(t)I]X_t^u + \beta(t)u_t)dt + \alpha(t)dW_t, \tag{168}$$

$$X_0^u = x. \tag{169}$$

We assume that the function $H : [0,T] \longmapsto \mathcal{L}(\Re^n; \Re^n)$ is bounded, measurable, skew symmetric, and the functions $\alpha : [0,T] \longmapsto \Re$, $\beta : [0,T] \longmapsto \Re$, and $\gamma : [0,T] \longmapsto \Re$ are bounded, measurable, with α, β continuous. We assume, as in [18], [22], and [23], that the dimensions of the Brownian motion, the trajectory of the system, and the control are equal. That is, $d = n = k$. We want to minimize the criterion $J : \mathcal{U} \longmapsto \Re$ defined by

$$J(u) = E\left[\int_0^T l(t, X^u(t))dt + k(X^u(T))\right]. \tag{170}$$

In the above definition, $l : [0,T] \times \Re^n \longmapsto \Re$ is given by

$$l(t,x) = \eta(t)|x|^2,$$

where $\eta : [0,T] \longmapsto [0,\infty)$ is bounded and measurable. On the other hand, $k : \Re^n \longmapsto \Re$ is given by

$$k(y) = |y|^2.$$

This problem was first solved by Beneš [2].

3.8.2. Assumptions. Following the notation of §1, we note that $U \equiv \{v \in \mathcal{R}^k : |v| \leq 1\}$, $V \equiv \Re^n$, $f(t,x,u) \equiv [H(t) + \gamma(t)I]x + \beta(t)u$, $g(t,x,u) \equiv \alpha(t)I$, $L(t,x,u) \equiv l(t,x) = \eta(t)|x|^2$, and $\Psi(x) \equiv k(x) = |x|^2$. Since we have a stochastic control problem with linear dynamics, convex cost criterion, and unconstrained state, we may apply the general theory of §1 to solve this problem. It is not hard to check that all the assumptions made in this section are also satisfied. *Hence, we may apply Theorem* 3.2 *to solve the Beneš problem.* The details of the application of Theorem 3.2 to solve this problem can be found in Chapter 6 of [8]. The first component of the adjoint process is found to satisfy

$$\frac{p(t)}{|p(t)|} = -\frac{\hat{X}(t)}{|\hat{X}(t)|}, \tag{171}$$

and the optimal control is found to be

$$\hat{u}(t) = -\text{sgn}(\beta(t))\frac{\hat{X}(t)}{|\hat{X}(t)|}. \tag{172}$$

Acknowledgments. We are grateful to Professors Ulrich G. Haussmann, Shige Peng, and X. X. Xue for their comments, and to Professor Mark Saksonov for advancing his preprint [38]. The two referees made valuable suggestions that improved the manuscript considerably; we thank both of them.

REFERENCES

[1] V. Arkin and M. Saksonov, *Necessary optimality conditions for stochastic differential equations*, Soviet Math. Dokl., 20 (1979), pp. 1–5.

[2] V. Beneš, *Composition and invariance methods for solving some stochastic control problems*, Adv. Appl. Prob., 7 (1975), pp. 299–329.

[3] ———, *Full "bang" to reduce predicted miss is optimal*, SIAM J. Control Optim., 14 (1976), pp. 62–84.

[4] A. Bensoussan, *Lectures on stochastic control*, in Lecture Notes in Mathematics, 972 Springer-Verlag, Berlin, 1981, pp. 1–62.

[5] J. M. Bismut, *Conjugate convex functions in optimal stochastic control*, J. Math. Anal. Appl., 44 (1973), pp. 384–404.

[6] ———, *Linear quadratic optimal control with random coefficients*, SIAM J. Control Optim., 14 (1976), pp. 419–444.

[7] ———, *An introductory approach to duality in optimal stochastic control*, SIAM Rev., 20 (1978), pp. 62–78.

[8] A. Cadenillas, *Contributions to the stochastic version of Pontryagin's maximum principle*, Ph.D. thesis, Columbia University, New York, 1992.

[9] A. Cadenillas and U. G. Haussman, *The stochastic maximum principle for a singular control problem*, Stochastics Stochastics Rep., 49 (1994), pp. 211–237.

[10] J. Detemple, *Linear recursive integral equations: A solution procedure*, 1993, preprint.

[11] J. Ekeland and R. Temam, *Convex Analysis and Variational Problems*, North-Holland, Amsterdam, and American Elsevier, New York, 1976.

[12] R. J. Elliott, *The optimal control of diffusions*, Appl. Math. Optim., 22 (1990), pp. 229–240.

[13] R. J. Elliott and M. Kohlmann, *The second order minimum principle and adjoint process*, Stochastics Stochastics Rep., 46 (1994), pp. 25–39.

[14] W. Fleming and R. Rishel, *Deterministic and Stochastic Optimal Control*, Springer-Verlag, New York, 1975.

[15] U. G. Haussmann, *General necessary conditions for optimal control of stochastic systems*, Math. Programming Stud., 6 (1976), pp. 30–48.

[16] ———, *On the stochastic maximum principle*, SIAM J. Control Optim., 16 (1978), pp. 236–251.

[17] ———, *On the adjoint process for optimal control of diffusion processes*, SIAM J. Control Optim., 19 (1981), pp. 221–243.

[18] ———, *Some examples of optimal stochastic controls*, SIAM Rev., 23 (1981), pp. 292–307.

[19] U. G. Haussmann, *Extremal controls for completely observable diffusions*, in Lecture Notes in Control and Information Sciences, 42, Springer-Verlag, Berlin, New York, 1982, pp. 149–160.

[20] ———, *Optimal control of partially observed diffusions via the separation principle*, in Lecture Notes in Control and Information Sciences, 43, Springer-Verlag, Berlin, New York, 1982, pp. 302–311.

[21] ———, *A Stochastic Maximum Principle for Optimal Control of Diffusions*, Longman Scientific and Technical, Essex, U.K., 1986.

[22] N. Ikeda and S. Watanabe, *A comparison theorem for solutions of stochastic differential equations and its applications*, Osaka J. Math., 14 (1977), pp. 619–633.

[23] ———, *Stochastic Differential Equations and Diffusion Processes*, North–Holland, Amsterdam, 1981.

[24] I. Karatzas, *Optimization problems in the theory of continuous trading*, SIAM J. Control Optim., 27 (1989), pp. 1221–1259.

[25] I. Karatzas and S. Shreve, *Brownian Motion and Stochastic Calculus*, Springer-Verlag, New York, 1988.

[26] I. Karatzas, J. Lehoczky, S. Sethi, and S. Shreve, *Explicit solution of a general consumption/investment problem*, Math. Oper. Res., 11 (1986), pp. 261–294.

[27] I. Karatzas, J. Lehoczky, and S. Shreve, *Optimal portfolio and consumption decisions for a small investor on a finite horizon*, SIAM J. Control Optim., 25 (1987), pp. 1557–1586.

[28] I. Karatzas, J. Lehoczky, S. Shreve, and G. Xu, *Martingale and duality methods for utility maximization in an incomplete market*, SIAM J. Control Optim., 29 (1991), pp. 702–730.

[29] I. Karatzas, D. Ocone, and J. Li, *An extension of Clark's formula*, Stochastics Stochastics Rep., 37 (1991), pp. 127–131.

[30] N. V. Krylov, *Controlled Diffusion Processes*, Springer-Verlag, Berlin, 1980.

[31] H. J. Kushner, *On the stochastic maximum principle: Fixed time of control*, J. Math. Anal. Appl., 11 (1965), pp. 78–92.

[32] ———, *Necessary conditions for continuous parameter stochastic optimization problems*, SIAM J. Control, 10 (1972), pp. 550–565.

[33] E. Pardoux and S. Peng, *Adapted solution of a backward stochastic differential equation*, Systems Control Lett., 14 (1990), pp. 55–61.

[34] S. Peng, *A general stochastic maximum principle for optimal control problems*, SIAM J. Control Optim., 28 (1990), 966–979.

[35] ———, *Backward stochastic differential equations and applications to optimal control*, Appl. Math. Optim., 27 (1993), pp. 125–144.

[36] L. S. Pontryagin, *Optimal regulation processes*, Amer. Math. Soc. Transl., Ser. 2, 18 (1961), pp. 321–339.

[37] L. S. Pontryagin, V. G. Boltyanskii, R. V. Gamkrelidze, and E. F. Mischenko, *The Mathematical Theory of Optimal Processes*, Wiley-Interscience, New York, 1962.

[38] M. Saksonov, *The variational principles in stochastic control*, Dushanbe State Educational Institute, 1989, preprint.

[39] M. Schäl, *A selection theorem for optimization problems*, Arch. Math., 25 (1974), pp. 219–224.

[40] S. Sethi, M. Taksar, and E. Presman, *Explicit solution of a general consumption/portfolio problem with subsistence consumption and bankruptcy*, J. Econom. Dynamics Control, 16 (1992), pp. 747–768.

[41] P. Whittle, *A risk-sensitive maximum principle*, Systems Control Lett., 15 (1990), pp. 183–192.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 33, No. 2, pp. 625–642, March 1995

©1995 Society for Industrial and Applied Mathematics

014

PERSISTENCY OF EXCITATION IN IDENTIFICATION USING RADIAL BASIS FUNCTION APPROXIMANTS*

A. J. KURDILA†, FRANCIS J. NARCOWICH‡, AND JOSEPH D. WARD‡

Abstract. In this paper, identification algorithms whose convergence and rate of convergence hinge on the regressor vector being persistently exciting are discussed. It is then shown that if the regressor vector is constructed out of radial basis function approximants, it will be persistently exciting, provided a kind of "ergodic" condition is satisfied. In addition, bounds on parameters associated with the persistently exciting regressor vector are provided; these parameters are connected with both the convergence and rates of convergence of the algorithms involved.

Key words. persistency of excitation, identification, radial basis functions

AMS subject classifications. Primary 93B30, 41A30; secondary 93D20, 41A63

1. Introduction.

In applications to neural networks and adaptive control, the use of radial basis function (RBF) approximants has become increasingly popular over the past few years. This popularity can be attributed to several factors. First, because data is collected in control applications in (near) real-time, uniform sampling of trajectories in time naturally leads to scattered data in phase space. The utility of radial basis approximants for scattered data is well documented [2]–[7], [10]–[15], [17]. Second, RBF approximants appear to be particularly well suited to tracking control. Papers by Sanner and Slotine [22]–[24], Tzirkel-Hancock and Fallside [28], and the present authors [8] provide examples in which direct adaptive tracking control using RBFs is achieved. Finally, several papers have appeared in the literature that use RBFs as the foundation of neural network architectures designed for identification and control [1], [9]. The relationship of some of these architectures and classical questions of "hypersurface reconstruction" in approximation theory has been addressed in [16].

Algorithms for handling identification problems have been studied extensively [22]–[25], [28]. In this paper we discuss two such algorithms, a standard least-squares method and a gradient-descent/dead-zone method introduced by Sanner and Slotine [22]–[24], and show that persistency of excitation of the regressor vector employed in each of them is a sufficient condition for them to converge at an exponential rate. (The gradient-descent/dead-zone algorithm always converges whether or not the regressor vector is perisistently exciting [22]–[24].) We then proceed to show that if the regressor vector is constructed out of RBF approximants, it will be persistently exciting, provided a kind of "ergodic" condition is satisfied. Finally, we will provide bounds on the parameters associated with the persistently excited regressor vector; these parameters are associated with convergence and rates of convergence.

The remainder of the paper is organized as follows. To complete the introduction, we will briefly discuss radial basis functions. In §2 we will provide a description of the identification problem and discuss several algorithms for dealing with it. In §3 we will prove the main theorem, Theorem 3.5, alluded to above. In §4 we will obtain bounds on quantities associated with a persistently excited regressor vector.

*Received by the editors June 2, 1992; accepted for publication (in revised form) November 23, 1993. This work was supported by Air Force Office of Scientific Research grant F49620-92-J-0403DEF.

†Department of Aerospace Engineering, Texas A&M University, College Station, Texas 77843.

‡Department of Mathematics, Texas A&M University, College Station, Texas 77843.

Radial basis functions. In this paper, we will be concerned with two classes of RBFs. The first class comprises every nonconstant $F(r)$ that can be represented via the formula

$$F(r) = \int_0^\infty e^{-r^2\rho} d\beta(\rho), \tag{1.1}$$

where $d\beta(\rho)$ is a finite, nonnegative Borel measure on $[0,\infty)$ and $r \in [0,\infty)$. (Such functions are closely related to *completely monotonic* functions; see [12], [26], [29].) RBFs have an "order" $m = 0, 1, \ldots$, associated with them. The functions described in (1.1) are "order-0" RBFs, and they include the Gaussian RBFs as well as the inverse multiquadric, $(1+r^2)^{-1/2}$. It should be mentioned that RBFs of the form (1.1) do not exhaust the class of all possible order-0 RBFs, but they do include all of the ones in use.

The second class comprises those order-1 RBFs that can be represented via a formula similar to (1.1). Let $d\eta(\rho)$ be a nonnegative Borel measure on $[0,\infty)$ for which $\int_1^\infty \rho^{-1} d\eta(\rho) < \infty$. All functions of the form

$$F(r) = F(0) + \int_0^\infty \frac{1 - e^{-r^2\rho}}{\rho} d\eta(\rho) \tag{1.2}$$

such that $r^{-1}F'(r)$ is nonconstant are order-1 RBFs that make up the second class. We remark that the case when $r^{-1}F'(r)$ is a constant is degenerate. Also, it should be noted that although the constant $F(0)$ in (1.2) can be arbitrary, it is taken to be nonnegative. The reason for making this assumption involves the invertibility of a certain interpolation matrix [12], [10], [13]–[15]. Finally, it is easy to show that if F is in the first class, then $F(0) - F(r)$ is in the second. On the other hand, there are many order-1 RBFs in the second class that do not arise in this way. Since the only order-0 RBFs considered here are those of the form (1.1), and the only order-1 RBFs are those of the form (1.2), we will simply refer to functions of the first class as "order-0 RBFs," and to those of the second class as "order-1 RBFs." No confusion should result from this minor abuse of terminology.

The most prominent members of the second class that are not in the first are the Hardy multiquadric function, $\sqrt{1+r^2}$, a function studied by Dyn [3], $\log(1+r^2)$, and the distance function r. As was mentioned above in connection with the first class, RBFs of the form (1.2) do not exhaust the class of all possible order-1 RBFs, but they do include all the ones of practical importance. For a complete discussion of the definition, classification, and representation of RBFs, we refer the interested reader to the papers by Powell [17], [18].

Two important features concerning RBFs are the stability of their associated interpolation and least-squares matrices [13], [15], [19], [27], and their ability to provide uniform approximations to smooth functions on compact sets. A typical interpolant/approximant $L(x)$ is formed in the following way. Pick a set of distinct points in $\mathbb{R}^s$, $\Xi := \{\xi_j\}_{j=1}^N$; these are called the *centers*. The function L is then selected from the span of the family $\{F(x - \xi_j)\}_{j=1}^N$, where F is a fixed RBF.

For the case in which one wishes to use RBFs to uniformly approximate a smooth function f defined on a compact set $\Omega \subset \mathbb{R}^s$, one first uniformly approximates f by a band-limited function f_B (see [24] for a discussion). For uniform approximation of f_B with RBFs, one may use a result of Madych and Nelson [11, Thm. 4.4].

Let the centers $\xi_1, \ldots, \xi_N$ be scattered throughout Ω in a way that makes the quantity

$$d = d(\Xi, \Omega) := \sup_{y\in\Omega} \inf_{x\in\Xi} \|y - x\|_2$$

small. Because Ξ is a subset of Ω, d turns out to be the standard Hausdorff distance between compact sets. Indeed, when the centers are placed at the vertices of an s-cube whose sides have length h, it is easy to show that $d = \frac{1}{2}\sqrt{s}\,h$. The result in [11] is as follows. For certain RBFs F (including Gaussian and multiquadrics) and arbitrary k, there exist coefficients $\{a_j\}$ such that

$$\sup_{x\in\Omega}\left|f_B(x) - \sum_j a_j F(x - x_j)\right| = \mathcal{O}(d^k).$$

Thus, one may uniformly approximate the band-limited function f_B by linear combinations of translates of RBFs. The usual "up-over-and-around" argument then implies that one can uniformly approximate the original function f on Ω by such linear combinations.

2. An identification problem.

Algorithms for handling identification problems have been studied extensively [22]–[25], [28]. In this section, we wish to discuss two such algorithms, a standard least-squares method and a gradient-descent/dead-zone method introduced by Sanner and Slotine [22]–[24]. Specifically, we wish to analyze the problem of trying to determine the functional form of a continuous output funtion $G : \Omega \to \mathbb{R}$, where Ω is a compact subset of the state space, $\mathbb{R}^s$. The information available is the signal $Y(t) = G(x(t))$, where $x(t)$ denotes a continuous function of t taking values in Ω.

To deal with this identification problem, assume that $\mathfrak{W} = \{w_k(x)\}_{k=1}^N$ is a fixed set of continuous functions (RBFs with various centers, for example) defined on the state space. The basis is chosen so that $G(x)$ can be uniformly approximated to within an acceptable level of error by (time-independent) linear combinations of the w_k's. More precisely, given an error-level $\Phi > 0$, we suppose that there exist coefficients c_k^* for which

$$\left\|G(x) - \sum_{k=1}^N c_k^* w_k(x)\right\|_{\infty,\Omega} \leq \Phi,$$

where $\|\cdot\|_{\infty,\Omega}$ denotes the supremum norm in Ω. Whether a given basis provides an acceptable approximation is an issue that must be addressed. For RBFs, we discussed the matter when we reviewed them earlier.

The coefficients c_k^* may not be unique, even if they make $\sum_{k=1}^N c_k^* w_k(x)$ a "best approximant" to $G(x)$, because the unit ball in L^∞ is not strictly convex. Among all sets of coefficients that yield a best approximant, one may uniquely specify a set by imposing the additional requirement that $\sum_k (c_k^*)^2$ be a minimum.

Because knowing $c_1^*, \ldots, c_N^*$ is equivalent to knowing the function G, at least to within the error-level, Φ, our problem amounts to using the available information on $x(t)$ and the signal $Y(t) = G(x(t))$ described above to identify approximately the c_k^*'s. We wish to discuss both continuous and discrete-time methods to identifty these coefficients.

The idea common to all of these methods is to choose time-dependent coefficients $c_k(t)$, $k = 1, \ldots, N$, in such a way that each $c_k(t)$ converges to a neighborhood of the c_k^* as $t \to \infty$. Thus, for large t, the time-dependent approximation to G,

$$\tilde{G}(t,x) = \sum_{k=1}^N c_k(t) w_k(x),$$

will then be sufficiently close to

$$G^*(x) := \sum_k c_k^* w_k(x)$$

to render the same level of approximation to G as G^* itself.

We remark that, to bring the quantity $|Y(t) - \tilde{G}(t, x(t))|$ (the prediction error) to a neighborhood of zero along a particular trajectory, it is *not* necessary to have each $c_j(t)$ converge to a neighborhood of the corresponding c_j^*, at least in the case of the gradient-descent/dead-zone algorithm [22]–[24]. For this algorithm, the convegence of $c_j(t)$ will be an added feature that occurs when the regressors are persistently exciting. See Proposition 2.2 below.

These methods are better formulated with the aid of additional notation, which we now introduce. First of all, let

$$\Delta(x) := G(x) - G^*(x) = G(x) - \sum_{k=1}^{N} c_k^* w_k(x).$$

Our error criterion then becomes

$$\|\Delta\|_{\infty,\Omega} \le \Phi.$$

With a slight abuse of notation, set $\Delta(t) = \Delta(x(t))$. In addition, define the following quantities:

$$w(t) := \begin{pmatrix} w_1(x(t)) \\ \vdots \\ w_N(x(t)) \end{pmatrix}, \qquad c(t) := \begin{pmatrix} c_1(t) \\ \vdots \\ c_N(t) \end{pmatrix}, \qquad c^* := \begin{pmatrix} c_1^* \\ \vdots \\ c_N^* \end{pmatrix}.$$

In this notation, we have $Y(t) = w(t)^T c^* + \Delta(t)$ and $\tilde{G}(t, x(t)) = w(t)^T c(t)$. We now turn to our first identification method.

Least squares. Let $\alpha > 0$ be a positive constant that is at our disposal, and let $c \in \mathbb{R}^N$. The object is to minimize

$$E(t) := \alpha\|c\|^2 + \int_{[0,t]} (w(\tau)^T c - Y(\tau))^2 d\mu(\tau)$$

where $d\mu(\tau) = d\tau$ in the continuous case, and $d\mu(\tau) = \sum_{i=0}^{\infty} \delta(\tau - i) d\tau$ in the discrete case. (We will try to treat the discrete and continuous cases simultaneously when this is possible. Doing so may require that we think of a sequence as a continuous function evaluated at integers.) This is a straightforward task that yields [25, p. 50]

$$c(t) = \left(\alpha I + \int_{[0,t]} w(\tau) w(\tau)^T d\mu(\tau)\right)^{-1} \int_{[0,t]} Y(\tau) w(\tau) d\mu(\tau). \tag{2.1}$$

The parameter error is defined to be $\phi(t) := c(t) - c^*$. If one uses $Y(\tau) = w(\tau)^T c^* + \Delta(t)$, then one may put the last equation in terms of $\phi(t)$,

$$\phi(t) = \left(\alpha I + \int_{[0,t]} w(\tau) w(\tau)^T d\mu(\tau)\right)^{-1} \left(-\alpha c^* + \int_{[0,t]} \Delta(\tau) w(\tau) d\mu(\tau)\right). \tag{2.2}$$

The regressor vector $w(t)$ is *persistently exciting* [25, p. 72] if there exist positive constants δ, α_1, and α_2 such that

$$\alpha_2\|c\|^2 \geq \int_{[t_0,t_0+\delta]} |w(\tau)^T c|^2 d\mu(\tau) \geq \alpha_1\|c\|^2, \tag{2.3}$$

where $\|\cdot\| = \|\cdot\|_2$, holds for every $t_0 \geq 0$ and every constant column vector $c = (c_1 \cdots c_N)^T \in \mathbb{R}^N$. If persistency of excitation holds in the case at hand, then we have the following result.

PROPOSITION 2.1. *In the least-squares algorithm, if the regressor vector $w(t)$ is persistently exciting, then*

$$\limsup_{t\to\infty} \|\phi(t)\| \leq \frac{\sqrt{\alpha_2\delta}\|\Delta\|_{\infty,\Omega}}{\alpha_1} \leq \frac{\sqrt{\alpha_2\delta}\,\Phi}{\alpha_1}. \tag{2.4}$$

Proof. In (2.2), the parameter error has the form $\phi(t) = A(t)^{-1}(-\alpha c^* + v(t))$. The matrix $A(t)$ being inverted in (2.2) is real, self-adjoint, and positive definite. Using (2.3) and $t \geq m\delta$, we may estimate its lowest eigenvalue as $\alpha + m\alpha_1$. Consequently, the norm of its inverse is $(\alpha + m\alpha_1)^{-1}$, and so

$$\|\phi(t)\| \leq (\alpha + m\alpha_1)^{-1}(\alpha\|c^*\| + \|v(t)\|). \tag{2.5}$$

To obtain the norm of $v(t)$, let $b \in \mathbb{R}^N$, and note that

$$b^T v(t) = \int_{[0,t]} \Delta(\tau) b^T w(\tau) d\mu(\tau).$$

Applying Schwarz's inequality, $t \leq (m+1)\delta$, the inequality $|\Delta(\tau)| \leq \|\Delta\|_{\infty,\Omega}$, and (2.3), we arrive at

$$(b^T v(t))^2 \leq (m+1)\delta\|\Delta\|^2_{\infty,\Omega} \int_{[0,(m+1)\delta]} (b^T w(\tau))^2 d\mu(\tau) \leq (m+1)^2\delta\|\Delta\|^2_{\infty,\Omega}\alpha_2\|b\|^2,$$

from which it follows that

$$\|v(t)\| = \left\| \int_{[0,t]} \Delta(\tau) w(\tau) d\mu(\tau) \right\| \leq (m+1)\|\Delta\|_{\infty,\Omega}\sqrt{\alpha_2\delta} \quad \text{if } t \leq (m+1)\delta. \tag{2.6}$$

Combining (2.5) and (2.6) yields

$$\|\phi(t)\| \leq \frac{\alpha\|c^*\| + \sqrt{\alpha_2\delta}\,(m+1)\|\Delta\|_{\infty,\Omega}}{\alpha + m\alpha_1}. \tag{2.7}$$

Letting $t \to \infty$ above yields (2.4). □

We proved above that if the regressor vector $w(t)$ is persistently exciting, then as $t \to \infty$, the parameter error $\phi(t)$ enters a neighborhood of the origin, and $c(t)$ enters a neighborhood of c^*. It is clear that persistency of excitation is crucial to convergence of the method.

We remark that the algorithm treated above is least squares in its "classic" form. In general, one would use a modified, recursive form of the algorithm [25]. Rather than treat one of these, we will turn to the the algorithm below, which is naturally formulated in a recursive way.

Gradient-descent/dead-zone algorithm. Sanner and Slotine [22]–[24] have discussed and analyzed a class of gradient-descent algorithms that employ a dead zone. If $c(t)$ is the parameter vector at time t and $\phi(t)$ is the parameter error, then define the prediction error to be

$$(2.8)\quad \epsilon(t) := \begin{cases} w(t)^T c(t) - Y(t) = w(t)^T \phi(t) - \Delta(t) & \text{(continuous time)}, \\ w(t)^T c(t-1) - Y(t) = w(t)^T \phi(t-1) - \Delta(t) & \text{(discrete time)}, \end{cases}$$

and the dead-zone prediction error to be

$$(2.9)\quad \epsilon_d(t) = \begin{cases} \epsilon(t) - \Phi \operatorname{sgn}(\epsilon(t)) & \text{if } |\epsilon| \geq \Phi, \\ 0 & \text{if } |\epsilon| \leq \Phi. \end{cases}$$

To update $c(t)$, we require that for continuous time,

$$(2.10)\quad \dot{c}(t) = -\eta \epsilon_d(t) w(t),$$

and for discrete time,

$$(2.11)\quad c(t) = c(t-1) - \eta \epsilon_d(t) w(t).$$

Here, η is another positive constant that is at our disposal. Of course, we may replace the left side in each of these equations by $\phi(t) = c(t) - c^*$ because c^* is constant.

For the case in which the basis comprised Gaussians centered on a discrete lattice, both the selection of the error-level Φ and the stability of the algorithm were analyzed in detail in [23], [24]. While more work is required in the error-level selection for the scattered-data case, their stability analysis applies to the case at hand almost verbatim. Here is a brief summary of that analysis.

Choose the function $V(t) := \frac{1}{2}\|\phi(t)\|^2$ as a Lyapunov function. In the continuous case, the derivative of V is negative; indeed, it satisfies

$$(2.12)\quad \dot{V} \leq -\eta \epsilon_d(t)^2.$$

A similar result applies in the discrete case, although the analysis forces a restriction on the constant η. If $0 < \eta < 2(\sum_k \|w_k(x)\|^2_{\infty,\Omega})^{-1}$, then

$$(2.13)\quad V(t) - V(t-1) \leq -\eta \epsilon_d(t)^2.$$

Thus, in either case, $V(t)$ is decreasing, and the algorithm is stable.

An obvious by-product of the analysis is that, in the continuous case, the dead-zone prediction error is square integrable, and, in the discrete case it is square summable. This observation has some important consequences. In the discrete case, $\epsilon_d(t)$ being square summable implies that $\epsilon_d(t) \to 0$ as $t \to \infty$. In the continuous case, the same conclusion can be arrived at if a few mild regularity assumptions are made. Suppose that $x(t)$ has a uniformly bounded, piecewise continuous derivative on $[0, \infty)$, and that $G(x)$ and the basis functions $w_k(x)$ are continuously differentiable on Ω. It easily follows that the continuous-case dead-zone prediction error $\epsilon_d(t)$ is continuous and has a bounded, piecewise continuous derivative $\dot{\epsilon}_d(t)$. Following the argument in

[24, p. 850] then yields that $\epsilon_d(t) \to 0$ as $t \to \infty$. Note that persistency of excitation is *not* required for these convergence results.

As in the least-squares algorithm, persistency of excitation guarantees that the parameter error vector will asymptotically enter a neighborhood of the origin—equivalently, $c(t)$ will asymptotically enter a neighborhood of c^*—with the neighborhood's size being controlled by the level of error, Φ. Indeed, we have this result.

PROPOSITION 2.2. *In the gradient-descent/dead-zone algorithm, if the regressor vector $w(t)$ is persistently exciting, then*

$$\limsup_{t\to\infty} \|\phi(t)\| \le 2\Phi\sqrt{\frac{\delta}{\alpha_1}}. \tag{2.14}$$

Proof. To do the continuous case, integrate (2.10) to get the equation

$$\phi(t') - \phi(t) = -\eta \int_t^{t'} \epsilon_d(\tau) w(\tau) d\tau.$$

Fix t. By multiplying both sides of this equation by $w(t')^T$ and then rearranging and manipulating its terms, we obtain

$$w(t')^T\phi(t) = \epsilon_d(t') + \gamma(t') + \eta\zeta(t'), \tag{2.15}$$

where $\gamma(t') := \epsilon(t') - \epsilon_d(t') + \Delta(t')$ and $\zeta(t') = \int_t^{t'} \epsilon_d(\tau)w(t')^T w(\tau)d\tau$. From (2.8) and (2.9), coupled with $|\Delta(t')| \le \Phi$, we see that

$$|\gamma(t')| \le 2\Phi. \tag{2.16}$$

Let δ be as in (2.3), which holds because $w(t)$ is persistently exciting. Compute the $H = L^2([t, t+\delta])$-norm of both sides, and use the upper estimate on the norm of the right side obtained with the triangle inequality to get

$$\|w(\cdot)^T\phi(t)\|_H \le \|\epsilon_d(\cdot)\|_H + \|\gamma(\cdot)\|_H + \eta\|\zeta(\cdot)\|_H. \tag{2.17}$$

From (2.3), we have that the left side of (2.17) satisfies

$$\|w(t')^T\phi(t)\|_H = \sqrt{\int_t^{t+\delta} (w(t')^T\phi(t))^2 dt'} \ge \sqrt{\alpha_1}\|\phi(t)\|. \tag{2.18}$$

In addition, (2.16) implies that

$$\|\gamma(\cdot)\|_H \le 2\Phi\sqrt{\delta}. \tag{2.19}$$

To estimate $\|\zeta(\cdot)\|_H$, first estimate $\zeta(t')$ using Schwarz's inequality and (2.3) as follows:

$$\begin{aligned}
|\zeta(t')| &\le \sqrt{\int_t^{t'} \epsilon_d(\tau)^2 d\tau}\sqrt{\int_t^{t'} (w(t')^T w(\tau))^2 d\tau} \\
&\le \sqrt{\int_t^{\delta} \epsilon_d(\tau)^2 d\tau}\sqrt{\int_t^{\delta} (w(t')^T w(\tau))^2 d\tau} \\
&\le (\|\epsilon_d(\cdot)\|_H)\,(\sqrt{\alpha_2}\|w(t')\|) \\
&\le \sqrt{\alpha_2}\|\epsilon_d(\cdot)\|_H \left(\sum_k \|w_k(x)\|_{\infty,\Omega}\right),
\end{aligned}$$

where $\|w(x)\|_{\infty,\Omega}$ is finite because the $w_k(x)$'s are continuous on the compact set Ω. From this inequality, we easily get an upper bound on $\|\zeta(\cdot)\|_H$:

$$\|\zeta(\cdot)\|_H \leq \sqrt{\alpha_2\delta}\|\epsilon_d(\cdot)\|_H \left(\sum_k \|w_k(x)\|_{\infty,\Omega}\right). \tag{2.20}$$

Combining (2.18), (2.19), and (2.20) yields

$$\begin{aligned}\|\phi(t)\| \leq & \frac{2\Phi\sqrt{\delta}+[1+\eta\sqrt{\alpha_2\delta}(\sum_k \|w_k(x)\|_{\infty,\Omega})]\|\epsilon_d(\cdot)\|_H}{\sqrt{\alpha_1}} \\ \leq & \frac{2\Phi\sqrt{\delta}+[1+\eta\sqrt{\alpha_2\delta}(\sum_k \|w_k(x)\|_{\infty,\Omega})]\sqrt{\int_t^{t+\delta}\epsilon_d(\tau)^2 d\tau}}{\sqrt{\alpha_1}}.\end{aligned} \tag{2.21}$$

Since, as we mentioned above, $\epsilon_d \in L^2[0,\infty]$, we have that $\int_t^{t+\delta}\epsilon_d(\tau)^2 d\tau \to 0$ as $t \to \infty$. Using this fact and taking the lim sup as $t \to \infty$ in (2.21), we arrive at (2.14). The discrete case differs only in minor ways from the continuous one, so the proof in that case will be omitted. □

In the continuous case, the convergence to such a neighborhood will happen at an exponential rate, at least in the time-regime during which the dead-zone prediction error $\epsilon_d(t)$ is strictly bounded away from 0. The lemmas that follow are aimed toward showing this.

LEMMA 2.3. *If $\epsilon_d(t) \neq 0$, then the function*

$$\sigma(t) := \frac{\epsilon_d(t)}{w(t)^T\phi(t)} \tag{2.22}$$

satisfies the inequality

$$1 \geq \sigma(t) \geq \frac{|\epsilon_d(t)|}{|\epsilon_d(t)|+2\Phi}. \tag{2.23}$$

In particular, if $|\epsilon_d(t)| \geq \Phi/\kappa$, where κ is a positive integer, then

$$1 \geq \sigma(t) \geq \frac{1}{2\kappa+1}. \tag{2.24}$$

Proof. From (2.8) and (2.9), it follows that when $\epsilon_d(t) \neq 0$, we have $|\epsilon(t)| > \Phi$, $\operatorname{sgn}(\epsilon_d) = \operatorname{sgn}(\epsilon)$, and

$$w(t)^T\phi(t) = \epsilon_d(t) + \Delta(t) + \Phi\operatorname{sgn}(\epsilon_d(t)). \tag{2.25}$$

This equation and the inequality $|\Delta(t)| \leq \Phi$ imply that

$$\operatorname{sgn}(\epsilon_d(t))w(t)^T\phi(t) = |\epsilon_d(t)| + (\Phi + \Delta(t)\operatorname{sgn}(\epsilon_d(t))) \geq |\epsilon_d(t)| > 0,$$

so that $w(t)^T\phi(t) \neq 0$ and $\operatorname{sgn}(w(t)^T\phi(t)) = \operatorname{sgn}(\epsilon_d(t))$. Thus, the function $\sigma(t)$ defined in (2.22) is finite and positive. Divide (2.25) by $w(t)^T\phi(t)$ and manipulate the resulting equation as follows:

$$\begin{aligned} 1 =& \sigma(t) + \frac{\Delta(t)+\Phi\operatorname{sgn}(\epsilon_d(t))}{\epsilon_d(t)}\sigma(t), \\ 1 =& \left(1+\frac{\Delta(t)\operatorname{sgn}(\epsilon_d(t))+\Phi\operatorname{sgn}^2(\epsilon_d(t))}{\epsilon_d(t)\operatorname{sgn}(\epsilon_d(t))}\right)\sigma(t), \\ 1 =& \left(1+\frac{\Phi+\Delta(t)\operatorname{sgn}(\epsilon_d(t))}{|\epsilon_d(t)|}\right)\sigma(t).\end{aligned}$$

Using the last equation and the inequality $\Phi \geq |\Delta(t)|$, we easily get (2.23). Inequality (2.24) is a direct consequence of the fact that the term on the right-hand side in (2.23) decreases with decreasing $|\epsilon_d|$. □

As long as we are working with a time interval, starting at $t = 0$, during which $\epsilon_d \neq 0$, we may change the time variable to

$$\tau = \int_0^t \sigma(s)ds. \tag{2.26}$$

The next lemma addresses the persistency of excitation for the regressor vector $w(t(\tau))$.

LEMMA 2.4. *Let $w(t)$ be persistently exciting, so that* (2.3) *holds, and let α_1, α_2, and δ be as in* (2.3). *If $t(\tau_0+\delta)$ is in a time interval $[0, t_\kappa]$ during which $|\epsilon_d(t)| \geq \Phi/\kappa$, then $w(t(\tau))$ satisfies*

$$\underbrace{(2\kappa+1)^{-1}\alpha_1}_{\alpha_1'} \|c\|^2 \leq \int_{\tau_0}^{\tau_0+\delta} |w(t(\tau'))^T c|^2 d\tau' \leq \underbrace{(2\kappa+1)\alpha_2}_{\alpha_2'} \|c\|^2. \tag{2.27}$$

Proof. Suppose that $t_0 \leq t_1 \leq t_\kappa$. Making the change of variables $\tau \longrightarrow t$ in the integral $\int_{\tau_0}^{\tau_1} |w(t(\tau'))^T c|^2 d\tau'$ and using (2.24) in the resulting integral yields

$$(2\kappa+1)^{-1} \int_{t_0}^{t_1} |w(t)^T c|^2 dt \leq \int_{\tau(t_0)}^{\tau(t_1)} |w(t(\tau'))^T c|^2 d\tau' \leq \int_{t_0}^{t_1} |w(t)^T c|^2 dt. \tag{2.28}$$

On the other hand, from (2.24) and (2.26), we see that

$$\frac{1}{2\kappa+1}(t_1 - t_0) \leq \tau(t_1) - \tau(t_0) = \int_{t_0}^{t_1} \sigma(s)ds \leq t_1 - t_0,$$

and so

$$\tau(t_1) - \tau(t_0) \leq t_1 - t_0 \leq (2\kappa+1)(\tau(t_1) - \tau(t_0)). \tag{2.29}$$

Note that the last equation implies that if $\tau(t_1) - \tau(t_0) = \delta$, then

$$\delta \leq t_1 - t_0 \leq (2\kappa+1)\delta.$$

The desired inequality (2.27) is a straightforward consequence of the last inequality, (2.28), and (2.3). □

If we make the change $t \longrightarrow \tau$ in the evolution equation (2.10), then we get a new evolution equation,

$$\frac{d\phi}{d\tau} = -\eta w(t(\tau))w(t(\tau))^T \phi, \tag{2.30}$$

which is prototypical of equations that are guaranteed to be exponentially stable with exponential decay in the norm of the parameter error vector, provided the regressor vector is persistently excited [25, p. 75]. In the case we are dealing with here, equation (2.30) is only valid for τ's that are bounded by $\tau_\kappa = \int_0^{t_\kappa} \sigma(t)dt$. If this is taken into account, one easily sees that the analysis given in [25, §2.5] applies as long as δ is less than τ_κ. More precisely, we have the following result.

PROPOSITION 2.5. *Let α_1, α_2, and δ be as in (2.3), and let t_κ be as in Lemma 2.4. If $\delta < \int_0^{t_\kappa} |\epsilon_d(s)|(|\epsilon_d(s)| + \Phi)^{-1} ds$, then for all $t \leq t_k$ the norm of the parameter error vector $\phi(t)$ satisfies*

$$\|\phi(t)\| \leq \exp\left\{-\alpha_\kappa \left(\int_0^t \frac{|\epsilon_d(s)|}{|\epsilon_d(s)| + \Phi} ds - \delta\right)\right\} \|\phi(0)\|. \tag{2.31}$$

The rate α_κ is given by

$$\alpha_\kappa = \frac{1}{2\delta} \ln \left[\frac{1}{1 - \dfrac{2\eta(2\kappa+1)^{-1}\alpha_1}{(1+\sqrt{s}\,\eta(2\kappa+1)\alpha_2)^2}}\right]. \tag{2.32}$$

Proof. If $\delta < \int_0^{t_\kappa} |\epsilon_d(s)|(|\epsilon_d(s)|+\Phi)^{-1} ds$, then (2.23) implies $\delta < \int_0^{t_\kappa} \sigma(s) ds =: \tau_k$. The argument used to prove [25, Thm. 2.5.3, p. 75] applies to the evolution equation (2.30), provided that $\tau < \tau_\kappa$. Combining [25, Thm. 1.5.2, p. 33] with [25, Thm. 2.5.3, p. 75], with the constants α_1 and α_2 in [25, Eq. 2.5.12] replaced by α_1' and α_2' from Lemma 2.4, yields

$$\|\phi(t(\tau))\| \leq \exp\{-\alpha_\kappa(\tau - \delta)\}\|\phi(0)\|. \tag{2.33}$$

The inequality in (2.31) follows from (2.23) and the definition of $\tau(t)$ given in (2.26). □

In closing our discussion of the gradient-descent/dead-zone algorithm, we wish to point out that, in the discrete case, a similar algorithm was analyzed by N. Sadegh [21], who also obtained an exponential rate of convergence for the associated parameters when the regressor vector is persistently exciting.

One unique feature of the identification algorithms presented above is the central role played by the persistency of excitation of the regressor vector. We now turn to an analysis of conditions sufficient to guarantee persistency of excitation when the basis $\mathfrak{W}$ used to construct the regressor vector w consists of RBFs centered at (possibly) scattered sites.

3. Sufficient conditions for persistency of excitation. We begin by stating a slightly broadened version of the definition of persistency of excitation, one that encompasses the discrete and continuous cases employed in the algorithms analyzed above. Let μ be a positive, Σ-finite Borel measure on $[0,\infty)$. We will say that a continuous, vector-valued function $w : [0,\infty) \longrightarrow \mathbb{R}^N$ is *persistently exciting* [25, p. 72] if there exist positive constants δ, α_1, and α_2 such that

$$\alpha_2\|c\|^2 \geq \int_{t_0}^{t_0+\delta} |w(\tau)^T c|^2 d\mu(\tau) \geq \alpha_1\|c\|^2$$

holds for every $t_0 \geq 0$ and every constant column vector $c = (c_1 \cdots c_N)^T \in \mathbb{R}^N$.

When RBFs are employed in solving the identification problem described in §2, the vector $w(t)$ has the form

$$w(t) = (F(\|x(t) - \xi_1\|) \cdots F(\|x(t) - \xi_N\|))^T, \tag{3.1}$$

where F is the RBF, each ξ_i is a given point in state space, termed a center, and $x(t)$ is the target trajectory. The function $x(t)$ is a continuous map from $[0,\infty)$ into $\mathbb{R}^s$, the state space. Furthermore, we assume that $x(t)$ remains in a bounded subset of $\mathbb{R}^s$. The N centers $\xi_1,\ldots,\xi_N$ are distinct points in $\mathbb{R}^s$. We wish to show that a sufficient condition for w of the form (3.1) to be persistently exciting amounts to an "ergodic" condition on $x(t)$, which we will precisely state in Theorem 3.5 below. We begin by proving three lemmas.

LEMMA 3.1. *Let $c \in \mathbb{R}^N$ and let $x \in \mathbb{R}^s$ be fixed. If $F(r)$ is an order-0 RBF of the form (1.1) or an order-1 RBF of the form (1.2), then*

$$\left|\sum_{j=1}^{N} F(\|x-\xi_j\|)c_j\right|^2 \le \left(\sum_{j=1}^{N} F(\|x-\xi_j\|)^2\right)\|c\|^2 \le \begin{cases} F(0)^2N\|c\|^2 & \text{for order } 0,\\ F(R)^2N\|c\|^2 & \text{for order } 1,\end{cases}$$

where in the order-1 case R is any number larger than the diameter of the set $\{x,\xi_1,\ldots,\xi_N\}$.

Proof. The proof of either inequality begins with an application of Schwarz's inequality. The top inequality then follows from the fact that an order-0 RBF attains its maximum at $r = 0$, and the bottom from the fact that an order-1 RBF is an increasing function of r. □

Remark 3.2. For many different order-0 RBFs, including the Gaussians, bounds not involving N are available. See Corollary 4.2 in §4 below.

LEMMA 3.3. *Let $x_i \in \mathbb{R}^s$ for $i = 1,\ldots,N$. If F is an RBF of the form (1.1) or (1.2), and if*

$$A = A(x_1,\ldots,x_N) = \begin{pmatrix} F(\|x_1-\xi_1\|) & \cdots & F(\|x_1-\xi_N\|)\\ \vdots & \ddots & \vdots\\ F(\|x_N-\xi_1\|) & \cdots & F(\|x_N-\xi_N\|)\end{pmatrix},$$

then there exists a number $\epsilon > 0$ and a number $\theta = \theta(\epsilon,\xi_1,\ldots,\xi_N) > 0$ such that

$$\|Ac\| \ge \theta\|c\|$$

for all $c \in \mathbb{R}^N$ and for every set of x_i's that satisfy $|x_i - \xi_i| \le \epsilon$ for $i = 1,\ldots,N$.

Proof. It is clear that θ^2 is a lower estimate for the smallest eigenvalue $\lambda(x_1,\ldots,x_N)$ of $A(x_1,\ldots,x_N)^T A(x_1,\ldots,x_N)$, whose components are obviously real, continuous functions of the x_i's. One can easily modify a theorem of Rellich [20, p. 40] to show that λ is also a continuous function of the x_i's. Moreover, $\lambda(\xi_1,\ldots,\xi_N) > 0$, because $A(\xi_1,\ldots,\xi_N)$ is invertible [12], [10], [26]. One may therefore choose $\epsilon > 0$ so that

$$\lambda(x_1,\ldots,x_N) > \frac{1}{2}\lambda(\xi_1,\ldots,\xi_N) > 0$$

whenever $|x_i - \xi_i| \le \epsilon$, $i = 1,\ldots,N$. Choosing $\theta = \sqrt{\frac{1}{2}\lambda(\xi_1,\ldots,\xi_N)}$ completes the proof. □

The lemma above is certainly not the best possible, because no estimates on the size of θ are given. In the next section, we shall discuss precise estimates for θ as a function of ϵ and the ξ_i's. For present purposes, this is not necessary.

We point out that if some choice of $\epsilon > 0$ works in Lemma 3.3, every smaller one will work as well. For purposes of the next lemma, we restrict our choice of ϵ so that

$$0 < \epsilon < q := \frac{1}{2} \min_{i \neq j} \|\xi_i - \xi_j\|. \tag{3.2}$$

Our next lemma is crucial to the proof of the main result.

LEMMA 3.4. *Let I be a bounded, μ-measurable subset of $[0,\infty)$, and also let the sets I_i be defined by*

$$I_i = \{t \in I \, : \, \|x(t) - \xi_i\| \leq \epsilon\} \quad \text{with } i = 1, \dots, N,$$

where ϵ is as in (3.2). If $\mu(I_i) \geq \tau_0$ for $i = 1, \dots, N$, then with $\theta > 0$ as in Lemma 3.3,

$$\int_I |w(\tau)^T c|^2 d\mu(\tau) \geq \tau_0 \theta^2 \|c\|^2 \tag{3.3}$$

holds for every constant column vector $c = (c_1 \cdots c_N)^T \in \mathbb{R}^N$.

Proof. The sets I_i are disjoint, because with $\epsilon < q = \frac{1}{2}\min_{i \neq j} \|\xi_i - \xi_j\|$ the balls with center ξ_i and radius ϵ are nonintersecting. Clearly, we thus have that

$$\int_I |w(\tau)^T c|^2 d\mu(\tau) \geq \sum_{i=1}^N \int_{I_i} |w(\tau)^T c|^2 d\mu(\tau). \tag{3.4}$$

Since we have that

$$|w(\tau)^T c|^2 = \left| \sum_{j=1}^N F(\|x(\tau) - \xi_j\|) c_j \right|^2,$$

and since $\tau \in I_i$ implies that $\|x(\tau) - \xi_i\| \leq \epsilon$, we immediately obtain the inequality

$$\begin{aligned} &\max_{|x - \xi_i| \leq \epsilon} \left\{ \left| \sum_{j=1}^N F(\|x - \xi_j\|) c_j \right|^2 \right\} \int_{I_i} d\mu(\tau) \\ &\geq \int_{I_i} |w(\tau)^T c|^2 d\mu(\tau) \geq \min_{|x - \xi_i| \leq \epsilon} \left\{ \left| \sum_{j=1}^N F(\|x - \xi_j\|) c_j \right|^2 \right\} \int_{I_i} d\mu(\tau), \end{aligned}$$

where the maximum and minimum are taken over all x in the ball $\|x - \xi_i\| \leq \epsilon$. The inequality above and the continuity of $|\sum_{j=1}^N F(\|x - \xi_j\|) c_j|^2$ over this compact and connected ball allow application of the intermediate value theorem, from which we may deduce that there exist vectors $x_i \in \mathbb{R}^n$ such that $\|x_i - \xi_i\| \leq \epsilon$ and

$$\int_{I_i} |w(\tau)^T c|^2 \, d\mu(\tau) = \left| \sum_{j=1}^N F(\|x_i - \xi_j\|) c_j \right|^2 \mu(I_i).$$

By assumption, $\mu(I_i) \geq \tau_0$ for $i = 1, \dots, N$, so

$$\int_{I_i} |w(\tau)^T c|^2 \, d\mu(\tau) \geq \left| \sum_{j=1}^N F(\|x_i - \xi_j\|) c_j \right|^2 \tau_0.$$

Inequality (3.4) and the last inequality imply that

$$\int_I |w(\tau)^T c|^2 \, d\mu(\tau) \geq \|Ac\|^2 \tau_0, \tag{3.5}$$

where A is an $N \times N$ matrix with i-j entry $A_{ij} = F(\|x_i - \xi_j\|)$. From the fact that $\|x_i - \xi_i\| \leq \epsilon$, where ϵ has been chosen to satisfy the conditions of Lemma 3.3, and from Lemma 3.3 itself, we have that

$$\|Ac\|^2 \geq \theta^2 \|c\|^2, \tag{3.6}$$

where θ depends upon ϵ and the ξ_i's, but not on the x_i's or c. Using (3.6) to replace the right-hand side of (3.5) yields the desired inequality, (3.3). □

THEOREM 3.5. *Let ϵ be as in Lemma* 3.3, *subject to restriction* (3.2). *In addition, for every $t_0 \geq 0$ and every $\delta > 0$, let*

$$I_i = \{t \in [t_0, t_0 + \delta] \, : \, \|x(t) - \xi_i\| \leq \epsilon\}.$$

If there exists a δ such that $\mu(I_i)$ is bounded below by a positive constant that is independent of t_0 and i, and if $\mu([t_0, t_0 + \delta]) \leq \delta$, then $w(t)$ is persistently exciting.

Proof of Theorem 3.5. Take the interval $I = [t_0, t_0 + \delta]$, and choose ϵ as in (3.2) and Lemma 3.3. By the hypotheses of the theorem, the sets I_i satisfy $\mu(I_i) \geq \tau_0 > 0$, where τ_0 depends upon $\delta, \epsilon, \xi_1, \ldots, \xi_N$, but does not depend on t_0 or i. Consequently, we may apply Lemma 3.3 to get

$$\int_{t_0}^{t_0+\delta} |w(\tau)^T c|^2 \, d\mu(\tau) \geq \alpha_1 \|c\|^2 \quad \text{with } \alpha_1 = \theta^2 \tau_0. \tag{3.7}$$

On the other hand, Lemma 3.1 implies that

$$\int_{t_0}^{t_0+\delta} |w(\tau)^T c|^2 d\mu(\tau) \leq \alpha_2 \|c\|^2 \quad \text{with } \alpha_2 = \begin{cases} F(0)^2 \delta N & \text{for order 0,} \\ F(R)^2 \delta N & \text{for order 1.} \end{cases} \tag{3.8}$$

Here, R is the diameter of the set $\{x(t) \, : \, t \in [0, \infty)\} \cup \{\xi_1, \ldots, \xi_N\}$. (This is a finite number, because the trajectory $x(t)$ is bounded.) Since both α_1 and α_2 are independent of t_0, $w(t)$ is persistently exciting. □

As we mentioned in §1, for the regressor vector to be persistently exciting, Theorem 3.5 requires a kind of ergodic condition. We wish to briefly discuss some consequences of Theorem 3.5 for the continuous and discrete algorithms mentioned in §2.

Continuous algorithms. The theorem essentially requires that in each time interval $[t_0, t_0 + \delta]$, the vector $x(t)$ must visit a sufficiently small ϵ-ball about each ξ_i for a minimum amount of time that is independent of t_0. As a simple example of how these conditions might be met, suppose that $x(t)$ is periodic with period T, and μ is the Lebesgue measure. If in the time interval $[0, T]$ the trajectory $x(t)$ stays within a distance ϵ of each center ξ_i for an amount of time that is at least $\tau_0 > 0$, then $w(t)$ is persistently exciting.

Discrete algorithms and neural networks. The theory derived above yields a strengthened version of certain results of Sanner and Slotine [23] concerning the persistency of excitation of the sequence of hidden layer outputs of a neural network.

The regressor vector $w(\tau)$ defined in (3.1) is called $g(\tau)$ in [23]; the function $F(\|x(\tau) - \xi_j\|)$ represents the jth hidden layer node. In [23], it was shown that by choosing the input function $x : [0, \infty) \to \mathbb{R}^s$ so that for any integer t, $\{x(\tau)\}_{\tau=t}^{\tau=N+t-1} = \{\xi_j\}_1^N$ (equality of sets), the corresponding sequence of hidden layer outputs will be persistently exciting.

To extend these results, first recall that the Hausdorff distance between two compact sets A and B in $\mathbb{R}^s$ is given by

$$d(A, B) = \max_{x \in A, y \in B} \{\operatorname{dist}(x, B),\ \operatorname{dist}(y, A)\},$$

where $\operatorname{dist}(x, B) = \inf_{z \in B} \|x - z\|$. Suppose that the sequence of network inputs satisfies an ergodic condition, namely, the input function $x(\tau)$ should have the property that for any integer t,

$$d(\{x(\tau)\}_{\tau=t}^{\tau=t+N-1}, \{\xi_j\}_1^N) < \epsilon.$$

Using Theorem 3.5 and taking the ϵ there to be the one used in the inequality above, one obtains that the system is still persistently exciting.

4. Estimates on parameters. What the identification algorithms discussed in §2 converge to, and at what rate they converge, depend quantitatively on the parameters δ, α_1, and α_2 that appear in the inequality defining persistency of excitation, (2.3), and indirectly on τ_0 from Lemma 3.4. The parameter δ is a "recurrence time." τ_0 represents the minimum time spent in an ϵ-ball about each of the centers, and is dependent on the system one is working with. On the other hand, from the proof of Theorem 3.5, one sees that the two ratios α_1/τ_0 and α_2/δ are independent of the behavior of $x(t)$; indeed, using the best bounds one can get from Lemmas 3.1–3.4, they have the following form:

$$\frac{\alpha_1}{\tau_0} = \theta(\epsilon, \xi_1, \dots, \xi_N)^2, \tag{4.1}$$

$$\frac{\alpha_2}{\delta} = \sup_{x \in \Omega} \left(\sum_{j=1}^{N} F(\|x - \xi_j\|)^2 \right). \tag{4.2}$$

In this section, we wish to estimate these ratios.

Let us begin with the problem of estimating α_1/τ_0. This problem hinges on the problem mentioned after Lemma 3.3—estimating θ as a function of ϵ and the ξ_i's. Fortunately, this problem was treated in detail in [27]. Since the technicalities involved in it are formidable, we will confine our discussion to the case of Gaussian RBFs,

$$F_\rho(r) = e^{-\rho r^2}, \tag{4.3}$$

where $\rho > 0$ is a parameter that classifies the choice of Gaussian. Concerning these RBFs, the following result holds.

THEOREM 4.1. *Let $\rho > 0$ be fixed, and let s and q be as in* §3; *set $z = \rho q^2$. For the Gaussian* RBF *F_ρ given in* (4.3), *we may take as a lower bound in Lemma* 3.3

$$\theta = \frac{1}{2} C_s z^{-s/2} e^{-\sigma^2/z}, \tag{4.4}$$

where C_s and σ are constants given by

$$\sigma := 12 \left(\frac{\pi \Gamma^2(\frac{s+2}{2})}{9} \right)^{1/(s+1)} \quad \textit{and} \quad C_s := \frac{\sigma^s}{2^{s+1} \Gamma(\frac{s+2}{2})},$$

provided $\epsilon > 0$ *in Lemma* 3.3 *is chosen to satisfy*

$$\epsilon \le \min\left\{q, \frac{q}{2}\frac{C_s e^{-\sigma^2/z}}{z^{1+s/2}}[1+6sf_s(z)]\right\}, \quad \text{where } f_s(z) := \sum_{j=0}^{\infty}(j+3)^s e^{-zj^2}. \tag{4.5}$$

Proof. We will adopt the notation used in Lemma 3.3, and we will follow the pattern of proof used to establish similar results in [27]. To begin, split $A(x_1,\dots,x_N)$ into its symmetric and antisymmetric parts, $A = A_{\text{sym}} + A_{\text{anti}}$. If $c \in \mathbb{R}^N$, then, from standard matrix theory, we have $c^TAc = c^TA_{\text{sym}}c$. Applying the Cauchy–Schwarz inequality results in $\|Ac\|\|c\| \ge c^TA_{\text{sym}}c$. Because the entries of A are real, the minimum of $\|Ac\|/\|c\|$ occurs for $c \in \mathbb{R}^N$. Hence, we obtain the inequality

$$\frac{\|A(x_1,\dots,x_N)c\|}{\|c\|} \ge \min_{\|\alpha\|=1} \alpha^T A_{\text{sym}}(x_1,\dots,x_N)\alpha, \tag{4.6}$$

and our task is reduced to estimating the symmetric quadratic form on the right above.

To estimate this quadratic form, write it as

$$\begin{aligned}\alpha^T A_{\text{sym}}(x_1,\dots,x_N)\alpha =&\alpha^T A_{\text{sym}}(\xi_1,\dots,\xi_N)\alpha\\ &+ [\alpha^T A_{\text{sym}}(x_1,\dots,x_N)\alpha - \alpha^T A_{\text{sym}}(\xi_1,\dots,\xi_N)\alpha]\\ =&I_\rho + J_\rho,\end{aligned}$$

where I_ρ and J_ρ denote the obvious quadratic forms, with the many arguments involved suppressed. From Proposition 3.6(iv) in [27], we have that

$$I_\rho \ge C_s\rho^{-s/2}q^{-s}e^{-\sigma^2q^{-2}\rho^{-1}} = C_s z^{-s/2}e^{-\sigma^2/z}.$$

An estimate for $|J_\rho|$ is given in Proposition 3.12 in [27], provided $\epsilon < q$. The estimate in our setting (where the quantity $\|d_0\|$ in [27] is ϵ here, and the incorrect factor π^s appearing in [27] is omitted here) becomes

$$|J_\rho| \le \rho q\epsilon[1+6sf_s(\rho q^2)] = \frac{z\epsilon}{q}[1+6sf_s(z)], \quad \text{where } f_s(z) := \sum_{j=0}^{\infty}(j+3)^s e^{-zj^2}.$$

Putting the various estimates together yields

$$\begin{aligned}\min_{\|\alpha\|=1} \alpha^T A_{\text{sym}}(x_1,\dots,x_N)\alpha &\ge I_\rho - |J_\rho|\\ &\ge C_s z^{-s/2}e^{-\sigma^2/z} - \frac{z\epsilon}{q}[1+6sf_s(z)].\end{aligned}$$

Picking ϵ so that it satisfies (4.5) results in

$$\min_{\|\alpha\|=1} \alpha^T A_{\text{sym}}(x_1,\dots,x_N)\alpha \ge \frac{1}{2}C_s z^{-s/2}e^{-\sigma^2/z}.$$

The theorem then follows on combining our last estimate with (4.6). □

We now turn to estimates for the ratio α_2/δ. When the RBF is a Gaussian F_ρ, the estimate for α_2/δ used in Theorem 3.5 can be made independent of N. To see why, note that from (4.2) it suffices to show that the vector $w_\rho(x) = (F_\rho(x-\xi_1)\cdots F_\rho(x-\xi_N))^T$ has length independent of x and N. Set

$$S_k(x) := \{\xi_j \mid kq \le \|x-\xi_j\| \le (k+1)q\}.$$

Observe that the cardinality of S_k cannot be any more than the ratio of the volume of the spherical shell with radii $(k-1)q$ and $(k+2)q$ to the volume of a sphere of radius q, because $\|\xi_j - \xi_k\| \ge 2q$. It is easy to see that

$$\operatorname{card}(S_k) \le (k+2)^s - (k-1)^s \le 3s(k+2)^{s-1},$$

where the inequality on the right follows from applying the mean value theorem. In addition, we obviously have

$$\bigcup_{k=0}^{\infty} S_k \supset \{\xi_1, \dots, \xi_N\}.$$

Furthermore, note that if $\xi_j \in S_k$, then

$$F_\rho(\|x-\xi_j\|) = e^{-\rho\|x-\xi_j\|^2} \le e^{-\rho(kq)^2}.$$

We thus have
(4.7)

$$\|w_\rho(x)\|^2 = \sum_{j=1}^{N} F_\rho^2(\|x-\xi_j\|) \le \sum_{k=0}^{\infty} \operatorname{card}(S_k) e^{-2\rho q^2 k^2} \le \sum_{k=0}^{\infty} 3s(k+2)^{s-1} e^{-2\rho q^2 k^2},$$

which gives us the following result.

COROLLARY 4.2. *If* $w_\rho(x) = (F_\rho(x-\xi_1)\cdots F_\rho(x-\xi_N))^T$, *then* $\|w_\rho(x)\|^2 \le 3sf_{s-1}(2\rho q^2)$, *where* f_s *is defined in* (4.5), *and the expression for* α_2/δ *in* (3.8) *may be replaced by*

$$\frac{\alpha_2}{\delta} = 3sf_{s-1}(2\rho q^2). \tag{4.8}$$

Proof. In the proof of Theorem 3.5, instead of using Lemma 3.1, use the expression for $\|w_\rho(x)\|^2$ given in (4.7). In addition, note that since $k+2 < k+3$,

$$\sum_{k=0}^{\infty} 3s(k+2)^{s-1} e^{-2\rho q^2 k^2} \le \sum_{k=0}^{\infty} 3s(k+3)^{s-1} e^{-2\rho q^2 k^2}.$$

Replacing the series on right side of this inequality by $f_{s-1}(2\rho q^2)$, where f_s is given in (4.5), completes the proof. □

Remark 4.3. For many order-0 RBFs, one can use Corollary 4.2 to obtain an upper estimate on α_2. To see this, note that from (1.1) and the expression for w_ρ given above, we have

$$w(x) = (\,F(x-\xi_1) \quad \cdots \quad F(x-\xi_N)\,)^T = \int_0^\infty w_\rho(x) d\beta(\rho).$$

Obviously, from Corollary 4.2, we also have

$$\|w(x)\| \le \int_0^\infty \|w_\rho(x)\| d\beta(\rho) \le \int_0^\infty \sqrt{3sf_{s-1}(2\rho q^2)}\, d\beta(\rho),$$

and so we arrive at

$$\alpha_2 \leq \left(\int_0^\infty \sqrt{3sf_{s-1}(2\rho q^2)}\, d\beta(\rho) \right)^2 \delta, \tag{4.9}$$

which will hold whenever the right side above is finite. Similar, somewhat more complicated results can be derived for α_1; we will not state them here.

Acknowledgments. The authors thank B. J. C. Baxter and the referee for helpful suggestions.

REFERENCES

[1] D. S. Broomhead and D. Lowe, *Multivariable functional interpolation and adaptive networks*, Complex Systems, 2 (1988), pp. 321–355.

[2] J. Duchon, *Splines minimizing rotation invariant semi-norms in Sobolev spaces*, in Constructive Theory of Functions of Several Variables, Lecture Notes in Mathematics 571, W. Schempp and K. Zeller, eds., Springer-Verlag, Berlin, New York, 1977, pp. 85–100.

[3] N. Dyn, *Interpolation and approximation by radial and related functions*, in Approximation Theory VI: Vol. I, C. K. Chui, L. L. Schumaker, and J. D. Ward, eds., Academic Press, New York, 1989.

[4] N. Dyn, D. Levin, and S. Rippa, *Numerical procedures for global surface fitting of scattered data by radial functions*, SIAM J. Sci. Statist. Comput., 7 (1986), pp. 639–659.

[5] R. Franke, *Scattered data interpolation: Tests of some methods,* Math. Comp., 38 (1982), pp. 181–199.

[6] R. L. Hardy, *Multiquadric equations of topography and other irregular surfaces*, J. Geophys. Res., 76 (1971), pp. 1905–1915.

[7] ———, *Theory and applications of the multiquadric-biharmonic method,* Comput. Math. Appl., 19 (1990), pp. 163–208.

[8] A. J. Kurdila, *Convergence of Hamiltonian control systems: Adaptive tracking control via radaial basis functions*, preprint.

[9] J. G. McWhirter, D. S. Broomhead, and T. J. Shepard, *A systolic array for nonlinear adaptive filtering and pattern recognition,* J. VLSI Signal Process., 3 (1991), pp. 69–75.

[10] W. R. Madych and S. A. Nelson, *Multivariate interpolation and conditionally positive definite functions*, Approx. Theory Appl., 4 (1988), pp. 77–79.

[11] ———, *Multivariate interpolation and conditionally positive definite functions* II, Math. Comp., 54 (1990), pp. 211–230.

[12] C. A. Micchelli, *Interpolation of scattered data: Distances, matrices, and conditionally positive definite functions*, Constr. Approx., 2 (1986), pp. 11–22.

[13] F. J. Narcowich and J. D. Ward, *Norms of inverses and condition numbers for matrices associated with scattered data*, J. Approx. Theory, 64 (1991), pp. 69–94.

[14] ———, *Norms of inverses for matrices associated with scattered data*, in Curves and Surfaces, P. J. Laurent, A. Le Méhauté, and L. L. Schumaker, eds., Academic Press, Boston, 1991.

[15] ———, *Norm estimates for the inverses of a general class of scattered-data radial-function interpolation matrices*, J. Approx. Theory, 69 (1992), pp. 84–109.

[16] T. Poggio and F. Girosi, *A theory of networks for approximating and learning*, A. I. Memo No. 1140, Artificial Intelligence Laboratory, Massachusetts Institute of Technology, Cambridge, MA, 1989.

[17] M. J. D. Powell, *Radial basis functions for multivariable approximation*, in Algorithms for Approximation, J. C. Mason and M. G. Cox, eds., Oxford University Press, Oxford, 1987.

[18] ———, *The theory of radial basis approximation in* 1990, in Wavelets, Subdivision and Radial Functions, W. Light, ed., Oxford University Press, 1990.

[19] E. Quak, N. Sivakumar, and J. D. Ward, *Least squares approximation with radial functions*, SIAM J. Math. Anal., 24 (1993), pp. 1043–1066.

[20] F. Rellich, *Perturbation Theory of Eigenvalue Problems*, Gordon and Breach, New York, 1969.

[21] N. Sadegh, *Nonlinear identification and control via neural networks*, in Control of Systems with Inexact Dynamic Models, DSC-Vol. 33, N. Sadegh and Y.-H. Chen, eds., American Society of Mechanical Engineers, New York, 1991.

[22] R. M. Sanner and J.-J. E. Slotine, *Gaussian networks for direct adaptive control,* in Proc. 1991 American Control Conference, Vol. 3, Boston, MA, IEEE 1991, pp. 2153–2157.

[23] ———, *Stable recursive identification using radial basis function networks*, in the Proc. 1992 American Control Conference, Vol. 3, Chicago, IL, IEEE 1992, pp. 1829–1833.

[24] ———, *Gaussian networks for direct adaptive control*, IEEE Trans. Neural Networks, 3 (1992), pp. 837–863.

[25] S. Sastry and M. Bodson, *Adaptive Control: Stability, Convergence, and Robustness*, Prentice–Hall, Englewood Cliffs, NJ, 1989.

[26] I. J. Schoenberg, *Metric spaces and completely monotone functions*, Ann. of Math., 39 (1938), pp. 811–841.

[27] N. Sivakumar and J. D. Ward, *On the best least squares fit by radial functions to multidimensional scattered data*, Numer. Math., 65 (1993), pp. 219–243.

[28] E. Tzirkel-Hancock and F. Fallside, *Stable control of nonlinear systems using neural networks*, Tech. report CUED/F-INENG/TR.81, Cambridge University Engineering Department, Cambridge, England, 1991.

[29] D. V. Widder, *The Laplace Transform,* Princeton University Press, Princeton, 1941.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 33, No. 2, pp. 643–666, March 1995

© 1995 Society for Industrial and Applied Mathematics

015

MAXIMIZING ROBUSTNESS IN NONLINEAR ILLPOSED INVERSE PROBLEMS*

KAZUFUMI ITO† AND KARL KUNISCH‡

Abstract. A framework for maximizing the robustness of nonlinear illposed inverse problems by choosing appropriate system inputs is presented. This framework is based on maximizing the lowest eigenvalues of the linearized and regularized nonlinear mapping. Stability and sensitivity of the eigenvalues of the linearization are studied. The results are applied to parameter estimation problems for elliptic partial differential equations. Numerical examples illustrating the results are given.

Key words. nonlinear illposed inverse problems, optimal input, sensitivity measures, regularization

AMS subject classifications. 49B40, 35R25

1. Introduction.

Let Φ_f be a family of nonlinear mappings from a subset of a Hilbert space X into a Hilbert space Y. The problem under investigation in this paper is the choice of an optimal f such that the inversion of Φ_f becomes as stable as possible. The mapping $a \to \Phi_f(a)$ can typically arise as the solution mapping for a partial differential equation with a representing a coefficient or an inhomogeneity of the differential equation. The functional parameter f represents a design parameter that is chosen from a class of admissible input functions. We then study a problem of optimal experimental design: Which is the best choice for the input f for the reconstruction of the coefficient a given knowledge of $\Phi_f(a)$, the state of the differential equation.

A criterion for the selection of the optimal parameter f must be specified. Here we proceed as follows. Let $\Phi'_f(a)$ be the linearization of Φ_f at some reference point $a \in X$; and consider

$$\sup_{f \in \mathcal{F}} \inf_{\substack{h \in V \\ h \neq 0}} \frac{|\Phi'_f(a)h|_Y^2}{|h|_X^2}, \tag{1.1}$$

where $V \subset X$ is a Hilbert space continuously embedded in X. Thus we propose to maximize the lowest singular value of $\Phi'_f(a)$ (considered as unbounded operator between X and Y) as f varies in F. The saddle point problem (1.1) is the focus of attention in this paper.

Let us turn to the infimum problem

$$\inf_{\substack{h \in V \\ h \neq 0}} \frac{|\Phi'_f(a)h|_Y^2}{|h|_X^2} \tag{1.2}$$

in (1.1). Since the focuses of attention are mappings Φ_f that are illposed in the sense of lack of continuous invertibility, one has to guard against the infimum being zero for all $f \in \mathcal{F}$. First, of course, this requires injectivity of $\Phi'_f(a)$ for at least one $f \in \mathcal{F}$.

* Received by the editors May 26, 1992; accepted for publication October 25, 1993.

† The research of this author was supported in part by Air Force Office of Scientific Research grant AFOSR-90-0091 and National Science Foundation grant UINT-8521208.

‡ This author was supported in part by the Fonds zur Förderung der wissenschaftlichen Forschung, Austria, under P-8146-PHY.

Second, for (1.2) to be nonzero an estimate of the type

$$(1.3)\qquad |\Phi'_f(a)h|_Y \geq k_f |h|_X, \quad \text{for all } h \in V,$$

is necessary with k_f a positive constant, possibly depending on f. For illposed problems (1.3) is not feasible and the infimum in (1.2), while it may not be attained, will be zero. To overcome this difficulty we need to change the problem formulation in such a way that an estimate of the type (1.3) holds while simultaneously the underlying structure of the inverse problem remains as unchanged as possible.

To be specific, let us consider the problem of estimating the functional coefficient a in

$$(1.4)\qquad \begin{cases} -\text{div}(a \text{ grad } u) &= f \quad \text{in } \Omega, \\ u|\partial\Omega &= 0 \end{cases}$$

given knowledge of $u(a)$, where Ω is a bounded domain in $\mathbf{R}^n, n = 1, 2,$ or 3. In this case $\Phi_f(a) = u(a)$ and the linearization of Φ_f is given by

$$\Phi'_f(a)h = A^{-1} \text{ div}(h \text{ grad } u(a)),$$

where $A(a) : H_0^1(\Omega) \to H^{-1}(\Omega)$ is defined by $A(a)u = -\text{div}(a \text{ grad } u), a(x) \geq \nu > 0$. For the choice $X = Y = L^2(\Omega)$ an estimate of the type (1.3) is impossible. If X and Y are chosen as $L^2(\Omega)$ and $H^1(\Omega)$ respectively and if f satisfies additional assumptions, then for every $K > 0$ there exists $k_f > 0$ such that

$$(1.5)\qquad |\Phi'_f(a)h|_{H^1} \geq k_f |h|^2_{L^2} \quad \text{for all } h \in L^2(\Omega), \text{ with } |h|_{H^1} \leq K$$

[IK1]. While this estimate does not give (1.3) with X, Y replaced by $L^2(\Omega)$ and $H^1(\Omega)$, it indicates what can be achieved by changing the norm in image and preimage space. In passing let us note that in the one-dimensional case (1.5) holds with the constraint $|h|_{H^1} \leq K$ replaced by the assumption that h lies in a subspace of L^2 with codimension 1. To obtain an estimate of the type (1.3) we need to make an additional change in the problem formulation by using a regularization term.

Returning to the general formulation, (1.2) is replaced by

$$(1.6)\qquad \inf_{h\neq 0} \frac{|\Phi'_f(a)h|^2_Y + \beta\tilde{\sigma}(h,h)}{|h|^2_X}$$

where $\tilde{\sigma}$ is a nonnegative Hermitian form on $V \times V$ and β is a small, positive parameter. The original saddle point problem becomes

$$(1.7)\qquad \sup_{f\in\mathcal{F}} \inf_{h\neq 0} \frac{|\Phi'_f(a)h|^2_Y + \beta\tilde{\sigma}(h,h)}{|h|^2_X}.$$

We can now describe our proposal to solve the optimal design problem stated at the beginning of this section: First, find a problem formulation by restricting the class of design functions $\mathcal{F}$ and by appropriate choice of norms for the preimage and the image spaces X and Y of $\Phi'_f(a)$ to make this linearized problem as "well posed as possible." Second, use regularization, if necessary.

Let us turn to (1.7). First, this is a difficult saddle point problem for which questions of existence and differentiability of the infimum problem with respect to f need to be studied. This will be the focus of the present paper. Second, if regularization is used, the question arises whether the infimum is still sufficiently sensitive with respect to different choices of $f \in \mathcal{F}$. We have no analysis to answer this question, but numerical experiments with specific problems suggest a positive answer. Finally, in view of the fact that the mapping Φ_f can be nonlinear, (1.7) is not satisfactory in as far as it requires linearization about a reference parameter, which in practice must be chosen as good approximation to a^*, the unknown preimage of $\Phi_f(a^*)$ and updated by an appropriate iterative procedure.

Let us also mention that this paper is a continuation of our work on optimal design for inverse problems, with the first paper [IK2] focusing on the one-dimensional version of (1.4). By allowing only a restrictive class of inputs $\mathcal{F}$, the analysis of [IK2] does not necessitate the use of a regularization term and the infimum problem can be solved explicitly in specific situations, thus allowing a rather detailed analysis of (1.7) with $\beta = 0$.

While in our analysis we have chosen the lowest singular value of Φ'_f as a measure for distinguishing the behaviour of the inverse of Φ_f as f varies, other criteria may also be of interest. One of them would be the trace of a discretization of $\mathcal{T} = (\Phi'_f(a))^* \Phi'_f(a)$ or the sum of the eigenvalues of $\mathcal{T}$. Different choices will lead to different optimal inputs f. This can be seen, for example, from numerical experiments, where in some cases we calculated the complete spectrum of a discretization of $\mathcal{T}$. Minimizing the lowest singular value represents a conservative measure (robustness). For parameter estimation problems such as (1.4) it requires that the reconstruction of a^* from $u(a^*)$ is uniformly "as well posed as possible" throughout the domain Ω.

The paper is organized in the following way. Section 2 contains a study of the saddle point problem (1.7). In §3 we apply these results to the problem of determining a from $u(a)$, with u the solution of (1.4). Numerical results are given in §4.

2. The general framework. Let V and X be Hilbert spaces with V densely and compactly imbedded in X, such that

$$V \subset X \subset V^*$$

forms a Gelfand triple, with V^* denoting the antidual of V. The norm and the inner product on V are denoted by $\| \cdot \|$ and $\langle\langle \cdot, \cdot \rangle\rangle$ and similarly those on X by $| \cdot |$ and $\langle \cdot, \cdot \rangle$. Further for a bounded linear operator T from V into a Hilbert space Y, we shall write $T \in \mathcal{L}(V, Y)$. By $\sigma(\cdot, \cdot)$ we denote a continuous Hermitian form on $V \times V$. In particular $\sigma(x, y) = \overline{\sigma(y, x)}$ and there exist a constant $c > 0$ such that

$$|\sigma(x, y)| \leq c \, \| x \| \, \| y \| \, .$$

This implies the existence of an operator $\mathcal{N} \in \mathcal{L}(V, V^*)$ with

$$\sigma(x, y) = \langle \mathcal{N} x, y \rangle_{V^*, V} \quad \text{for all } x \text{ and } y \in V,$$

where $\langle \cdot, \cdot \rangle_{V^*, V}$ denotes the antiduality between V^* and V. Since $T \in \mathcal{L}(V, Y)$, it defines a continuous Hermitian form on $V \times V$ via $(x, y) \to \langle Tx, Ty \rangle_Y$.

The operator $\mathcal{T} \in \mathcal{L}(V, V^*)$ representing this form satisfies

$$\langle Tx, Ty \rangle_Y = \langle \mathcal{T} x, y \rangle_{V^*, V} \quad \text{for } x \text{ and } y \in V$$

and is given by $\mathcal{T} = T^* T$, where $T^* \in \mathcal{L}(Y, V^*)$ denotes the adjoint of T. We are now prepared to introduce the Hermitian form $a : V \times V \to C$, which will be the focus of this section. It is defined by

$$a(x, y) = \langle Tx, Ty \rangle_Y + \sigma(x, y),$$

and it will be assumed that

(H1) There exists $\mu > 0$ such that $a(x,x) = |Tx|_Y^2 + \sigma(x,x) \geq \mu \parallel x \parallel^2$ for all $x \in V$.

In the context of the introduction the present formalism will be used with $\mathcal{T} = \Phi'_f(a)$ and with the regularization term given by $\beta\tilde{\sigma}(x,x) = \sigma(x,x)$. Neither $\mathcal{T}$ nor σ separately are assumed to be coercive. Condition (H1) requires strict coercivity of their sum. For $(x,y) \in V \times V$ we have

$$a(x,y) = \langle(\mathcal{T} + \mathcal{N})x, y\rangle_{V^*,V},$$

and by the Lax-Milgram theorem $A := \mathcal{T} + \mathcal{N}$ is an isomorphism of V onto V^*. In the usual way [DL, II p. 369 and III p. 39], A can be considered a closed linear selfadjoint operator in X with domain $D_A = \{x \in X : Ax \in X\} = \{x \in V : y \to a(x,y)$ with $y \in V$, is continuous form X to $C\}$. Alternatively, a can be treated as a densely defined Hermitian closed form in $X \times X$.

Example. As an example for the above formalism we may consider the special case $X = Y = H^0(\Omega)$, $V = H^1(\Omega)$,

$$\sigma(x,y) = \langle\nabla x, \nabla y\rangle_{H^0},$$

where H^i denotes the Sobolev space of order i. If $T \in \mathcal{L}(H^1(\Omega), H^0(\Omega))$ satisfies

$$|Tx|_{H^0} \geq \bar{\mu}|x|_{H^0} \quad \text{for some } \bar{\mu} > 0 \text{ and all } x \in H^1(\Omega)$$

(or if only $|T1|_{L^2} \geq \bar{\mu}$ with 1 the constant function with value 1), then (H1) holds.

LEMMA 2.1. *Let* (H1) *hold. Then the spectrum of the operator A is point spectrum $\sigma_p(A)$ with real eigenvalues λ^i satisfying*

$$0 < \mu \leq \lambda^1 \leq \lambda^2 \leq \cdots,$$

and no finite point of accumulation. The eigenvectors x^n, normalized in X and associated with λ^n, satisfy

$$a(x^n, v) = \lambda^n\langle x^n, v\rangle \quad \textit{for all } v \in V.$$

The eigenvectors form an orthonormal basis for X. For a proof see [DL, III p. 39].

We next introduce a family of operators A_f, and we investigate continuity and differentiability results of their spectrum.

Let $\mathcal{F}$ be a metric space with metric ρ, and for each $f \in \mathcal{F}$ let $T_f \in \mathcal{L}(V,Y)$. Assume that

(H1*) There exists $\mu > 0$ such that $a_f(x,x) = |T_f(x)|_Y^2 + \sigma(x,x) \geq \mu \parallel x \parallel^2$ for all $x \in V$ and $f \in \mathcal{F}$.

We introduce the notation $A_f = T_f^* T_f + \mathcal{N} \in \mathcal{L}(V, V^*)$ and

$$a_f(x,y) = \langle T_f x, T_f y\rangle + \sigma(x,y).$$

Henceforth let $f_0 \in \mathcal{F}$, let $\{f_k\}_{k=1}^{\infty}$ be a sequence in $\mathcal{F}$ with $\lim_{k\to\infty} \rho(f_k, f_0) = 0$, let λ_k^n, x_k^n be the eigenvalues and eigenfunctions of A_{f_k} enumerated as in Lemma 2.1, and similarly let λ_0^n, x_0^n be the eigenvalues and eigenfuctions of A_{f_0}.

By P_k^n we denote the orthogonal (in X) projection from X into the eigenspace M_k^n associated with the eigenvalue λ_k^n. We shall require the following condition:

(H2) $\lim_{k\to\infty} T_{f_k} = T_{f_0}$ in $\mathcal{L}(V,Y)$.

PROPOSITION 2.2. *Let* $\lim_{k\to\infty} \rho(f_k, f_0) = 0$, *and let* (H1*) *and* (H2) *hold. Then*

(i) $\lim_{k\to\infty} \lambda_k^n = \lambda_0^n$ *for every* $n = 1, 2, \ldots.$
(ii) $\lim_{k\to\infty} P_k^n = P_0^n$ *in* $\mathcal{L}(X)$ *for every* $n = 1, 2, \ldots.$
(iii) $\dim M_k^n = \dim M_0^n$ *for every* $k = k(n)$ *sufficiently large and* $n = 1, 2, \ldots.$
(iv) *If* $\dim M_0^n$, *then*

$$\lim_{k\to\infty} x_k^n = x_0^n \quad \text{in } V,$$

where $x_k^n \in M_k^n$ *are eigenfunctions of* A_k^n *for the eigenvalue* λ_k^n *normalized by* $\langle x_0^n, x_k^n\rangle = 1$.

For the proof we refer to results in [K]; see also [C]. Specifically, the densely defined closed symmetric forms a_{f_k} are relatively bounded with respect to the form a_{f_0}. It follows that A_{f_k} converges in X to A_{f_0} in the generalized sense and any finite system of eigenvalues of A_{f_0} is stable [K, pp. 202, 212, 338–340]. In particular this implies (i)–(iii). Next assume that the multiplicity of the eigenvalue λ_0^n is 1. Then $\dim M_k^n = 1$ and $x_k^n = \alpha_k^{-1} P_k^n x_0^n$ for all k sufficiently large, where $\alpha_k = \langle P_k^n x_0^n, x_0^n\rangle$. Since $\lim_{k\to\infty} P_k^n x_0^n = x_0^n$ in X, it follows that $\lim_{k\to\infty} \alpha_k = 1$. We find

$$\lim_{k\to\infty} |x_k^n - x_0^n| = \lim_{k\to\infty} \left|\frac{1-\alpha_k}{\alpha_k} P_k^n x_0^n\right| + \lim_{k\to\infty} |P_k^n x_0^n - x_0^n| = 0$$

and thus $\lim_{k\to\infty} x_k^n = x_0^n$ in X. To show the final assertion of the theorem note that

$$a_{f_k}(x_k^n, x_k^n) = \lambda_k^n \langle x_k^n, x_k^n\rangle. \tag{2.1}$$

Due to (H1*) and (2.1) it follows that $\{x_k^n\}_{k=1}^{\infty}$ is bounded in V. Moreover, taking the limit in (2.1) and by (H2)

$$\begin{aligned}|a_{f_0}(x_0^n, x_0^n) - a_{f_0}(x_k^n, x_k^n)| \le & |a_{f_0}(x_0^n, x_0^n) - a_{f_k}(x_k^n, x_k^n)| \\ & + |a_{f_k}(x_k^n, x_k^n) - a_{f_0}(x_k^n, x_k^n)| \to 0 \quad \text{as } k \to \infty.\end{aligned}$$

Consequently $a_{f_0}(x_k^n, x_k^n) \to a_{f_0}(x_0^n, x_0^n)$ and $x_k^n \to x_0^n$ weakly in V as $k \to \infty$. Since $\sqrt{a_{f_0}(\cdot,\cdot)}$ defines an equivalent norm on V, it follows that $\lim_{k\to\infty} x_k^n = x_0^n$ in V.

Strenghtening (H2) to

(H3i) There exist constants K and $\gamma > 0$ such that $\| T_{f_k} - T_{f_0} \|_{\mathcal{L}(V,Y)} \le K\rho(f_k, f_0)^\gamma$, respectively,

(H3ii) There exist constants K and $\gamma > 0$ such that $|T_{f_k}x - T_{f_0}x|_Y \in K\rho(f_k, f_0)^\gamma |x|_X$ for all $x \in V$, allows us to assert

PROPOSITION 2.3. *Let* $\lim_{k\to\infty} \rho(f_k, f_0) = 0$, *and let* (H1*) *and* (H3i) *hold. Then for every* n *there exists* K_n *such that*

$$\| P_k^n - P_0^n \|_{\mathcal{L}(X)} \le K_n \rho(f_k, f_0)^\gamma$$

and

$$|\lambda_k^n - \lambda_0^n| \le K_n \rho(f_k, f_0)^\gamma.$$

If (H3ii) *holds in place of* (H3i), *then there exists* $\hat{K}$ *independent of* n *such that*

$$|\lambda_k^n - \lambda_0^n| \le \hat{K} \rho(f_k, f_0)^\gamma.$$

For the proof one notes that (H3i) implies the existence of a constant M such that

$$|(a_{f_0} - a_{f_k})(x,x)| \leq M\rho(f_k, f_0)^\gamma \parallel x \parallel^2 \leq \frac{M}{\mu}\rho(f_k, f_0)^\gamma a_0(x,x),$$

for all $x \in V$. The assertion under condition (H3i) then follows from the same results in [K] as cited above. The proof of convergence of the eigenvalues under condition (H3ii) can be based on an argument using the Raleigh quotient representation of eigenvalues or on [K, Thm. V. 4.10]

We turn next to investigating differentiability properties of the lowest eigenvalue–eigenvector pair at a reference value $f_0 \in F$, where F is now assumed to be a normed linear space with norm $|\cdot|_F$. We drop the superscript 1 for the largest eigenvalue–eigenvector pair and assume that the multiplicity of the smallest eigenvalue λ_{f_0} of A_{f_0} is one. Let x_f denote the eigenvectors associated with the lowest eigenvalue λ_f of A_f normalized by

$$\langle x_f, x_{f_0}\rangle = 1, \qquad |x_{f_0}| = 1.$$

For $f \in \mathcal{F}$ we put

$$\mathcal{T}_f = T_f^* T_f \in \mathcal{L}(V, V^*).$$

We shall employ the following hypotheses:

(H2*) The mapping $f \longrightarrow T_f$ from F to $\mathcal{L}(V,Y)$ is continuous at $f_0 \in \mathcal{F}$.

(H4) The mapping $f \longrightarrow \mathcal{T}_f$ from $\mathcal{F}$ into $\mathcal{L}(V,V^*)$ has a Gateaux differential at $f_0 \in \mathcal{F}$ in the direction $f_1 \in F$; i.e., $f_0 + \tau f_1 \in \mathcal{F}$ for $|\tau|$ sufficiently small and there exists $\dot{\mathcal{T}}(f_0, f_1) \in \mathcal{L}(V,V^*)$ such that

$$\dot{\mathcal{T}}(f_0, f_1) = \lim_{\tau \longrightarrow 0} \frac{1}{\tau}[\mathcal{T}_{f_0+\tau f_1} - \mathcal{T}_{f_0}] \quad \text{in } \mathcal{L}(V, V^*).$$

Concerning the terminology of differentiability in infinite-dimensional spaces we follow [W]. The proofs of the subsequent results of this section are given in the appendix.

THEOREM 2.4. *Assume that* (H1*), (H2*), *and* (H4) *hold and that the smallest eigenvalue associated with* A_{f_0} *is simple. Then the mapping* $f \longrightarrow (\lambda_f, x_f)$, *where* λ_f *is the smallest eigenvalue of* A_f *and* x_f *is the corresponding eigenfunction of* A_f *normalized by* $\langle x_{f_0}, x_f\rangle = 1$, *from* F *to* $\mathbf{R} \times V$ *has a Gateaux differential* $(\dot\lambda, \dot x) = (\dot\lambda(f_0, f_1), \dot x(f_0, f_1))$ *at* f_0 *in direction* f_1. *It is the unique solution of*

$$(2.2) \qquad \left\{ \begin{array}{rcl} (A_{f_0} - \lambda_{f_0})\dot x - \dot\lambda x_{f_0} & = & -\dot{\mathcal{T}}(f_0, f_1)x_{f_0}, \\ \langle \dot x, x_{f_0}\rangle_X & = & 0. \end{array} \right.$$

Stronger differentiability assumptions for $f \longrightarrow \mathcal{T}_f$ carry over to improved differentiability properties of $f \longrightarrow (\lambda_f, x_f)$. Throughout the remainder of this section we assume that f_0 is an interior point of $\mathcal{F}$. We shall use the following hypotheses:

(H5) $f \in \mathcal{F} \longrightarrow T_f \in \mathcal{L}(V,V^*)$ has a Gateaux differential at $f_0 \in \mathcal{F}$ in every direction $f_1 \in F$, and $f_1 \longrightarrow \dot{\mathcal{T}}(f_0, f_1)x_{f_0}$ from F to V^* is linear.

(H6) $f \in \mathcal{F} \subset F \longrightarrow \mathcal{T}_f \in \mathcal{L}(V,V^*)$ has a Fréchet derivative at f_0.

With these conditions we obtain the following.

COROLLARY 2.5.

(i) *Let* (H1*), (H2*), (H4), *and* (H5) *hold. Then* $f \to (\lambda_f, x_f)$ *from* $\mathcal{F}$ *to* $\mathbf{R} \times V$ *has a Gateaux derivative at* f_0.

(ii) *Let* (H1*), (H2*), *and* (H6) *hold. Then* $f \to (\lambda_f, x_f)$ *from* F *to* $\mathbf{R} \times V$ *has a Fréchet derivative at* f_0.

The differentiability properties of the higher eigenvalue–eigenvector pairs can be studied with no additional effort. We obtain the following corollary.

COROLLARY 2.6.

(i) *Assume that* (H1*), (H2*), *and* (H4) *hold and that the eigenvalues of* A_{f_0} *are simple. Then the mapping* $\mathcal{F}^n : f \to (\lambda^n_f, x^n_f)$, *where* $\langle x^n_f, x^n_{f_0}\rangle = 1$, *from* F *to* $\mathbf{R} \times V$ *has a Gateaux differential* $(\dot\lambda^n(f_0, f_1), \dot x^n(f_0, f_1))$ *at* f_0 *in direction* f_1. *It is the unique solution of*

$$\begin{aligned}(A_{f_0} - \lambda^n_{f_0})\dot x - \dot\lambda x^n_{f_0} &= -\dot{\mathcal{T}}(f_0, f_1)x^n_{f_0},\\ \langle \dot x, x^n_{f_0}\rangle &= 0.\end{aligned} \tag{2.3}$$

(ii) *If* (H1*), (H2*), (H4), *and* (H5) *hold, then* $\mathcal{F}^n$ *has a Gateaux derivative at* f_0. *If* (H1*), (H2*), *and* (H6) *hold, then* $\mathcal{F}^n$ *has a Fréchet derivative at* f_0.

In the final results of this section we give conditions on

$$f \to T_f \text{ from } \mathcal{F} \subset F \text{ to } \mathcal{L}(V, Y),$$

which imply (H5) and (H6). We shall use the following:

(A1) $f \to T_f$ is continuous and has a Gateaux derivative at the point f_0.

(A2) $f \to T_f$ has a Fréchet derivative depending continuously on f at the point f_0.

LEMMA 2.7. *If* (A1) *holds, then* $f \to \mathcal{T}_f$ *from* F *to* $\mathcal{L}(V, V^*)$ *has a Gateaux derivative at* f_0 *and*

$$\dot{\mathcal{T}}(f_0, f_1) = T^*_{f_0}\dot T(f_0, f_1) + \dot T^*(f_0, f_1)T_{f_0} \tag{2.4}$$

or equivalently

$$\begin{aligned}\langle \dot{\mathcal{T}}(f_0, f_1)\varphi, \psi\rangle_{V^*,V} =& \langle \dot T(f_0, f_1)\varphi, T_{f_0}\psi\rangle_Y\\ &+\langle T_{f_0}\varphi, \dot T(f_0, f_1)\psi\rangle_Y\end{aligned} \tag{2.5}$$

for all $f_1 \in F$ *and* $\varphi, \psi \in V$. *Here* $\dot T(f_0, f_1)$ *denotes the Gateaux differential of* T_f *at* f_0 *in direction* f_1. *In particular,* (H5) *holds.*

LEMMA 2.8. *If* (A2) *holds, then* $f \to \mathcal{T}_f$ *from* F *to* $\mathcal{L}(V, V^*)$ *has a Fréchet derivative at* f_0.

3. Optimal input.

In this section we demonstrate the applicability of the general results of §2 for the linearization $\Phi'(a)$ of the parameter-to-solution mapping $\Phi(a)$ given by the elliptic boundary value problem (1.4).

First we reconsider the saddle point problem (1.7) in the form

$$\sup_{f\in\mathcal{F}} G_f \tag{3.1}$$

where

$$G_f = \inf_{h\neq 0} \frac{|\Phi'_f(a)h|^2_Y + \beta\tilde\sigma(h, h)}{|h|^2_X}.$$

To apply the results of §2 we make the identification $T_f h = \Phi'_f(a)h$ and $\sigma(h_1, h_2) = \beta\tilde{\sigma}(h_1, h_2)$. Let us now assume that $\mathcal{F}$ is compact, that (H1*) holds, that $f \longrightarrow T_f$ from $\mathcal{F}$ to $\mathcal{L}(V, Y)$ is continuous, and that the smallest eigenvalues λ_f of $A_f, f \in \mathcal{F}$, are distinct. In view of the Rayleigh principle [DL]

$$G_f = \lambda_f = a_f(h_f, h_f),$$

where h_f denotes the eigenfunction associated with λ_f. Due to the continuity of $f \longrightarrow \lambda_f$ established in Proposition 2.2, (3.1) has at least one solution $f^* \in \mathcal{F}$. An approximate evalution of G_f can be achieved by Galerkin approximation and solution of the resulting generalized eigenvalue problems. For an approximate solution of (3.1) one can rely on iterative techniques that require a gradient. Sufficient conditions for the existence and characterization of the Gateaux and the Fréchet derivatives of $f \longrightarrow G_f$ were given in Theorem 2.4 and Corollary 2.5.

The following technical lemma will be required later. We denote by 1 the constant function with value 1 as well as the real number 1.

LEMMA 3.1. (i) *Let $\Omega \subset \mathbf{R}^n, n \in \mathbf{N}$, be a domain satisfying the cone property if $n > 1$; and let $l \in H^1(\Omega)^*$ with the property that $l(1) \neq 0$. Then $h \longrightarrow (|\nabla h|^2_{L^2} + |l(h)|^2)^{1/2}$ defines a norm on $H^1(\Omega)$ that is equivalent to the common Hilbert space norm on $H^1(\Omega)$.*

(ii) *Let A be a compact metric space, and let $l_\alpha \in H^1(\Omega)^*$ with the property that $l_\alpha(1) \neq 0$ for all $\alpha \in A$ and such that $\alpha \longrightarrow l_\alpha$ from A to $H^1(\Omega)^*$ is continuous. Then there exists $\kappa > 0$ such that*

$$|\nabla h|^2_{L^2} + |l_\alpha(h)|^2 \geq \kappa |h|^2_{H^1} \quad \textit{for all } \alpha \in A \textit{ and } h \in H^1(\Omega). \tag{3.2}$$

Proof. Part (i) follows from [M, p. 27]. We turn to (ii). If (3.2) was false, then there would exist $\{\alpha_i\}_{i=1}^\infty \subset A$ and $\{h\}_{i=1}^\infty \subset H^1(\Omega)$ with $|h_i|_{H^1} = 1$ such that

$$\lim_{i \to \infty} \left(|\nabla h_i|^2_{L^2} + |l_{\alpha_i}(h_i)|^2\right) = 0. \tag{3.3}$$

Due to compactness of A it can be assumed that $\{\alpha_i\}_{i=1}^\infty$ is convergent with limit α_0. Continuity of $\alpha \longrightarrow l_\alpha$ from A to $H^1(\Omega)^*$ implies that

$$\lim_{i \to \infty} [l_{\alpha_i}(h_i) - l_{\alpha_0}(h_i)] = 0. \tag{3.4}$$

From (3.3) and (3.4) it follows that

$$\lim(|\nabla h_i|^2_{L^2} + |l_{\alpha_0}(h_i)|^2) = 0,$$

which is a contradiction. □

3.1. The one-dimensional case.

Here we consider the two-point boundary value problem

$$\begin{aligned} -(au_x)_x &= f \text{ in } (0,1), \\ u(0) = u(1) &= 0, \end{aligned} \tag{3.5}$$

where $a \in H^1, a(x) \geq \nu > 0$ on $(0,1)$, and $f \in H^{-1}$. In this subsection all function spaces are considered over the interval $(0,1)$. We introduce the operator

$$A(a)u = -(au_x)_x;$$

it is an isomorphism form H_0^1 to H^{-1}. The solution $u = u(a, f)$ to (3.5) is given by $u(a, f) = A^{-1}(a)f$. The mapping $a \rightarrow u(a, f)$ (for $a \geq \nu$) is Fréchet differentiable from H^1 to H_0^1 and the Fréchet derivative at a in the direction $h \in H^1$ is given by

$$u_a(a, f)(h) = A^{-1}(a)D(hDu(a, f)). \tag{3.6}$$

Here D denotes differentiation. Henceforth a will be fixed. In the context of §2 we choose the spaces

$$X = L^2, \qquad V = H^1, \qquad Y = H_0^1, \qquad F = H^{-1}, \tag{3.7}$$

and define the operators

$$T_f h = A^{-1}(a)D(hD(u(a, f))). \tag{3.8}$$

Clearly $T_f \in \mathcal{L}(V, Y)$ for every $f \in \mathcal{F}$. For T_f to be well defined and for the subsequent discussion it suffices to have $a \in L^\infty$ and $a \geq \nu > 0$ a.e. on $(0, 1)$. Above we choose $a \in H^1$ only for the sake of consistency since the variations h of a are taken in $V = H^1$. We further define

$$\sigma(h, \tilde{h}) = \beta\langle h_x, \tilde{h}_x\rangle_{L^2}, \tag{3.9}$$

where β is a small positive parameter. To assertain continuity of the eigenvalues and eigenfunctions of $f \rightarrow A_f$ we verify (H1*) and (H2*). Let us assume that $\hat{f} \neq 0$. Then

$$h \rightarrow \left(|DT_{\hat{f}}h|^2_{L^2} + \beta|Dh|^2_{L^2}\right)^{1/2}$$

defines a norm on H^1 that is equivalent to the usual H^1-norm. This implies the existence of $\mu > 0$ such that

$$a_{\hat{f}}(h, h) = |DT_{\hat{f}}h|^2_{L^2} + \beta|Dh|^2_{L^2} \geq 2\mu|h|^2_{H^1} \quad \text{for all } h \in H^1. \tag{3.10}$$

Using (3.10) and continuity of $f \rightarrow u(f)$ from H^{-1} to H_0^1 it can be shown that there exists a neighborhood $\mathcal{F}$ of $\hat{f}$ in H^{-1}, such that

$$a_f(h, h) \geq \mu|h|^2_{H^1} \quad \text{for all } f \in \mathcal{F}\text{and } h \in H^1, \tag{3.11}$$

and (H1*) holds. Next we show the continuity of the linear mapping $f \rightarrow T_f$ from F to $\mathcal{L}(V, Y)$. Let $M = \| DA^{-1}(a)D \|_{\mathcal{L}(L^2)}$, and let k_1 be the embedding constant from H^1 into L^∞. For $f \in H^{-1} = F$ and $h \in H^1 = V$ we find

$$\begin{aligned} (3.12)\, |T_f h|_{H_0^1} &= |DA^{-1}(a)D(hDu(a, f))|_{L^2} \leq Mk_1|h|_{H^1}|u(a, f)|_{H_0^1} \\ &\leq Mk_1|h|_{H^1}|A^{-1}(a)f|_{H_0^1} \leq Mk_1|h|_{H^1} \parallel A^{-1}(a) \parallel_{\mathcal{L}(H^{-1}, H_0^1)} |f|_{H^{-1}}. \end{aligned}$$

This estimate implies continuity of $f \rightarrow T_f$. In particular (H2*) holds (with $f_0 = \hat{f}$), and Proposition 2.2 is applicable. Moreover, (H3i) holds with $\gamma = 1$. Turning to the differentiability properties of the eigenvalues and eigenfunctions of T_f, we note that $f \rightarrow T_f$ is in $\mathcal{L}(F, \mathcal{L}(V; Y))$. Thus (A2) holds at every $f \in F$ and $f \rightarrow T_f$ has a Fréchet derivative at every $f \in F$. Thus the conclusions of Theorem 2.4 and Corollaries 2.5 and 2.6 are applicable.

Specific examples. We discussed the case where $\mathcal{F}$ is a subset of H^{-1}. Next we turn to the situation that is of practical importance where f varies in a parameterized set.

(i) Let $\epsilon \in (0,1)$ and $\mathcal{F} = [\epsilon, 1-\epsilon] \subset \mathbf{R}$. For $\alpha \in \mathcal{F}$ let δ_α denote the delta impulse at α. For a as above and $\alpha \in \mathcal{F}$ let $u(a, \delta_\alpha) \in W^{1,\infty}$ denote the solution of (3.5) with $f = \delta_\alpha$. The general theory will be applied to the operators

$$h \rightarrow T_{\delta_\alpha}(h) = A^{-1}(a)D(hDu(a, \delta_\alpha)), \quad \alpha \in \mathcal{F},$$

which are elements of $\mathcal{L}(H^1, H_0^1)$, with X, V, Y and σ as in (3.7), (3.9) and $F = \mathbf{R}$. It is simple to argue that $h \rightarrow |T_{\delta_\alpha} h|$ satisfies the assumptions for l_α of Lemma 3.1 with $A = [\epsilon, 1-\epsilon]$. Consequently there exists $\mu > 0$ such that

$$a_{\delta_\alpha}(h,h) \geq \mu |h|^1_{H^1} \quad \text{for all } h \in H^1 \text{ and } \alpha \in \mathcal{F},$$

i.e., (3.11) and hence (H1*) holds. Concerning the continuity assumption (H2), we note that T_{δ_α} is the composition of the mappings $\alpha \rightarrow \delta_\alpha \rightarrow A^{-1}(a)D(hDu(a, \delta_\alpha))$, which is Hölder continuous from $\mathcal{F}$ to $\mathcal{L}(H^1, H_0^1)$ using (3.12). In particular, Proposition 2.2 is applicable. Since $\alpha \rightarrow \delta_\alpha$ is not differentiable from $\mathcal{F} \subset \mathbf{R}$ to H^{-1}, the mapping $\alpha \rightarrow (\lambda_{\delta_\alpha}, h_{\delta_\alpha})$ is not differentiable for $Y = H_0^1$ as output space. In [IK2] we derived an explicit formula for the lowest eigenvalue of $T^*_{\delta_\alpha} T_{\delta_\alpha}$ and we argued directional differentiability of the smallest eigenvalues by means of the explicit representation. If one chooses the output space Y to be L^2, then the continuity assumptions (H1*), (H2*), and (H3i) as well as the differentiability hypothesis (A2) hold. This is a consequence of the fact that $\alpha \rightarrow \delta_\alpha$ is differentiable from $\mathcal{F} \subset \mathbf{R}$ to H^{-2} and of Hölder continuity and continuous differentiability of $\alpha \rightarrow T_{\delta_\alpha}$ from $\mathcal{F}$ to $\mathcal{L}(H^1, L^2)$.

(ii) Again we choose $\epsilon \in (0,1)$ and put $\mathcal{F} = [\epsilon, 1-\epsilon]$. For $\alpha \in \mathcal{F}$ we define

$$\eta_\alpha(x) = \begin{cases} 1 & \text{for } x \in \left[a - \frac{\epsilon}{2}, a + \frac{\epsilon}{2}\right], \\ 0 & \text{otherwise.} \end{cases}$$

The inhomogeneities f in (3.5) are now chosen in the class of functions $\{\eta_\alpha : \alpha \in \mathcal{F}\}$. We define the operators $T_{\eta_\alpha} h = A^{-1}(a)D(hDu(a; \eta_\alpha))$. The spaces X, Y, and V and the form σ are chosen as above and $\beta > 0$. Using Lemma 3.1 one can show that there exists $\mu > 0$ such that

$$a_{\eta_\alpha}(h,h) = \langle T_{\eta_\alpha} h, T_{\eta_\alpha} h \rangle_{L^2} + \beta \langle Dh, Dh \rangle_{L^2} \geq \mu |h|^2_{H^1}$$

for all $h \in H^1$ and $\alpha \in \mathcal{F}$. Thus (H1*) holds and (H2), (H2*), and (H3i) are simple to check. Turning to (H3ii) we choose $\alpha, \tilde{\alpha} \in \mathcal{F}$ and find for all $h \in L^2$

$$\begin{aligned} (3.13) \quad |T_{\eta_\alpha} h - T_{\eta_{\tilde{\alpha}}} h|_{H^1} &\leq M |hDu(a, \eta_\alpha - \eta_{\tilde{\alpha}})|_{L^2} \\ &\leq M |h|_{L^2} |u(a, \eta_\alpha - \eta_{\tilde{\alpha}})|_{W^{1,\infty}} \leq k_2 M |h|_{L^2} |\eta_\alpha - \eta_{\tilde{\alpha}}|_{L^2}, \end{aligned}$$

where k_2 is a constant independent of $\alpha, \tilde{\alpha} \in \mathcal{F}$ and M was defined immediately following (3.11). A short calculation gives $|\eta_\alpha - \eta_{\tilde{\alpha}}|_{L^2} \leq \sqrt{2} |\alpha - \tilde{\alpha}|^{1/2}$. Combining this estimate with (3.13) we have

$$|T_{\eta_\alpha} h - T_{\eta_{\tilde{\alpha}}} h|_{H^1} \leq \sqrt{2} k_2 M |h|_{L^2} |\alpha - \tilde{\alpha}|^{1/2} \quad \text{for all } h \in H^1 \text{ and } \alpha, \tilde{\alpha}, \in \mathcal{F}.$$

This is (H3ii) with $\gamma = \frac{1}{2}$, and hence Proposition 2.3 is applicable. Concerning differentiability we first note that $\alpha \rightarrow \eta_\alpha$ from $\mathcal{F} \subset \mathbf{R} \rightarrow H^{-1}$ is differentiable with the derivative given by $\frac{d}{d\alpha} \eta_\alpha = \delta_{\alpha + \epsilon/2} - \delta_{\alpha - \epsilon/2}$. Moreover, $\alpha \rightarrow T_{\eta_\alpha}$ is differentiable with

$$\frac{d}{d\alpha} T_{\eta_\alpha} = T_{\delta_{\alpha+\epsilon/2} - \delta_{\alpha-\epsilon/2}},$$

and Lemma 2.7 implies

$$\frac{d}{d\alpha}\mathcal{T}_{\eta_\alpha} = \frac{d}{d\alpha}(T^*_{\eta_\alpha} T_{\eta_\alpha}) = T^*_{\eta_\alpha} T_{\delta_{\alpha+\epsilon/2}-\delta_{\alpha-\epsilon/2}} + T^*_{\delta_{\alpha+\epsilon/2}-\delta_{\alpha-\epsilon/2}} T_{\eta_\alpha}.$$

We conclude Fréchet differentiability of the eigenvalues and eigenvectors of A_{η_α} according to Corollaries 2.5 and 2.6.

3.2. The multidimensional case. We turn to

$$(3.14) \qquad \begin{cases} -\mathrm{div}(a \text{ grad } u) = f & \text{in } \Omega, \\ u|\partial\Omega = 0, \end{cases}$$

where Ω is a bounded domain in $\mathbf{R}^n, n \in \{2,3,4\}$, with boundary $\partial\Omega$ of class $C^{1,\delta}$ for some $\delta \in (0,1)$. The choice of n guarantees that $H^1 \subset L^4$. Unless specified otherwise the function spaces in this subsection are considered over the domain Ω. The coefficient a is assumed to satisfy $a \in L^\infty, a(x) \geq \nu > 0$ a.e. on Ω, and $f \in H^{-1}$. Additional regularity for a and f will be required later. The operator $A(a) : H^1_0 \to H^{-1}$ defined by

$$A(a)u = -\mathrm{div}(a \text{ grad } u)$$

is an isomorphism. In particular this implies that (3.14) has a unique solution $u(a,f) = A^{-1}(a)f$. The mapping $a \to u(a,f)$ is Fréchet differentiable from L^∞ to H^{-1} with

$$(3.15) \qquad u_a(a,f)h = A^{-1}(a)\nabla \cdot (h\nabla u(a,f)) \quad \text{for } h \in L^\infty.$$

To extend (3.15) to a continuous linear operator on H^1 additional regularity of $u(a,f)$ is required. For the dimensions under consideration it suffices to require $u(a,f) \in W^{1,4}$. This is implied by the additional assumptions

$$(3.16) \qquad f \in W^{-1,4} = (W_0^{1,4/3})^* \quad \text{and} \quad a \in C^{0,\delta}$$

[A, T]. Henceforth we fix $a \in C^{0,\delta}, a \geq \nu > 0$, and for $f \in W^{-1,4}$ we introduce $T_f : H^1 \to H^1_0$ defined by

$$T_f(h) = A^{-1}(a)\nabla \cdot (h\nabla u(a,f)).$$

Conditions were given in [IK1] that guarantee injectivity of T_f (and $|T_f h|_{H^1_0} \geq \kappa |h|_{L^2}$, where $\kappa > 0$ is uniform with respect to h in bounded subsets of H^1). We shall discuss the "wellposedness" criterion (H1*) and the continuity assumption (H2*) with the choice

$$(3.17) \qquad X = L^2,\ Y = H^1_0,\ V = H^1,\ F = W^{-1,4},\ \sigma(h,\tilde{h}) = \beta\langle\nabla h, \nabla\tilde{h}\rangle_{L^2},$$

and

$$a_f(h,\tilde{h}) = \langle T_f h, T_f \tilde{h}\rangle_{H^1_0} + \beta\langle\nabla h, \nabla\tilde{h}\rangle_{L^2}, \qquad \text{for } h, \tilde{h} \in H^1 \text{ and } f \in F.$$

Let $0 \neq \hat{f} \in F$. Then by Lemma 3.1 there exists $\mu > 0$ such that

$$a_{\hat{f}}(h,h) \geq 2\mu|h|^2_{H^1} \quad \text{for all } h \in H^1.$$

We shall make use of the fact that there exists a constant C such that

$$|u(f)|_{W_0^{1,4}} \leq C|f|_{W^{-1,4}} \text{ for all } f \in W^{-1,4} \tag{3.18}$$

[A, p. 48], [T, p. 179]. Due to (3.18) and the continuous embedding of H^1 into L^2 there exists a neighborhood $\mathcal{F}$ of $\hat{f}$ in $W^{-1,4}$ such that

$$a_f(h,h) \geq \mu |h|^2_{H^1} \quad \text{and} \quad f \in \mathcal{F}. \tag{3.19}$$

This is (H1*). For the following estimate let $M = \| \nabla A^{-1}(a)\nabla \|_{\mathcal{L}(L^2_n)}$, with $L^2_n = \otimes_{i=1}^n L^2$, and let k_2 denote the embedding constant of H^1 into L^4. For $h \in H^1$ we find

$$\begin{aligned} |T_f h|_{H_0^1} &= |\nabla A^{-1}(a)\nabla \cdot (h\nabla u(a,f))|_{L^2} \leq M|h\nabla u(a,f)|_{L^2} \\ &\leq M|h|_{L^4}|\nabla u(a,f)|_{L^4_n} \leq Mk_2|h|_{H^1}|u(a,f)|_{W^{1,4}} \\ &\leq Mk_2C|h|_{H^1}|f|_{W^{-1,4}}. \end{aligned} \tag{3.20}$$

This estimate implies that $f \to T_f \in \mathcal{L}(W^{-1,4}, \mathcal{L}(H^1, H_0^1))$ is Lipschitz continuous. It follows that (H2*), (H3i), and (A2) hold and that Propositions 2.2 and 2.3(i), Theorem 2.4, and Corollaries 2.5 and 2.6 are applicable for the choice of spaces given in (3.7).

Specific example. As in the one-dimensional case we consider a specific situation where f varies in a parametrized set. Let $n = 2$; let Ω be as specified above; and choose $\mathcal{F} = [c,d]$, a compact interval in $\mathbf{R}$. For every $\alpha \in \mathcal{F}$, Γ_α denotes a (nontrivial) Lipschitzian curve in Ω and f_α is given by

$$\langle f_\alpha, \varphi \rangle = \int_{\Gamma_\alpha} \varphi \, ds \quad \text{for all } \varphi \in H_0^1. \tag{3.21}$$

We shall make use of the fact that $(W_0^{1,4/3})^* = W^{-1,4}$, with $W^{-1,4}$ densely injected into H^{-1}. Moreover, if G is a bounded open subset of $\mathbf{R}^n$ with Lipschitz boundary ∂G, then the zero-order trace operator has a unique continuous extension to an operator form $W^{1,4/3}$ onto $W^{1/4,4/3}(\partial\Omega)$ and $W^{1/4,4/3}(\partial\Omega) \subset L^2(\partial\Omega)$. This implies that (3.21) defines an element $f_\alpha \in W^{-1,4}(\Omega)$, and

$$h \to T_{f_\alpha}(h) = \nabla A^{-1}(a)\nabla(h\nabla u(a,f_\alpha))$$

is well defined for all $h \in H^1$. We next give a condition that implies continuity of $\alpha \to f_\alpha$ from $\mathcal{F}$ to $W^{-1,4}$.

> For $\alpha_0 \in (c,d)$ choose $\epsilon = \epsilon(\alpha_0) > 0$ such that $I_{\alpha_0} = [\alpha_0, \alpha_0 + \epsilon] \subset [c,d]$ and define the sets
>
> $$\Omega_\epsilon = \operatorname{int} \bigcup_{\beta \in I_{\alpha_0}} \Gamma_\beta \text{ and } \Omega_\alpha = \operatorname{int} \bigcup_{\beta \in [\alpha_0, \alpha]} \Gamma_\beta, \quad \text{for } \alpha \in I_{\alpha_0}. \tag{3.22}$$
>
> It is assumed that Ω_ϵ and Ω_α are connected domains in $\mathbf{R}^2$ with $\partial\Omega_\alpha$ Lipschitzian for all $\alpha \in I_{\alpha_0}$ and such that the subdomains Ω_α and $\Omega_\epsilon - \Omega_\alpha$ lie on opposite sides of Γ_α and
>
> $$\partial\Omega_\alpha = \Gamma_{\alpha_0} \cup \Gamma_\alpha \cup \tilde{\Gamma}_\alpha,$$

with $\tilde{\Gamma}_\alpha$ consisting of at most two connected smooth components. Let v_α denote the unit normal to Γ_α pointing into $\Omega_\epsilon - \Omega_\alpha$. Assume that v_α defines a vectorfield $v \in W^{1,\infty}(\Omega_\epsilon, \mathbf{R}^n)$ as α varies in I_{α_0} and that it can be extended uniquely to a vectorfield on $\overline{\Omega_\epsilon}$ and that $\alpha \to \alpha_0$ implies $\int_{\Omega_\alpha} dx \to 0$ and $\int_{\tilde{\Gamma}_\alpha} ds \to 0$. Further assume that the analogous construction is possible with I_{α_0} replaced by $[\alpha_0 - \epsilon, \alpha_0]$.

Let $\alpha \to \alpha^0$, and without loss of generality assume that $\alpha \geq \alpha_0$. Then by Green's formula, applied to $v\psi \in W^{1,4/3}(\Omega_\alpha)$, we find for $\psi \in W^{1,4/3}(\Omega)$ and n_α the unit outer normal to Ω_α:

$$\begin{aligned}\int_{\Omega_\alpha} \operatorname{div}(v\psi)\,dx &= \int_{\Gamma_{\alpha_0}} n_\alpha v\psi\,ds + \int_{\Gamma_\alpha} n_\alpha v\psi\,ds + \int_{\tilde{\Gamma}_\alpha} n_\alpha v\psi\,ds \\ &= -\int_{\Gamma_{\alpha_0}} \psi\,ds + \int_{\Gamma_\alpha} \psi\,ds + \int_{\tilde{\Gamma}_\alpha} n_\alpha v\psi\,ds,\end{aligned}$$

and consequently

$$\begin{aligned}|\langle f_{\alpha_0} - f_\alpha, \psi\rangle| &= \left|\int_{\Gamma_{\alpha_0}} \psi\,ds - \int_{\Gamma_\alpha} \psi\,ds\right| \\ &\leq \left|\int_{\Omega_\alpha} \operatorname{div}(v\psi)\,dx\right| + \left|\int_{\tilde{\Gamma}_\alpha} n_\alpha v\psi\,ds\right| \\ &\leq K\left(|\psi|_{W^{1,4/3}}\left(\int_{\Omega_\alpha} dx\right)^{1/4} + \int_{\tilde{\Gamma}_\alpha} |\psi|\,ds\right),\end{aligned}$$

with K a constant independent of $\psi \in W^{1,4/3}$. It follows that

$$\begin{aligned}\left|\int_{\Gamma_{\alpha_0}} \psi\,ds - \int_{\Gamma_\alpha} \psi\,ds\right| &\leq K\left[\left|\int_{\Omega_\alpha} dx\right|^{1/4} |\psi|_{W^{1,4/3}} + \left(\int_{\tilde{\Gamma}_\alpha} ds\right)^{1/2}\left(\int_{\tilde{\Gamma}_\alpha} |\psi|^2\,ds\right)^{1/2}\right] \\ &\leq \tilde{K}\,|\psi|_{W^{1,4/3}(\Omega)}\left[\left(\int_{\Omega_\alpha} dx\right)^{1/4} + \left(\int_{\tilde{\Gamma}_\alpha} ds\right)^{1/2}\right],\end{aligned}$$

with $\tilde{K}$ independent of $\psi \in W^{1,4/3}$. By (3.22) continuity of $\alpha \to f_\alpha$ from $\mathcal{F}$ to $W^{-1,4}$ follows. Together with (3.20) we have

$$\alpha \to T_{f_\alpha} \text{ is continuous from } [c,d] \text{ to } \mathcal{L}(H^1, H_0^1). \tag{3.23}$$

Thus (H2*) holds. To verify (H1*) we note that $l_\alpha(1) \neq 0$ for all $\alpha \in [c,d]$. Together with compactness of $\mathcal{F}$ and (3.23), Lemma 3.1 implies the desired conclusion. Thus Proposition 2.2 can be applied. It implies continuity of the eigenvalues and eigenfuctions of A_{f_α} provided they are distinct. Moreover, since $\mathcal{F}$ is compact, the associated saddle point problem (3.1) has a solution.

Two examples in which (3.22) holds true are given by the domain $\Omega =$ unit ball in $\mathbf{R}^2$, and

(i)

$$\Gamma_\alpha = \left\{\begin{pmatrix} x \\ \alpha \end{pmatrix} : x \in \left[-\frac{1}{2}, \frac{1}{2}\right]\right\}, \qquad \alpha \in \left[-\frac{1}{2}, \frac{1}{2}\right];$$

and

(ii)

$$\Gamma_\alpha = \left\{ \begin{pmatrix} x \\ \alpha(x-\frac{1}{2})(x+\frac{1}{2}) \end{pmatrix} : x \in \left[-\frac{1}{2}, \frac{1}{2} \right] \right\}, \qquad \alpha \in [-2, 2].$$

3.3. The multidimensional case revisited. Once again we consider the problem of the optimal choice of f for the identification of a in (3.14), but now the output space is chosen to be $H^2 \cap H^1_0$ rather than H^1_0. This choice of output space enhances the sensitivity of the parameter-to-solution mapping. We shall consider the maximization of the measure

$$G^\beta_f = \inf_{\substack{h \in H^1 \\ h \neq 0}} \frac{|T_f h|^2_{H^2} + \beta |\nabla h|^2}{|h|^2_{L^2}} \tag{3.24}$$

with $\beta > 0$ (regularized case) and $\beta = 0$ (unregularized case).

Let Ω be a bounded domain in $\mathbf{R}^n$ with Lipschitzian boundary, and fix a such that $A(a) : H^2 \cap H^1_0 \to L^2$ is an isomorphism. Assume further that $\partial\Omega$ is sufficiently regular such that for an appropriately chosen normed linear space F the linear mapping

$$f \to u(a, f)$$

is continuous from F to $W^{1,\infty} \cap W^{2,1+\epsilon}$ with $\epsilon > 0$ if $n = 2$ and from F to $W^{2,n}$ if $n \geq 3$.

Here $u(a, f)$ denotes the solution of (3.14).

It follows that $h \to T_f(h) = A^{-1}(a)\nabla\cdot(h\nabla u(a, f))$ is an element of $\mathcal{L}(H^1, H^2 \cap H^1_0)$ and that there exists a constant C independent of $(h, f) \in H^1 \times F$ such that

$$|T_f(h)|_{H^2 \cap H^1_0} \leq C |h|_{H^1} |f|_F \quad \text{for all } (h, f) \in H^1 \times F. \tag{3.25}$$

In the context of the notation of §2 we take

$$X = L^2, \qquad Y = H^2 \cap H^1_0, \qquad Y = H^1, \qquad \sigma(h, \tilde{h}) = \beta \langle \nabla h, \nabla \tilde{h} \rangle_{L^2}, \tag{3.26}$$

F as specified above and

$$a_f(h, \tilde{h}) = \langle T_f h, T_f \tilde{h} \rangle_{H^2} + \beta \langle \nabla h, \nabla \tilde{h} \rangle_{L^2} \quad \text{for } h, \tilde{h} \in H^1 \text{ and } f \in F. \tag{3.27}$$

First let us argue that (H1*), (H2*), and (A2) hold for an appropriately chosen subset $\mathcal{F} \subset F$, if $\beta > 0$.

Let $0 \neq \hat{f} \in F$. Then by Lemma 3.1 there exists $\mu > 0$ such that

$$a_{\hat{f}}(h, h) \geq 2\mu |h|^2_{H^1} \quad \text{for all } \in H^1,$$

and due to (3.25) and bilinearity of $(h, f) \to T_f h$ there exists a neighborhood $\mathcal{F}$ of $\hat{f}$ such that

$$a_f(h, h) \geq \mu |h|^2_{H^1} \quad \text{for all } (h, f) \in H^1 \times \mathcal{F}.$$

Thus (H1*) is satisfied. As a consequence of (3.25) condition (H2*) holds with f_0 replaced by $\hat{f}$ and hence Propostion 2.2 is applicable. We note that (3.25) also implies (H3i) and (A2) so that, in particular, by Corollary 2.5 Fréchet differentiability of $f \to (\lambda_f, x_f)$ at $\hat{f}$ follows.

Now let us turn to the case $\beta = 0$ in (3.24). This requires the interpretation of T_f as an unbounded operator in L^2. In [IK] sufficient conditions are given that guarantee T_f with dom $T_f = H_0^1 \subset L^2$ and $T_f h = A^{-1}\nabla \cdot (h \nabla u(a,f))$ is a closable operator in L^2 with its closure (which is again denoted by T_f) satisfying

$$|T_f h|_{H^2} \geq 2K |h|_{L^2} \quad \text{for all } h \in \text{ dom } T_f,$$

for a constant $K > 0$ independent of h. Such a sufficient condition is given by assuming that $\Omega \subset \mathbf{R}^n, n \in \{1,2,3\}, \partial\Omega$ is $C^{1,1}$-smooth (if $n = 2,3$), and

$$(3.28) \qquad \begin{cases} u(a,f) \in W^{3,4}(\Omega) \\ \nabla u(x;a,f) \cdot n < 0 \text{ on } \partial\Omega, \\ \text{there exist constants } k_1 \in \mathbf{R} \text{ and } k_2 > 0 \text{ such that} \\ k_1 |\nabla u(x;a,f)|^2 + \Delta u(x;a,f) \geq k_2 \text{ on } \Omega. \end{cases}$$

Let us assume henceforth that the normed linear space F is chosen such that

$$f \longrightarrow u(a,f)$$

is continuous from F to $W^{3,4}$, and let $\hat{f}$ be a reference input for which (3.28) holds. Then the existence of a neighborhood $\mathcal{F}$ of $\hat{f}$ follows, such that T_f is closable for all $f \in \mathcal{F}$ and

$$(3.29) \qquad |T_f h|_{H^2} \geq K|h|_{L^2} \quad \text{for all } (h,f) \in \text{ dom } T_f \times \mathcal{F}.$$

In particular this implies that the measure

$$G_f = \inf_{\substack{h \in \text{ dom } T_f \\ h \neq 0}} \frac{|T_f h|^2_{H^2}}{|h|^2_{L^2}}$$

is nontrivial for all $f \in \mathcal{F}$. We shall argue that $f \longrightarrow G_f$ is upper semicontinuous, so that existence of a solution to

$$\max \ G_f \quad \text{over } f \in \mathcal{F}$$

is guaranteed if $\mathcal{F} \subset F$ is compact. We first show that

$$(3.30) \qquad \lim_{\beta \to 0^+} G_f^\beta = G_f \quad \text{for every } f \in \mathcal{F}.$$

Since $\beta \longrightarrow G_f^\beta$ is monotonically decreasing, $\lim_{\beta \to 0^+} G_f^\beta$ exists. Arguing by contradiction, assume that $\lim_{\beta \to 0^+} G_f^\beta = \mu_0 > G_f$. Since $H^1(\Omega)$ is dense in $L^2(\Omega)$, we find

$$(3.31) \qquad G_f = \inf_{\substack{h \neq 0 \\ h \in H^1}} \frac{|T_f h|^2_{H^2}}{|h|^2_{L^2}}.$$

Hence there exists $\bar{h} \neq 0$ in $H^1(\Omega)$ such that

$$\frac{|T_f \bar{h}|^2_{H^2}}{|\bar{h}|^2_{L^2}} \leq G_f + \delta,$$

where $\delta = \frac{1}{4}(\mu_0 - G_f)$. It follows that

$$G_f^\beta \le \frac{|T_f\bar{h}|^2_{H^1} + \beta|\nabla\bar{h}|^2_{L^2}}{|\bar{h}|^2_{L^2}} \le G_f + 2\delta = \frac{1}{2}(\mu_0 + G_f) \tag{3.32}$$

for all β with $\beta|\nabla\bar{h}|^2_{L^2} \le \delta|\bar{h}|^2_{L^2}$. Taking the limit with respect to β in (3.32) leads to a contradiction, and hence (3.30) holds. To prove upper semicontinuity of $f \to G_f$ let $\{f_n\}_{n=1}^\infty$ be a sequence in $\mathcal{F}$ with $\lim_{n\to\infty} f_n = f$, and let $\eta > 0$ be arbitrary. By (3.30) there exists β_0 such that

$$|G_f^{\beta_0} - G_f| < \eta.$$

Moreover, by Proposition 2.3 there exists N_0 such that

$$|G_{fn}^{\beta_0} - G_f^{\beta_0}| < \eta \quad \text{for all } n \ge N_0.$$

Consequently we have

$$\begin{aligned} G_f - G_{f_n} &\ge G_f - G_f^{\beta_0} + G_f^{\beta_0} - G_{f_n}^{\beta_0} + G_{f_n}^{\beta_0} - G_{f_n} \\ &\ge G_f - G_f^{\beta_0} + G_f^{\beta_0} - G_{f_n}^{\beta_0} \ge -2\eta \end{aligned}$$

for all $n \ge N_0$. Since $\eta > 0$ is arbitrary, upper semicontinuity of $f \to G_f$ follows.

4. Numerical experiments.

We carried out numerical tests to determine the optimal location of the unit impulse δ_α in

$$\begin{cases} -(au_x)_x = & \delta_\alpha \quad \text{in } (0,1), \\ u(0) = u(1) = & 0, \end{cases} \tag{4.1}$$

for the determination of the function $a(x)$. Theoretical aspects for this problem were considered in §3.1, Specific Examples (i). For $\alpha \in (0,1)$ the solution to (4.1) satisfies $u(a, \delta_\alpha) \in W^{1,\infty}$, and therefore $T_{\delta_\alpha} h = A^{-1}(a)D(hDu(a,\delta_\alpha))$ is well defined for $h \in L^2$. The Hermitian forms are given by

$$a_{\delta_\alpha}(h, \tilde{h}) = \langle T_{\delta_\alpha} h, T_{\delta_\alpha} \tilde{h}\rangle_{H_0^1} + \beta\langle h_x, \tilde{h}_x\rangle_{L^2},$$

and the saddle point problems are

$$\sup \inf_{\substack{h\in H^1 \\ h\neq 0}} \frac{a_{\delta_\alpha}(h,h)}{|h|^2_{L^2}}. \tag{4.2}$$

It is known [IK1] that a_{δ_α} does not satisfy (H1) for $\beta = 0$. To discretize these problems we choose subspaces $H^N = \{h^N = \sum_{i=1}^{N/2} h_i B_i^N : h_i\}$, where N is even and B_i^N are piecewise constant functions with value 1 on $\left[\frac{2(i-1)}{N}, \frac{2i}{N}\right]$ and value 0 outside of this interval. For given α and $a > 0$, the approximate solution $u^N(a, \delta_\alpha)$ to (4.1) was determined as the Galerkin solution with respect to the discretization of (4.1) by linear spline functions on the grid $\left\{\frac{i}{N}\right\}_{i=0}^N$. The approximate minimization problems are then given by

$$\min_{\substack{h^N\in H^N \\ h^N\neq 0}} \frac{a_{\delta_\alpha}^N(h^N, h^N)}{|h^N|^2_{L^2}}, \tag{4.3}$$

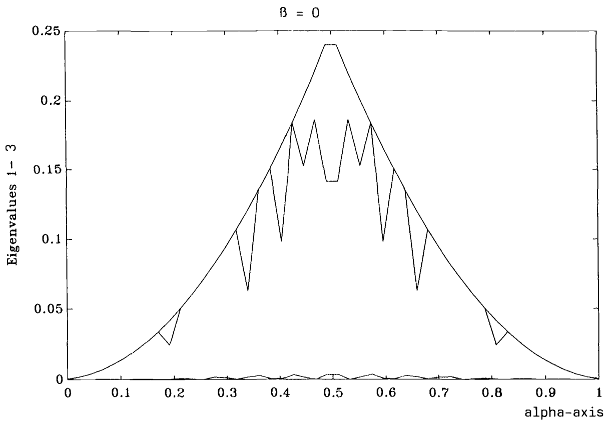


FIG. 1.

where $a^N_{\delta_\alpha}$ is defined like a_{δ_α} with $u(a, \delta_\alpha)$ replaced by $u^N(a, \delta_\alpha)$. The solution to (4.3) is characterized by the smallest eigenvalue of the generalized eigenvalue problem

$$A^N_{\delta_\alpha} \vec{h}^N = Q^N \vec{h}^N, \tag{4.4}$$

and the minimum in (4.3) is assumed at the eigenfunction associated with the smallest eigenvalue. In (4.4), $\vec{h}^N$ denotes the coordinate vector of h^N, and $A^N_{\delta_\alpha}, Q^N$ are the matrix representations of the forms $a^N_{\delta_\alpha}(h^N, \tilde{h}^N)$ and $\langle h^N, \tilde{h}^N \rangle$ on $H^N \times H^N$.

(i) In the first test example we choose $a = 1$. It is simple to see that 0 is an eigenvalue of $T^*_{\delta_\alpha} T_{\delta_\alpha}, \alpha \in (0,1)$, with T_{δ_α} considered as the operator form L^2 to H^1_0. Consequently we expect that for $\beta = 0$ the lowest eigenvalues of $A^N_{\delta_\alpha}, \alpha \in (0,1)$, approximate zero from above. Figure 1 shows the first three eigenvalues of A_{δ_α} as α varies in $(0,1)$ with $\beta = 0$ and $N = 128$. The curves were obtained by solving the generalized eigenvalue problem (4.4) for $\alpha_i = \frac{i}{49}, i = 1, \ldots, 48$. This plot was already analyzed in [IK2] where we noted that the eigenfunctions associated with the lowest eigenvalue approximate the step function in the spectrum of A_{δ_α} (considered as operator in L^2) and the eigenfunctions associated with the second eigenvalue approximate a δ-function. In [IK2] we also showed that the smallest eigenvalues of $T^*_{\delta_\alpha} T_{\delta_\alpha}$ when properly restriced to a subspace of codimension 1 (to eleminate 0 as eigenvalue) are given by

$$\lambda^0(\alpha) = \begin{cases} \alpha^2 & \text{on } \left(0, \frac{1}{2}\right), \\ (1-\alpha)^2 & \text{on } \left(\frac{1}{2}, 1\right). \end{cases}$$

The graph of λ^0 is almost identical with the third eigenvalue of $A_{\delta_\alpha}, \alpha \in (0,1)$, seen in Fig. 1. The elimination of a proper subspace to obtain (H1*) may be impractical in larger problems, and then regularization can be used. In Figs. 2 and 3 we show the results with $\beta = 10^{-4}$ and 10^{-3} respectively. The remaining specifications are those of Fig. 1. It can be observed that the lowest eigenvalue is now bounded away from 0 and that its maximum is attained at (the desired) location $\alpha^* = .5$. A comparison of all eigenvalues of A_{δ_α} with $\beta = 0$ and $\beta > 0$ also shows that regularization spreads the eigenvalues apart.

(ii) For this experiment we choose $a(x) = 1 + x$. It is known from [IK2] that the optimal location for the unit impulse is to the right of .5. Figure 4 depicts the

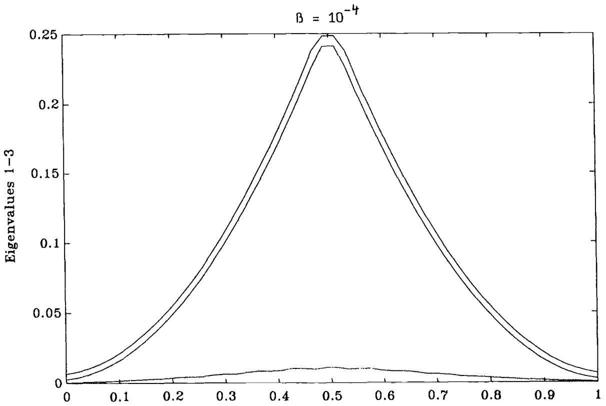


FIG. 2.

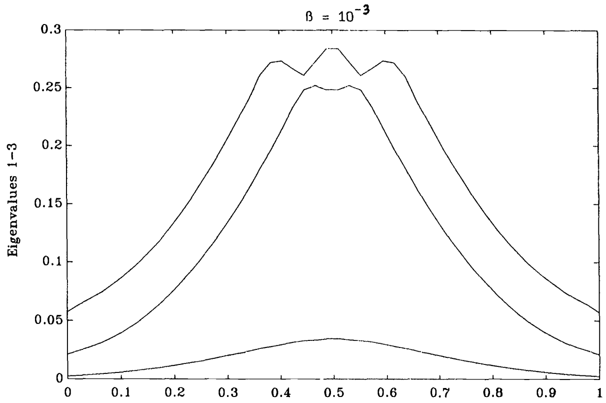


FIG. 3.

lowest eigenvalue of the generalized eigenvalue problem (4.4) for different values of β and for $N = 32$. The abscissa is again the α-axis, and the curves were obtained from results for $\alpha_i = \frac{i}{49}, i = 1, \ldots, 48$. Starting with the lowest curve, the graphs correspond successively to $\beta = 10^{-5}, 5 \times 10^{-5}, 10^{-4}, 5 \times 10^{-4}, 10^{-3}$. If β is sufficiently large, here $\beta \geq 5 \times 10^{-4}$, then we can see that $\alpha^* \approx .531$, which is larger than .5, as desired. The value for the optimal input $\hat{\alpha}$ calculated by restricting $T^*_{\delta_\alpha} T_{\delta_\alpha}$ to a proper subspace, so that (H1*) holds, was found to be $\hat{\alpha} \approx .545$ [IK2, Plot 2]. The zig-zag of the curves corresponding to small β-values is due to the relative distance of the gridpoints $\left\{\frac{i}{N}\right\}_{i=0}^{N}$ of the discretization for u and h and the location for the impulses at $\left\{\frac{j}{49}\right\}_{j=1}^{48}$, with smaller distance giving smaller eigenvalues.

(iii) In the final two plots we compare the eigenfunctions corresponding to the three smallest eigenvalues of the generalized eigenvalues problem (4.4) for $\beta = 0$ and $\beta = 10^{-4}$; see Figs. 5 and 6 respectively. Here $a(x) = 1+x, N = 128$, and $\alpha = \alpha^*$, the optimal input location found in (ii). As expected, the eigenfunctions for $\beta = 10^{-4}$ are much smoother than those for $\beta = 0$. We can also note that the first eigenfunction for $\beta = 10^{-4}$ is a smooth approximation to a multiple of the first eigenfunction for $\beta = 0$, which is almost a step function.

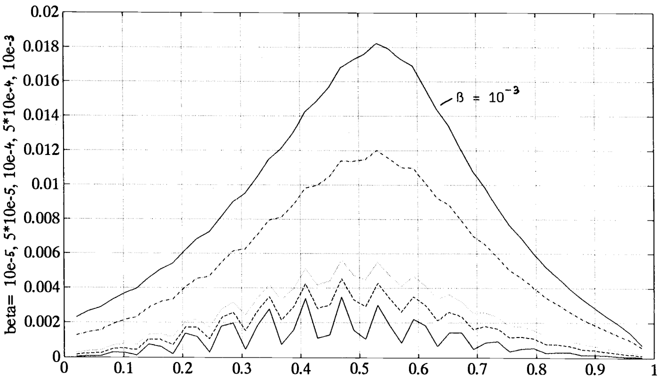


FIG. 4.

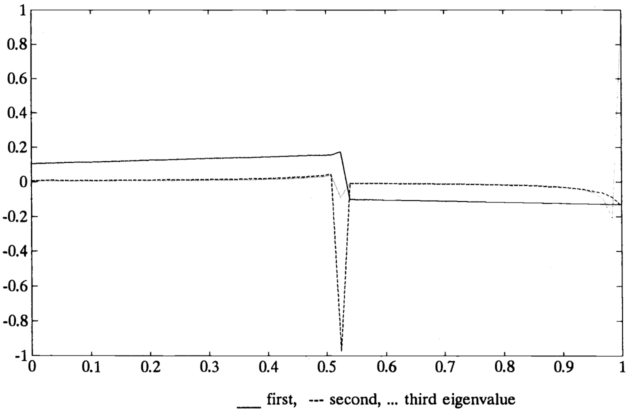


FIG. 5.

5. Appendix. Here we give the proof for the differentiability results of §2. The eigenvectors x_f^n are normalized by $\langle x_f^n, x_{f_0}^n \rangle = 1$ and put $(\lambda_0^n, x_0^n) = (\lambda_{f_0}^n, x_{f_0}^n)$. The superscript $n = 1$ is deleted.

Proof of Theorem 2.4. For simplicity of notation we put $f = f_0 + \tau f_1$ and $(\lambda_0, x_0) = (\lambda_{f_0}, x_{f_0})$. As a consequence of Proposition 2.2 and (H2*) we find

$$(\lambda_f, x_f) \to (\lambda_0, x_0) \quad \text{in } \mathbf{R} \times V, \text{ for } \tau \to 0. \tag{5.1}$$

By Lemma 2.1

$$a_f(x_f, v) = \lambda_f \langle x_f, v \rangle \tag{5.2}$$

and

$$a_{f_0}(x_0, v) = \lambda_0 \langle x_0, v \rangle \quad \text{for all } v \in V. \tag{5.3}$$

Subtracting (5.3) from (5.2) we obtain

$$\begin{aligned} \sigma(x_f - x_0, v) + \langle T_{f_0}(x_f - x_0), T_{f_0} v \rangle_Y - \lambda_0 \langle x_f - x_0, v \rangle_X \\ -(\lambda_f - \lambda_0) \langle x_f, v \rangle_X \\ = -\langle T_f x_f, T_f v \rangle_Y + \langle T_{f_0} x_f, T_{f_0} v \rangle_Y. \end{aligned} \tag{5.4}$$

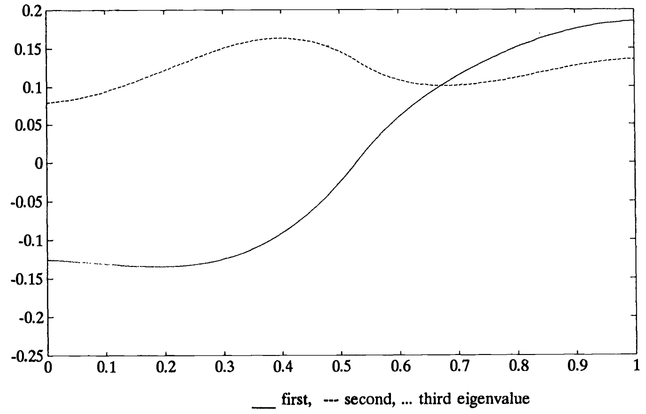


FIG. 6.

The normalizing equation $\langle x_f, x_{f_0}\rangle = 1$ together with (5.4) gives the following system of equations

$$(5.5)\qquad \begin{cases} (A_{f_0} - \lambda_0)(x_f - x_0) - (\lambda_f - \lambda_0)x_f & = -(\mathcal{T}_f x_f - \mathcal{T}_{f_0} x_f), \\ \qquad\qquad < x_f, x_f - x_0 >_X & = 0. \end{cases}$$

Introducting the operaters

$$C(x_f) : V \times \mathbf{R} \to V^* \times \mathbf{R}$$

defined by

$$C(x_f) = \begin{pmatrix} (A_{f_0} - \lambda_0) & -x_f \\ \langle x_0, \cdot\rangle & 0 \end{pmatrix}$$

it can be seen that (5.5) is equivalent to

$$(5.6)\quad C(x_0) = \begin{pmatrix} x_f - x_0 \\ \lambda_f - \lambda_0 \end{pmatrix} = [C(x_0) - C(x_f)] \begin{pmatrix} x_f - x_0 \\ \lambda_f - \lambda_0 \end{pmatrix} - \begin{pmatrix} \mathcal{T}_f x_f - \mathcal{T}_{f_0} x_f \\ 0 \end{pmatrix}.$$

Due to Proposition 2.2 and the special form of $C(x_f)$ we have

$$(5.7)\qquad \| \, C(x_0) - C(x_f) \, \|_{\mathcal{L}(V\times\mathbf{R}, V^*\times\mathbf{R})} \to 0 \quad \text{for } \tau \to 0.$$

It is simple to verify that $C(x_0)$ is injective and surjective. It follows that $C(x_0)$ has a bounded inverse and that (5.6) is equivalent to

$$(5.8)\qquad \left[I + C(x_0)^{-1} M_\tau\right] \begin{pmatrix} x_f - x_0 \\ \lambda_f - \lambda_0 \end{pmatrix} = -C(x_0)^{-1} \begin{pmatrix} \mathcal{T}_f x_f - \mathcal{T}_{f_0} x_f \\ 0 \end{pmatrix},$$

where $M_\tau = C(x_f) - C(x_0) \in \mathcal{L}(V \times \mathbf{R}, V^* \times \mathbf{R})$ and $\| \, M_\tau \, \| \to 0$ with $\tau \to 0$.

From (5.7) there exists $\tau_0 > 0$ and a constant K such that $I + C(x_0)^{-1} M_\tau$ is invertible in $\mathcal{L}(V \times \mathbf{R})$ and

$$(5.9)\qquad \| \, [I + C(x_0)^{-1} M_\tau]^{-1} \, \|_{\mathcal{L}(V\times\mathbf{R})} \le K$$

for all τ with $|\tau| \leq \tau_0$. Therefore, (5.8) is equivalent to

$$(5.10)\qquad \begin{pmatrix} x_f - x_0 \\ \lambda_f - \lambda_0 \end{pmatrix} = -[I + C(x_0)^{-1}M_\tau]^{-1}C(x_0)^{-1}\begin{pmatrix} \mathcal{T}_f x_f - \mathcal{T}_{f_0} x_f \\ 0 \end{pmatrix}.$$

Finally observe that

$$\begin{aligned}(5.11)\quad & \| \ \frac{1}{\tau}[I + C(x_0)^{-1}M_\tau]^{-1}C(x_0)^{-1}\begin{pmatrix} \mathcal{T}_f x_f - \mathcal{T}_{f_0} x_f \\ 0 \end{pmatrix} \\ & \quad -C(x_0)^{-1}\begin{pmatrix} \dot{\mathcal{T}}(f_0, f_1)x_0 \\ 0 \end{pmatrix} \|_{\mathcal{L}(V\times\mathbf{R})} \\ \leq \ & \| \ C(x_0)^{-1}M_\tau[I + C(x_0)^{-1}M_\tau]^{-1}C(x_0)^{-1}\begin{pmatrix} \frac{1}{\tau}(\mathcal{T}_f x_f - \mathcal{T}_{f_0} x_f) \\ 0 \end{pmatrix} \|_{\mathcal{L}(V\times\mathbf{R})} \\ + \ & \| \ C(x_0)^{-1} \|_{\mathcal{L}(V^*\times\mathbf{R}, V\times\mathbf{R})} \, |\frac{1}{\tau}[\mathcal{T}_f x_f - \mathcal{T}_{f_0} x_f] - \dot{\mathcal{T}}(f_0, f_1)x_f|_{V^*} \\ + \ & \| \ C(x_0)^{-1} \|_{\mathcal{L}(V^*\times\mathbf{R}, V\times\mathbf{R})} \, |\dot{\mathcal{T}}(f_0, f_1)(x_f - x_0)|_{V^*}.\end{aligned}$$

Due to (H4), (5.3), and the fact that $\| M_\tau \| \to 0$ for $\tau \to 0$, the first term on the right-hand side of (5.11) converges to zero. The second additive term tends to zero as a consequence of (H4). Finally due to (H4) and Proposition 2.2 the last term converges to zero as well. As a consequence of (5.10) and (5.11) it follows that $f \to (\lambda_f, x_f)$ has a Gateaux differential $(\dot{\lambda}, \dot{x})$ at f_0 in direction f_1. It is given by

$$(5.12)\qquad \begin{pmatrix} \dot{x} \\ \dot{\lambda} \end{pmatrix} = -C(x_0)^{-1}\begin{pmatrix} \dot{\mathcal{T}}(f_0, f_1)x_0 \\ 0 \end{pmatrix}.$$

This is (2.2). Uniqueness of the solution to (2.2) follows from injectivity of $C(x_0)$. □

Proof of Corollary 2.5. (i) We need to argue that $f_1 \to (\dot{\lambda}(f_0, f_1), \dot{x}(f_0, f_1))$ is linear [W]. This is a direct consequence of (H5) and (5.12).

(ii) Due to (5.12) and (H6) the mapping

$$(5.13)\qquad f_1 \to \begin{pmatrix} \dot{x}(f_0, f_1) \\ \dot{\lambda}(f_0, f_1) \end{pmatrix} = -C(x_0)^{-1}\begin{pmatrix} \dot{\mathcal{T}}(f_0, f_1)x_0 \\ 0 \end{pmatrix}$$

is an element of $\mathcal{L}(F; V \times \mathbf{R})$. To show that it satisfies the definition of Fréchet differentiability we estimate

$$(5.14)\qquad \begin{pmatrix} x_f - x_0 \\ \lambda_f - \lambda_0 \end{pmatrix} - (-C(x_0))^{-1}\begin{pmatrix} \dot{\mathcal{T}}(f_0, f_1)x_0 \\ 0 \end{pmatrix},$$

where we put $f = f_0 + f_1$. By Proposition 2.2 we have that

$$(\lambda_f, x_f) \to (\lambda_0, x_0) \quad \text{in } \mathbf{R} \times V \text{ if } |f_1|_F \to 0.$$

As in the proof of Theorem 2.4, $C(x_0)$ is an isomorphism from $V \times \mathbf{R}$ to $V^* \times \mathbf{R}$,

$$M(f) = C(x_f) - C(x_0) \to 0 \quad \text{in } \mathcal{L}(V \times \mathbf{R}, V^* \times \mathbf{R}) \text{ as } f \to f_0 \text{ in } F,$$

and $\| \left(I + C(x_0)^{-1}[C(x_f) - C(x_0)]\right)^{-1} \|_{\mathcal{L}(V\times\mathbf{R})}$ is uniformly bounded for all f sufficiently close to f_0. Returning to (5.14) we find (compare (5.8)

$$\begin{aligned}
&\begin{pmatrix} x_f - x_0 \\ \lambda_f - \lambda_0 \end{pmatrix} + C(x_0)^{-1}\begin{pmatrix} \dot{\mathcal{T}}(f_0, f_1)x_0 \\ 0 \end{pmatrix} \\
&\quad = \left\{-[I + C(x_0)^{-1}M(f)]^{-1} + I\right\} C(x_0)^{-1}\begin{pmatrix} \mathcal{T}_f x_f - \mathcal{T}_{f_0} x_f \\ 0 \end{pmatrix} \\
&\qquad -C(x_0)^{-1}\begin{pmatrix} \mathcal{T}_f x_f - \mathcal{T}_{f_0} x_f - \dot{\mathcal{T}}(f_0, f_1)x_0 \\ 0 \end{pmatrix} \\
&\quad = [I + C(x_0)^{-1}M(f)]^{-1}C(x_0)^{-1}M(f)C(x_0)^{-1}\begin{pmatrix} \mathcal{T}_f x_f - \mathcal{T}_{f_0} x_f \\ 0 \end{pmatrix} \\
&\qquad -C(x_0)^{-1}\begin{pmatrix} \mathcal{T}_f x_f - \mathcal{T}_{f_0} x_f - \dot{\mathcal{T}}(f_0, f_1)x_{f_0} \\ 0 \end{pmatrix}
\end{aligned}$$

and therefore

$$\begin{aligned}
(5.15)\qquad &\left\| \begin{pmatrix} x_f - x_0 \\ \lambda_f - \lambda_0 \end{pmatrix} + C(x_0)^{-1}\begin{pmatrix} \dot{\mathcal{T}}(f_0, f_1)x_0 \\ 0 \end{pmatrix} \right\|_{(V\times\mathbf{R})} \\
&\leq \; c_1 \, \| \, M(f) \, \|_{\mathcal{L}(V\times\mathbf{R}, V^*\times\mathbf{R})} \left\| C(x_0)^{-1}\begin{pmatrix} \mathcal{T}_f x_f - \mathcal{T}_{f_0} x_f \\ 0 \end{pmatrix} \right\|_{\mathcal{L}(V\times\mathbf{R})} \\
&\quad + c_2 \Big(\| \, \mathcal{T}_f - \mathcal{T}_{f_0} - \dot{\mathcal{T}}(f_0, f_1) \, \|_{\mathcal{L}(V,V^*)} \| \, x_f \, \|_V \\
&\quad + \| \, \dot{\mathcal{T}}(f_0, f_1) \, \|_{\mathcal{L}(V,V^*)} \| \, x_f - x_0 \, \|_V \Big),
\end{aligned}$$

where

$$c_1 = \| \, (I + C(x_0)^{-1}M(f))^{-1}C(x_0)^{-1} \, \|_{\mathcal{L}(V^*\times\mathbf{R}, V\times\mathbf{R})}$$

and

$$c_2 = \| \, C(x_0)^{-1} \, \|_{\mathcal{L}(V^*\times\mathbf{R}, V\times\mathbf{R})} \, .$$

Since $M(f) \to 0$ in $\mathcal{L}(V\times\mathbf{R}, V^*\times\mathbf{R})$ and due to (H6) the first term on the right-hand side of (5.15) behaves like $o(|f_1|_F)$. Due to (H6) and Proposition 2.3 the second term behaves like $o(|f_1|_F)$ as well. □

Proof of Corollary 2.6. For the proof it suffices to observe that (λ_f^n, x_f^n), $(\lambda_{f_0}^n, x_{f_0}^n)$ satisfy

$$C(x_{f_0}^n)\begin{pmatrix} x_f^n - x_{f_0}^n \\ \lambda_f^n - \lambda_{f_0}^n \end{pmatrix} = \left[C(x_{f_0}^n) - C(x_f^n)\right]\begin{pmatrix} x_f^n - x_{f_0}^n \\ \lambda_f^n - \lambda_{f_0}^n \end{pmatrix} - \begin{pmatrix} \mathcal{T}_f x_f^n - \mathcal{T}_{f_0} x_f^n \\ 0 \end{pmatrix};$$

compare (5.6), where $C(x_{f_0}^n) : V \times \mathbf{R} \to V^* \times \mathbf{R}$ is given by

$$C(x_f^n) = \begin{pmatrix} A_{f_0} - \lambda_0^n & -x_f^n \\ \langle x_{f_0}^n, \cdot \rangle & 0 \end{pmatrix}.$$

It is simple to check that $C(x_{f_0}^n)$ is an isomorphism. The assertions then follow with the same arguments as for Theorem 2.4 and Corollary 2.5. □

Proof of Lemma 2.7. Let $f_1 \in F$ and $\tau \in \mathbf{R}$, and consider

$$\begin{aligned}
&\left\| \frac{1}{\tau} \left(\mathcal{T}_{f_0+\tau f_1} - \mathcal{T}_{f_0} \right) - \dot{\mathcal{T}}(f_0, f_1) \right\|_{\mathcal{L}(V,V^*)} \\
&\quad = \sup_{|\varphi|_V=|\psi|_V=1} \left| \frac{1}{\tau} \langle T_{f_0+\tau f_1}\varphi, T_{f_0+\tau f_1}\psi \rangle - \frac{1}{\tau} \langle T_{f_0}\varphi, T_{f_0}\psi \rangle \right. \\
&\qquad\qquad \left. - \langle \dot{T}(f_0, f_1)\varphi, T_{f_0}\psi \rangle - \langle T_{f_0}\varphi, \dot{T}(f_0, f_1)\psi \rangle \right| \\
&\quad = \sup_{|\varphi|_V=|\psi|_V=1} \left| \left\langle \frac{1}{\tau} (T_{f_0+\tau f_0}\varphi - T_{f_0}\varphi) - \dot{T}(f_0, f_1)\varphi, T_{f_0+\tau f_1}\psi \right\rangle \right. \\
&\qquad\qquad - \langle \dot{T}(f_0, f_1)\varphi, T_{f_0+\tau f_1}\psi - T_{f_0}\psi \rangle \\
&\qquad\qquad \left. + \left\langle T_{f_0}\varphi, \frac{1}{\tau} (T_{f_0+\tau f_1}\psi - T_{f_0}\psi) - \dot{T}(f_0, f_1)\psi \right\rangle \right|.
\end{aligned}$$

From (A1) the expression on the right-hand side of the last equality converges to zero for $\tau \to 0$. Thus $f \to \mathcal{T}_f$ is Gateaux differentiable at f_0 in every direction $f_1 \in F$. From (2.3) it follows that $f_1 \to \dot{\mathcal{T}}(f_0, f_1)$ is linear, and therefore $f \to \mathcal{T}_f$ has a Gateaux derivative at f_0. □

Proof of Lemma 2.8. From (2.4) it is easily seen that $\dot{\mathcal{T}}(f_0, \cdot) \in \mathcal{L}(F, \mathcal{L}(V, V^*))$. To argue Fréchet differentiability at f_0 it suffices to demonstrate continuity of $f \to \dot{\mathcal{T}}(f, \cdot)$ from F to $\mathcal{L}(F, \mathcal{L}(V, V^*))$ at f_0. For $\hat{f}_0$ in a neighborhood of f_0 we have

$$\begin{aligned}
&\| \dot{\mathcal{T}}(f_0, \cdot) - \dot{\mathcal{T}}(\hat{f}_0, \cdot) \|_{\mathcal{L}(F,\mathcal{L}(V,V^*))} \\
&\quad = \sup_{|f_1|_F=|\varphi|_V=|\psi|_V=1} |\langle \dot{T}(f_0, f_1)\varphi - \dot{T}(\hat{f}_0, f_1)\varphi, T_{f_0}\psi \rangle_Y \\
&\qquad\qquad + \langle \dot{T}(\hat{f}_0, f_1)\varphi, T_{f_0}\psi - T_{\hat{f}_0}\psi \rangle_Y + \langle T_{f_0}\varphi - T_{\hat{f}_0}\varphi, \dot{T}(f_0, f_1)\psi \rangle_Y \\
&\qquad\qquad + \langle T_{\hat{f}_0}\varphi, \dot{T}(f_0, f_1)\psi - \dot{T}(\hat{f}_0, f_1)\psi \rangle_Y |,
\end{aligned}$$

and (A2) implies the claim. □

Acknowledgment. The authors would like to acknowledge the comments of the anonymous referee on an earlier version of this paper. The authors are thankful to M. Jaksche for his help with the MATLAB codes that were used to carry out the computations for this section.

REFERENCES

[A] R. A. Adams, *Sobolev Spaces*, Academic Press, New York, 1985.

[BBC] J. V. Beck, B. Blackwell, and C. R. St. Clair, *Inverse Heat Conduction, Illposed Problems*, Wiley-Interscience, New York, 1985.

[C] F. Chatelin, *Spectral Approximation of Linear Operators*, Academic Press, New York, 1989.

[DL] R. Dautray and J. L. Lions, *Mathematical Analysis and Numerical Methods for Science and Technology*, Vols. II, III, Springer, Berlin, 1990.

[IK1] K. Ito and K. Kunisch, *On the injectivity of the coefficient-to-solution mapping for elliptic boundary value problems and its linearization*, J. Math. Anal. Appl., to appear.

[IK2] K. Ito and K. Kunisch, *Sensitivity measures for the estimation of parameters in elliptic boundary value problems*, J. Math. Syst. Estimation Control, to appear.

[K] K. KATO, *Perturbation Theory of Linear Operators*, 2nd ed., Springer-Verlag, Berlin, 1980.

[M] V. G. MAZ'JA, *Sobolev Spaces*, Springer-Verlag, Berlin 1985.

[T] G. M. TROIANIELLO, *Elliptic Differential Equations and Obstacle Problems*, Plenum Press, New York, 1987.

[W] A. WOUK, *A Course of Applied Functional Analysis*, John Wiley and Sons, New York, 1979.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 33, No. 2, pp. 667–671, March 1995

© 1995 Society for Industrial and Applied Mathematics

016

CORRIGENDUM: LAGRANGE MULTIPLIERS IN STOCHASTIC PROGRAMMING*

SJUR DIDRIK FLÅM†

Abstract. This note deals with an error in S. D. Flåm, *SIAM J. Control Optim.*, 30 (1992), pp. 1–10. Using approximate subdifferentials the author corrects that error by deriving a Fritz John optimality condition for abstract programs constrained by cones which may have empty interior.

Key words. Fritz John optimality condition, nonsmooth analysis, approximate subdifferential, compactly Lipschitzian mappings, stochastic programming

AMS subject classifications. 90C30, 49J52, 26A27

1. Introduction. Several scholars, notably M. Ciligot-Travain and L. Thibault, have quickly and kindly pointed out that Theorem 3.2 in [7] concerning the Fritz John (FJ) optimality condition needs further qualification. Also, Glover and Craven [8] provide a counterexample to the theorem. As it stands, the theorem is valid only in finite dimensions (implying that the probability measure should have finite support), and would then result from earlier theorems of Clarke and Hiriart-Urruty [11]. In that context it can also be sharpened by replacing the Clarke subdifferential by the generally tighter approximate subdifferential of Ioffe [13], extensively used by Jourani and Thibault [16], [17] and Glover [9].

The error is that the set $\{\lambda \geq 0 : \rho(\lambda) \leq 1\}$ of multipliers λ used in [7] with $\rho(\cdot)$ stemming from a Clarkson–Rieffel renorming need not be bounded.

Therefore, a primary purpose of this note is to rectify these matters; a secondary aim is to complement the recent results in [8] and [10]. Thus, a new FJ condition is demonstrated in Theorem 2 below, partly motivated by stochastic programming problems, but of independent interest and applicable in other fields as well.

To make the note self-contained, we shall deal with the abstract and generic programming problem

$$\text{(P)} \qquad \text{minimize } f(x) \quad \text{subject to } x \in C \quad \text{and} \quad -g(x) \in K,$$

where $f : X \longrightarrow \mathbb{R}$, $g : X \longrightarrow Y$ are locally Lipschitz functions, X, Y are Banach spaces with Y separable, $C \subset X$ is closed, and $K \subset Y$ is a closed convex cone. In the topological dual space Y' of Y, denote by

$$\Lambda := \{y' \in Y' : \langle y, y' \rangle \geq 0 \quad \text{for all } y \in K\}$$

the positive dual cone of K. The following theorem, recently proven in [6, Thm. 2.6] and [10, Thm. 5], extends a similar result in [8] and serves as a point of reference.

THEOREM 1. *For problem* (P) *assume f is locally Lipschitz and g is locally compactly Lipschitz. Also, assume that K has nonempty interior. Then a necessary condition for x_0 to be a local minimum of* (P) *is that there exists* $\tau_0 \in [0,1]$, $\lambda_0 \in \Lambda$, $\|\lambda_0\| \leq 1$, $(\tau_0, \lambda_0) \neq 0$ *such that*

$$(1.1) \qquad 0 \in \tau_0 \partial_A f(x_0) + (1-\tau_0)\partial_A(\lambda_0 g)(x_0) + N_G(x_0, C), \qquad \langle \lambda_0, g(x_0) \rangle = 0.$$

Here ∂_A denotes the *approximate subdifferential*— developed by Ioffe [13], [14], [15], and Mordukhovich [20]—known to be weak* compact, occasionally nonconvex but always contained, and possibly strictly within the Clarke subdifferential. In (1.1) $N_G(x, C)$ is the normal cone to C at x in the sense that it equals the weak* closure of $\mathbb{R}_+\partial_A d(x, C)$, where

* Received by the editors June 30, 1994; accepted for publication July 26, 1994.

† Economics Department, Bergen University, 5007 Bergen, Norway.

$d(x, C) := \inf\{\|x-c\| : c \in C\}$ is the distance function to C; see [15]. Note that in Theorem 1 and throughout most of the section, the constraint function g must be *locally compactly Lipschitzian*—a notion introduced by Thibault [21], [22].

The assumption $\operatorname{int} K \neq \emptyset$ of Theorem 1 is often unfitting, however, it excludes equality constraints and it is at variance with stochastic programming and other fields where cones K like $L^p(\Omega, \mathbb{R}^m_+)$, $p \in [1, +\infty[$ may naturally arise. Admittedly, progress can sometimes be made without constraint qualifications; see [1] and [3]. However, it appears difficult to obtain results like Theorem 1 for spaces like $Y = L^p(\Omega, \mathbb{R}^m_+)$, $p \in [1, +\infty[$ unless some condition is placed on the criterion f also.

Therefore, we shall impose what we name an FJ *qualification*, motivated as follows. In his proof of the FJ condition, Clarke [2, Thm. 6.1.1] maximized for a real $\varepsilon > 0$ a Lagrangian-type expression over a nonempty weak*-compact set M of multipliers to obtain the desired conclusion by letting $\varepsilon \downarrow 0$. Two things, among others, were important then and in subsequent extensions of Clarke's technique [4], [8], [10] as well. First, in order to work with converging sequences of multipliers, M should be sequentially compact. This is ensured by the separability of Y. Second, limiting multipliers must be nonzero. For this property it suffices that $\operatorname{int} K \neq \emptyset$, but what seems more natural here is to let the entire problem structure, not only K, help in avoiding degenerate multipliers. Therefore we introduce the following.

FRITZ JOHN QUALIFICATION AT A POINT x_0. *There exists a w^*-compact convex set $M \subset \Lambda$ of multipliers such that*

$$\langle \lambda, y\rangle \le 0 \quad \textit{for all } \lambda \in M \Rightarrow -y \in K, \tag{1.2}$$

and such that among all sequences $(\varepsilon, x_\varepsilon, \lambda_\varepsilon, \tau_\varepsilon)$ *(omitting, for notational convenience, reference to the sequence index) satisfying* $\varepsilon \downarrow 0$ *in* $\mathbb{R}$, $x_\varepsilon \rightarrow x_0$ *strongly in* C, $\lambda_\varepsilon \rightarrow \lambda_0$ *weak* in* M, $1 \ge \tau_\varepsilon \rightarrow 0$ *in* $\mathbb{R}$,

$$\langle \lambda_\varepsilon, g(x_\varepsilon)\rangle = \max\{\langle m, g(x_\varepsilon)\rangle : m \in M\} \ge f(x_\varepsilon) - f(x_0) + \varepsilon > 0, \tag{1.3}$$

and

$$0 \in \tau_\varepsilon \partial_A f(x_\varepsilon) + (1-\tau_\varepsilon)\partial_A(\lambda_\varepsilon g)(x_\varepsilon) + \sqrt{\varepsilon} B^* + N_G(x_\varepsilon, C), \tag{1.4}$$

there exists, if any, at least one sequence for which the limiting multiplier λ_0 is nonzero.

We remark that frequently the burden of ensuring $\lambda_0 \neq 0$ is entirely carried by K or Y.

PROPOSITION 1. *The* FJ *qualification is satisfied if one of the conditions below holds*:

(i) $\operatorname{int} K \neq \emptyset$;

(ii) Y *is finite-dimensional*;

(iii) K *is the product of one cone having nonempty interior and another being finite dimensional*;

(iv) *There exists a weak* upper semicontinuous function* $c : \Lambda \rightarrow \mathbb{R}$, *and a number* χ *such that the set*

$$\{\lambda \in \Lambda : \|\lambda\| \le 1 \textit{ and } c(\lambda) \ge \chi\} \tag{1.5}$$

is nonempty, convex, does not contain 0, *and satisfies* (1.2).

Proof. In cases (i), (ii), and (iii), take $M = \{\lambda \in \Lambda : \|\lambda\| \le 1\}$. Then (1.3) implies $\|\lambda_\varepsilon\| = 1$, and any sequence $\{\lambda_\varepsilon\}$ contains a weak*-convergent subsequence with nonzero limit; see [8]. In case (iv), take M to equal the set defined in (1.5). □

Remark. In problem (P) one may accommodate more general sets K than closed convex cones. Also, to satisfy the FJ qualification it suffices that K be epi-Lipschitz-like at $-g(x_0)$. For details see Jourani [9].

2. Main result. We now state the announced result on the FJ condition, quite parallel to Theorem 1 and applicable to stochastic programming.

THEOREM 2. *For problem* (P) *assume f is locally Lipschitz, g is locally compactly Lipschitz, and the* FJ *qualification is satisfied at x_0. Then a necessary condition for x_0 to be a local minimum of* (P) *is that there exists $\tau_0 \in [0,1]$, $\lambda_0 \in M$, not both zero, such that* (1.1) *holds.*

Proof. The argument follows that of [10] with small modifications, but is given for completeness. Define a continuous sublinear function $k : Y \to \mathbb{R}$ by $k(y) = \max\{\langle m, y\rangle : m \in M\}$, where M is the nonempty set defined in the above FJ qualification. For any positive number ε consider the function

$$h_\varepsilon(x) := \max\{f(x) - f(x_0) + \varepsilon,\ k(g(x))\}$$

mapping X into $\mathbb{R}$. Observe that $h_\varepsilon(x) \leq 0$ implies $k(g(x)) \leq 0$; then (1.2) yields $-g(x) \in K$. It follows that $h_\varepsilon(x) > 0$ for all $x \in C$ sufficiently near x_0. Indeed, otherwise there would exist $x \in C$, arbitrarily close to x_0, satisfying $f(x) \leq f(x_0) - \varepsilon$ and $-g(x) \in K$. If so, this contradicts the local optimality of x_0. Now, the fact that $h_\varepsilon(x_0) = \varepsilon$ tells us that x_0 furnishes a local ε-minimum of h_ε, provided that this function is restricted to C. Since h_ε is lower semicontinuous, we may find, by Ekeland's variational principle [5], a point $x_\varepsilon \in C$ within $\sqrt{\varepsilon}$-distance from x_0 such that

$$h_\varepsilon(x_\varepsilon) < h_\varepsilon(x) + \sqrt{\varepsilon}\|x - x_\varepsilon\|$$

for all $x \in C$ different from x_ε and sufficiently close to x_0. In fact, a fortiori h_ε is Lipschitz near x_0 with some modulus L (independent of ε). It follows, therefore, by the exact penalization result of Clarke [2, Prop. 2.4.3], that the function

$$h_\varepsilon(x) + \sqrt{\varepsilon}\|x - x_\varepsilon\| + (L+\varepsilon)d(x, C)$$

attains an unconstrained local minimum at x_ε. Thus, since Fermat's rule applies to ∂_A, we get

$$\begin{aligned} 0 &\in \partial_A[h_\varepsilon(\cdot) + \sqrt{\varepsilon}\| \cdot - x_\varepsilon\| + (L+\varepsilon)d(\cdot, C)](x_\varepsilon) \\ &\subset \partial_A h_\varepsilon(x_\varepsilon) + \sqrt{\varepsilon}B^* + (L+\varepsilon)\partial_A d(x_\varepsilon, C), \end{aligned} \tag{2.6}$$

where B^* denotes the closed unit ball in Y'. We proceed to invoke the following chain rule of Jourani and Thibault [18].

Let $g : X \to Y$ be locally compactly Lipschitzian at x and let $k : Y \to \mathbb{R}$ be locally Lipschitz at $g(x)$. Then kg is locally Lipschitz at x and

$$\partial_A(kg)(x) \subset \bigcup\{\partial_A(\lambda g)(x) : \lambda \in \partial_A k(g(x))\}. \tag{2.7}$$

Letting k and g be as above, this chain rule (2.2) yields

$$\partial_A(kg)(x) \subset \bigcup\{\partial_A(\lambda g)(x) : \lambda \in M,\ \langle \lambda, g(x)\rangle = kg(x)\}. \tag{2.8}$$

The same rule (2.2) also implies that

$$\partial_A h(x) \subset \operatorname{conv}\{\partial_A h_i(x) : h_i(x) = h(x)\}, \tag{2.9}$$

when $h(x) := \max\{h_1(x), h_2(x)\}$ for locally Lipschitz functions $h_i : X \to \mathbb{R}$. For details on these last two assertions see [8]. Employing (2.3)–(2.4) in (2.1) we find $\tau_\varepsilon \in [0,1]$ and $\lambda_\varepsilon \in M$ such that

$$0 \in \tau_\varepsilon \partial_A f(x_\varepsilon) + (1 - \tau_\varepsilon)\partial_A(\lambda_\varepsilon g)(x_\varepsilon) + \sqrt{\varepsilon}B^* + (L+\varepsilon)\partial_A d(x_\varepsilon, C). \tag{2.10}$$

Now let $\varepsilon \downarrow 0$ to have $x_\varepsilon \to x_0$ strongly. We consider two cases.

Case 1. Assume $kg(x_\varepsilon) < f(x_\varepsilon) - f(x_0) + \varepsilon$ for some sequence $\varepsilon \downarrow 0$. Then $h_\varepsilon(x_\varepsilon) = f(x_\varepsilon) - f(x_0) + \varepsilon > 0$, so that $\tau_\varepsilon = 1$ by (2.4), and we may select $\lambda_\varepsilon = 0$ in (2.5). Passing to the limit in (2.5) we obtain (1.1) with $\tau_0 = 1$, $\lambda_0 = 0$.

Case 2. $kg(x_\varepsilon) \geq f(x_\varepsilon) - f(x_0) + \varepsilon > 0$ along every sequence $\varepsilon \downarrow 0$. Associated with any such sequence $\varepsilon \downarrow 0$, there are numbers $\tau_\varepsilon \to \tau_0 \in [0,1]$, and the multipliers $\lambda_\varepsilon \in M$ satisfying $\langle \lambda_\varepsilon, g(x_\varepsilon)\rangle = kg(x_\varepsilon)$ by (2.3). Since M is weak* sequentially compact (because Y is separable, see [12]), we may suppose $\lambda_\varepsilon \to \lambda_0 \in M$. We claim that limiting λ_0 satisfies the complementarity condition in (1.1). In fact, $\langle \lambda_0, g(x_0)\rangle \leq 0$ holds trivially, and the converse inequality follows from two facts: first, $\langle \lambda_\varepsilon, g(x_\varepsilon)\rangle = kg(x_\varepsilon) \to kg(x_0) \geq 0$, because $kg(x_\varepsilon) > 0$, and second, $\langle \lambda_\varepsilon, g(x_\varepsilon)\rangle \to \langle \lambda_0, g(x_0)\rangle$.

Next, for the inclusion in (1.1), observe that the correspondence $(\lambda, x) \to \partial_A(\lambda g)(x)$ has a closed graph in the (weak* × strong) product topology, see [6, Lem. 2.5] or [10].

Therefore, limit (2.5) gives the desired inclusion (1.1). Thus, it remains in this case only to exclude the degenerate outcome $(\tau_0, \lambda_0) = 0$. Clearly, if for some sequence $\varepsilon \downarrow 0$ it holds that $\tau_\varepsilon \to \tau_0 > 0$, then we are done. Otherwise the FJ qualification comes into play, ensuring that some $\lambda_0 \neq 0$. This completes the proof. □

Following [10] we remark that if g is *K-convex* in the sense that

$$\mu_1, \mu_2 \geq 0,\ \mu_1 + \mu_2 = 1 \Rightarrow \mu_1 g(x_1) + \mu_2 g(x_2) - g(\mu_1 x_1 + \mu_2 x_2) \in K \quad \forall x_1,\ x_2 \in X,$$

then the assumption that g is locally compactly Lipschitzian often becomes redundant.

COROLLARY. *For problem* (P) *assume that f, g are locally Lipschitz, g is K-convex, and the* FJ *qualification is satisfied at x_0. If Y has a strictly convex dual norm, then a necessary condition for x_0 to be a local minimum of* (P) *is that there exists $\tau_0 \in [0,1]$, $\lambda_0 \in M$, not both zero, such that* (1.1) *holds.*

REFERENCES

[1] J. M. Borwein and H. Wolkowicz, *Characterizations of optimality without constraint qualification for the abstract convex program*, Math. Programming Study, 19 (1982), pp. 77–100.

[2] F. H. Clarke, *Optimization and Nonsmooth Analysis*, John Wiley, New York, 1983.

[3] B. D. Craven, *Mathematical Programming and Control Theory*, Chapman and Hall, London, 1978.

[4] P. H. Dien, *On the regularity condition for the extremal problem under locally Lipschitz inclusion constraints*, Appl. Math. Optim., 13 (1985), pp. 151–161.

[5] I. Ekeland, *On the variational principle*, J. Math. Anal. Appl., 47 (1974), pp. 324–353.

[6] B. El Abdouni and L. Thibault, *Lagrange multipliers for Pareto nonsmooth programming problems in Banach spaces*, Optimization, 26 (1992), pp. 277–285.

[7] S. D. Flåm, *Lagrange multipliers in stochastic programming*, SIAM J. Control Optim., 30 (1992), pp. 1–10.

[8] B. M. Glover and B. D. Craven, *A Fritz John optimality condition using the approximate subdifferential*, Dept. of Math., Univ. of Melbourne, 1992, preprint.

[9] B. M. Glover, *Locally compactly Lipschitzian mappings in infinite dimensional programming*, Bull. Austral. Math. Soc., (1992).

[10] B. M. Glover, B. D. Craven, and S. D. Flåm, *A generalized Karush–Kuhn–Tucker optimality condition without constraint qualification using the approximate subdifferential*, Numer. Funct. Anal. Optim., 14 (1993), 333–353.

[11] J.-B. Hiriart-Urruty, *Refinements of necessary optimality conditions in nondifferentiable programming*. Appl. Math. Optim., 5 (1979), pp. 63–82.

[12] R. B. Holmes, *Geometric Functional Analysis and Applications*, Springer-Verlag, Berlin, 1975.

[13] A. D. Ioffe, *Approximate subdifferentials and applications* I: *The finite dimensional theory*, Trans. Amer. Math. Soc., 281 (1984), pp. 389–416.

[14] A. D. Ioffe, *Approximate subdifferentials and applications* II: *Functions on locally convex spaces*, Mathematika, 36 (1986), pp. 112–128.

[15] A. D. Ioffe, *Approximate subdifferentials and applications* III: *The metric theory*, Mathematika, 36 (1989), 1–38.

[16] A. Jourani and L. Thibault, *Approximate subdifferentials and metric regularity: The finite dimensional case*, Math. Programming, 47 (1990), pp. 203–218.

[17] A. Jourani and L. Thibault, *The use of metric graphical regularity in approximate subdifferential calculus rules in finite dimensions*, Optimization, 21 (1990), pp. 509–519.

[18] A. Jourani and L. Thibault, *Approximate subdifferential of composite functions*, Bull. Austral. Math. Soc., (1992).

[19] A. Jourani, *Qualification conditions for multivalued functions in Banach spaces with application to vector optimization problems*, Université de Bourgogne, Dijon, 1993 working paper.

[20] B. S. Mordukhovich, *Nonsmooth analysis with nonconvex generalized differentials and dual maps*, Dokl. Akad. Nauk. BSSR, 28 (1984), pp. 976–979.

[21] L. Thibault, *Subdifferentials of compactly Lipschitzian vector-valued functions*, Travaux du Séminaire d'analyse convexe, Vol. 8, Fascicule I, Montpellier, 1978.

[22] L. Thibault, *Subdifferentials of nonconvex vector-valued functions*, J. Math. Anal. Appl., 86 (1982), pp. 319–354.

© 1995 Society for Industrial and Applied Mathematics

THE RENDEZVOUS SEARCH PROBLEM[*]

STEVE ALPERN[†]

Abstract. The author considers the problem faced by two people who are placed randomly in a known search region and move about at unit speed to find each other in the least expected time. This time is called the rendezvous value of the region. It is shown how symmetries in the search region may hinder the process by preventing coordination based on concepts such as north or clockwise. A general formulation of the rendezvous search problem is given for a compact metric space endowed with a group of isometries which represents the spatial uncertainties of the players. These concepts are illustrated by considering upper bounds for various rendezvous values for the circle and an arbitrary metric network. The discrete rendezvous problem on a cycle graph for players restricted to symmetric Markovian strategies is then solved. Finally, the author considers the problem faced by two people on an infinite line who each know the distribution of the distance but not the direction to each other.

Key words. rendezvous, search

AMS subject classifications. 90B40, 90D26

1. Introduction. The purpose of this paper is to introduce and rigorously formulate a new form of cooperative optimal search, which we call the rendezvous search problem. This is the problem faced by two people placed in a known region according to some known distribution who wish to meet in minimum time. They attempt to find each other by moving with speed bounded by one until the first meeting time T, when the distance between them comes within a given detection radius. The least expected meeting time that can be achieved by allowable search strategies is called the rendezvous value R of the search region X. We give a formal definition of the rendezvous search problem, find upper bounds on R for certain regions or classes of regions, and consider a few related problems of rendezvous on the line and on a cycle graph. In addition, we give a number of questions that we hope will stimulate the development of a theory of rendezvous search.

As we shall see, there are in fact several different rendezvous values depending on which search strategies are allowed. We shall be primarily concerned with the symmetric rendezvous value R^s obtained by restricting the two players to using the same mixed search strategy. This corresponds to the case where the players have not previously agreed which role each should take (e.g., wait or search) in the event they become separated, or where, after becoming separated, they both take directions from a controller who does not know their names. Without this restriction we have the asymmetric rendezvous value R^a, which in general is smaller. In addition, the rendezvous value will depend on geometric aspects of the information each player has about his position in a symmetrical region X, which we will formalize by specifying a particular subgroup G of the isomorphism group Γ of the metric space (X, ρ). For example, the players may be lost on a circle, have no common notion of where $\theta = 0$ is, but have a common notion of clockwise (i.e., both know which way is up). In this case a strategy of *head for* $\theta = 0$ would not be allowable, but *go clockwise* would. In other words, the players may have some uncertainty about how to fit the picture of the search region that they see to a map common to both. It is the aim of this paper to formalize these notions of spatial symmetry, to give a rigorous definition of the rendezvous value(s), and to provide some estimates or determinations of R^s in some simple cases. We also suggest some extensions to the theory given here that we hope will stimulate further work on rendezvous search theory.

From a formal point of view rendezvous problems are very close to search games with mobile hiders (also called *princess and monster games*). As proposed in Isaacs's book on

[*] Received by the editors May 12, 1993; accepted for publication (in revised form) November 18, 1993.

[†] Department of Mathematics, London School of Economics, University of London, Houghton Street, London, England WC2A 2AE.

differential games [12], and extensively analyzed in Gal's book on search games [9] and in the literature [1], [2], [3], [4], [11], [10], [6], [13], [17], the latter are zero-sum games where the meeting time T is the payoff to the maximizing hider. One direct comparison is that the value of that game (called the search value) is obviously an upper bound for the rendezvous values. Our work on rendezvous search on the line is also closely related to the work of Beck (e.g., [7], [8]) on the linear search problem.

No introduction to a work on rendezvous search would be complete without reference to the examples and analysis given by Schelling in his book on the strategy of conflict [16]. His paradigm is the problem facing two parachutists who land in a field with various roads, buildings, a river, and a single bridge. He suggests that the essential problem is to find the unique *focal point*, which in this case is generally thought to be the bridge. The implication is that without focal points the problem cannot be formulated, let alone solved. Our perspective is diametrically opposed to Schelling's. We ask how to rendezvous when the search region is homogeneous. The difference in our analysis that enables us to tackle (sometimes) such questions is that while Schelling considered one-shot strategies (go to point x), we consider multistage strategies that continue to search after initial failed attempts to rendezvous. This is not to suggest that focal points are not important in practice—surely they are—but we hope to show that even in the absence of focal points rendezvous search is more than purely random wanderings.

The paper is organized as follows. In §2 we define the rendezvous value for a compact metric space endowed with a given subgroup of its isometries which reflects the players' spatial uncertainties. In §3 we give upper bounds for the rendezvous values of a metric network. In §4 we illustrate the definitions of §2 involving symmetries by considering various rendezvous values for the circle with respect to different isometry groups. Up to this point all the definitions and examples relate to continuous motion in continuous time, which is our main interest. However, interesting search problems can also be formulated on graphs with the players moving in integer time to adjacent nodes of a graph. An outstanding result in this direction is the paper of Anderson and Weber [5], which gives asymptotic results for the symmetric rendezvous value of a complete graph. In §5 we find the symmetric rendezvous value for the cycle graph on m nodes when the players are restricted to symmetric Markovian strategies. This result is a rendezvous version of the elegant analysis of Ruckle for the (zero-sum) search game with the same dynamics [14], [15]. In §6 we consider rendezvous search on the real line, where the players know the distribution of the distance between them but do not know the direction of the other player. These problems are a rendezvous version of the linear search problem [7], [8] mentioned above. We conclude with a short final section on suggestions for further research on rendezvous search.

2. Rendezvous on a compact metric space. In this section we define the rendezvous search problem (and associated values R^s and R^a) for a compact metric space (X, ρ) with a given detection radius δ and a given group G of isometries (distance preserving bijections) of X. Since it is the last of these elements that requires the most motivation, a brief discussion of the isometry group and its connection with spatial symmetries precedes the formal argument. Note that throughout this paper we are assuming that the players can see the whole region X but cannot see the other player. (Without this assumption we have rendezvous search in a maze, which will be the subject of another paper.)

Suppose the search region X is a plane (it's not compact, but that won't matter for this example) with a perfectly straight canal running through it, and a single bridge over the canal. If the water is flowing, then the two players could each determine their exact position on a map common to both. Thus a search strategy could, in principle, depend on the exact point where they are initially placed (where the parachutists land). Now suppose the bridge is not present.

The players can now tell their distance from the canal and which side of the canal they are on (the side from which the river flows right or left). The group G of isometries which describes this lack of information is the real line $\Re$, which acts on the planar region by translations in the direction of the canal. Each real number corresponds to a possible placement of the bridge along the canal. If the canal is the y-axis on the map, then each player's information on landing is his x-coordinate, and his information space $\tilde{X}$ is thus $\Re = X/G = \Re^2/\Re$, where the / indicates the equivalence classes of the set on the left via the action of the group on the right as defined below. Now suppose the river no longer flows. The players now know only their distance to the river, not which side they are on. Their information is now in $\tilde{X} = \Re^+ = [0, \infty)$, which equals X/G, where G now also includes the reflection isometry across the canal. Next, take away the canal. Now the region X is a featureless plane where all points look the same. Let us assume they landed on Earth so they know where up is. Then the relevant isometry group consists of translations and rotations of X, and the information space $\tilde{X}$ is a singleton, meaning no information is obtained upon landing. However, since they agree on up, they can coordinate to the extent of using, say, clockwise spirals in the symmetric game or opposite directional spirals in the asymmetric game. Finally, if we consider two astronauts landing on a planar region of space, perhaps from opposite sides, then they no longer have a common up, and a reflection of the plane must be added to the isometry group G. They can no longer agree on the meaning of clockwise. Note that in both of the last two scenarios the information space $\tilde{X}$ is trivial, but more strategies are available in the former than the latter. To summarize, the given group G represents the players' uncertainty (or information set) when trying to fit their view of the search region and their position on it onto a map common to both. We apologize to mathematicians for this lengthy discussion, but it is possible that these ideas are new to some who could otherwise follow the arguments of the paper.

To begin the formal analysis, we first define the set of paths P that the players may use for searching by $P = \{p : \Re^+ \to X, \rho(p(t_1), p(t_2)) \le |t_1 - t_2|\}$. The subset of paths which start at a point x is denoted by P_x. The *meeting time* $T : P \times P \to \Re^+$ of two paths p, q is defined by

$$T(p, q) = \min\{t : \rho(p(t), q(t)) \le \delta\}, \tag{1}$$

where δ is the given detection radius. We note that in most of our examples the space X is one-dimensional, and we will take $\delta = 0$. The group G induces an equivalence relation on X and P by $x \sim y \Leftrightarrow \exists g \in G,\ g(x) = y$ and $p \sim q \Leftrightarrow \exists g \in G, g(p(t)) \equiv q(t)$. Call the associated sets of equivalence classes $\tilde{X}$ and $\tilde{P}$, and let [ ] denote the equivalence class of an element. A search strategy is a map $s : \tilde{X} \to \tilde{P}$ such that there is some path $p \in s(x)$ with $p(0) = x$, and the set of all search strategies will be denoted by S. It is worth noting two important cases in which this formalism simplifies considerably. If there is no group of isometries given (or formally, if G is just the identity transformation), then a strategy is simply a map $s : X \to P$ satisfying $s(x) \in P_x$. If the group G acts transitively on X ($\forall x, y \in X, \exists g \in G : g(x) = y$), then $\tilde{X}$ is a singleton and $S = \tilde{P}$. The next problem is how to evaluate the (expected) meeting time for pairs of equivalence classes of paths. Let ν denote the Haar measure on G and define

$$\tilde{T}([p], [q]) = \int_G T(gp, q)\, d\nu(g). \tag{2}$$

The reason we must only have G act on one of the paths is that for $g,\ h \in G$, $T(gp, hq) = T(h^{-1}gp, q)$. To obtain the normal form for the search game, assume players are placed in X independently, according to the same measure μ. Then the normal form is given by the function $\hat{T} : S \times S \to \Re$ defined by

$$\hat{T}(s_1, s_2) = \int_{x \in X} \int_{y \in X} \tilde{T}(s_1(x), s_2(y))\, d\mu(x)\, d\mu(y). \tag{3}$$

The symmetric rendezvous search problem is finding the search strategy which, when used by both players, minimizes the expected meeting time. For some search problems (X, G) it may be that when the same path class is chosen by both players, for some choices of g in (2) the players' initial distance is preserved in time (e.g., if both go at speed 1 clockwise on the circle and g is the identity). The result may be an infinite expected meeting time. To get out of this enforced symmetry, the players can choose the same *mixed* strategy, breaking the symmetry by independent randomization. Thus even though this is essentially a one-person decision problem, randomization may be necessary. With this in mind, let S^* denote the set of all mixed search strategies, i.e., the set of all probability Borel measures on S. The topology on S is defined by giving P the topology of uniform convergence on compact sets. We then define the symmetric rendezvous value of the search problem (X, G, δ) by

$$R^s = R^s(X, G, \delta) = \min_{s^* \in S^*} \int_S \int_S \hat{T}(s_1, s_2)\, ds^*(s_1)\, ds^*(s_2). \tag{4}$$

The use of min rather than inf can be justified by observing that the meeting time function T is lower semicontinuous in each variable on the path space P when the latter is given the topology of uniform convergence on compact sets. The full argument is the same as for search games as in [9]. For the asymmetric rendezvous problem there is no need for mixed strategies, so the asymmetric rendezvous value is given by

$$R^a = R^a(X, G, \delta) = \min_{r,s \in S} \hat{T}(r, s). \tag{5}$$

Most of this paper is devoted to examining the efficacy of various search strategies and establishing upper bounds on the rendezvous values for particular or specific search spaces.

2.1. Homogeneous search regions. In most of the problems to be analyzed in the paper, the group G will act transitively on the search region X. In such cases we will say that X is homogeneous. (Roughly speaking, the region looks the same viewed from any point.) In these cases the definitions given above can be simplified. First, we may take the distributions of the players as given by the one induced by the Haar measure ν on G. In the cases of the circle, torus, or sphere, this will give a uniform distribution (normalized Lebesgue measure). The pure strategy space can be identified with the space of equivalence classes of paths $\tilde{P}$. Thus the normal form is given simply by (2).

2.2. Relation to search games. The reader familiar with search games with mobile hiders may wonder why the spatial symmetries of the search region are not so important in that theory. In search games, it is usually assumed that the hider and searcher can pick their starting points, hence their entire paths. If there is a symmetry, say rotationally on the circle, either player can symmetrize his strategy. If one player symmetrizes, it is the same as if the game had been played with a symmetrical initial placement. Now in a zero-sum game, if there is a condition (such as symmetrical-uniform initial placement) which *either* player can force, then there is no loss in generality if it is assumed. Hence there is no need to make this condition part of the definition of the game.

3. Rendezvous on a network. There are some classes of search strategies which may be applied to a whole family of search regions. They may be used to obtain general bounds on the rendezvous value (or the search game value V) for the whole family. For example, if the search region is a network of unit length arcs, the following elementary estimates are easily established.

PROPOSITION 1. *Let (X, ρ) be a connected network consisting of m unit arcs which intersect only at endpoints. Let G be the group of all isometries of (X, ρ), the detection radius δ be zero, and D denote the metric diameter of X. Then*

1. $V(X,G) \leq 2mD \leq 2m^2$,
2. $R^s(X,G) \leq 4m$, *and*
3. $R^a(X,G) \leq m$.

Proof. To obtain the first estimate, consider the following search strategy. In time intervals of length D, starting at time zero, search as follows: Pick a random arc and choose a path which arrives at the center of this arc in time D and one of the endpoints of this arc (equiprobably) in time $D - 1/2$. Such a path will meet any continuously moving player with probability at least $1/2m$. Hence the independent repetition of such paths meets any continuously moving player in expected time no more than $2mD$.

For the second estimate double every arc so that the resulting network is Eulerian. Thus it takes no more time than $2m$ to complete a circuit of the original network. Consider the following mixed search strategy called *Eulerian or Wait* (EW): With probability p follow some Eulerian path on the doubled graph. With probability $(1-p)$ wait for a time period $2m$. If p is taken to be 1/2 (which is not optimal but gives the best simplified estimate for this proof), then in each time period of length $2m$, two players using this strategy will meet with probability at least 1/2, i.e., the probability that one waits and one searches. Hence the expected meeting time (neglecting meetings when both are searching) is no more than $4m$.

In the asymmetric case, the players can agree on who will search and who will wait. If some Eulerian circuit of the doubled network is traversed equiprobably in either direction by the player who searches, the players will meet before time $2m$ and on average before time m, giving the third of these elementary results. □

Since $D \geq 2$ (if $m > 3$ and there are no multiple arcs with equality for the network based on the complete graph), the estimate for the rendezvous value is less than that for the search value. The second estimate can be improved for the network based on the complete graph by optimizing for p [5], and the first can be extended to networks with arcs of varying lengths [3]. One obvious improvement to investigate is for the waiter to move to a nearby node and wait, and for the searcher to use a traveling salesman route on the nodes. All these estimates are very crude, but it seems likely that they are better than those achieved in general by random walks. Of course, if the network has no symmetries ($G = \{\text{identity}\}$) then the players can pick a point at which to rendezvous and the rendezvous times are much reduced as shown for the circle in the next section.

4. Rendezvous on the circle. To illustrate the symmetry notions of §2 and show how the general network estimates of §3 can be improved for specific cases, we now investigate rendezvous on a circle C. For convenience we take C to have circumference 2. We will consider three isometry groups on C: group G_1 consisting of just the identity, group G_2 consisting of all rotations, and group G_3 consisting of all isometries of C (rotations and reflections). We define the allowability of a search strategy (or strategy pair in the asymmetric case) to be the highest indexed group G_i with which it is consistent. In this section we obtain estimates on the rendezvous values $R^j(C,G_i)$, $j \in \{s,a\}, i \in \{1,2,3\}$ by analyzing some particular search strategies. But before we begin this analysis we have two observations. Since the search value of C is known to be 3/2 (using the COHATU strategy described below) [17], [1], this is an upper bound for all the rendezvous values. Also, by observing that C can be viewed as a network with $m = 2$, and the last two estimates of Proposition 1 can be reduced by a factor of 2 for Eulerian networks, we see that $R^s(C,G) \leq 4$ and $R^a(X,G) \leq 1$ for any isometry group G. We now consider four search strategies (the last two are strategy pairs for use in the asymmetric problem):

- GOTOZ: This strategy takes the shortest route from the initial point to a given rendezvous point Z. The allowability of GOTOZ is clearly G_1, since its implementation requires the exact knowledge of both the initial point and Z. As a strategy for symmetric rendezvous on (C, G_1) it has expected meeting time 2/3 for any (agreed) Z.

- COHATU: This strategy is short for *coin half tour*, and has been shown to be optimal for *both* the hider and the seeker in the circle search game. It moves from the initial point to the antipodal point (distance 1) at unit speed, equiprobably via either route, and then continues to oscillate between these points equiprobably via one of the routes. Its allowability is G_3 and results in expected meeting time of 3/2 if either player adopts it.
- OPDIR: This is the strategy pair in which one player goes clockwise at unit speed and the other goes counterclockwise at unit speed. Its allowability is G_2, since it doesn't require knowledge of the initial position but does require knowledge of up, which is not preserved under reflections. The expected meeting time is 1/2.
- GOSTAY: This strategy pair has one player wait while the other player chooses a random direction to go around the circle at unit speed. Its allowability is G_3 and it has an expected meeting time of 1.

The examples given here provide upper bounds (which may be exact) on the various rendezvous values for the circle, which are summarized in Table 1.

TABLE 1
Rendezvous bounds for the circle.

	G_1	G_2	G_3
R^s	$\frac{2}{3}$ (GOTOZ)	$\frac{3}{2}$ (COHATU)	$\frac{3}{2}$ (COHATU)
R^a	$\frac{1}{2}$ (OPDIR)	$\frac{1}{2}$ (OPDIR)	1 (GOSTAY)

5. Rendezvous on a cycle graph.

The problem of rendezvous may also be formulated on a graph. Suppose at time $t = 0$ the two players are placed independently and equiprobably on the nodes of a given graph. At each integer time t they may move to an adjacent node until the first time T that they occupy the same node. (If they transpose their positions on adjacent nodes they remain unaware of the other's position.) The problem for a complete graph has been analyzed in [5].

In this section we recast a search problem of Ruckle [14], [15] in a rendezvous context. Ruckle considered the problem faced by two players placed randomly on the nodes of the cycle graph C_m with m nodes labelled $i = 0, 1, \ldots, m-1$, where node i is adjacent to node j if $|i-j| = 1 \pmod m$. He restricted the players to symmetric Markovian strategies. That is, each player picks some number p in (0,1/2) and moves counterclockwise with probability p, clockwise with probability p, and remains still with probability $1-2p$. Ruckle considered the zero-sum game with payoff as expected meeting time $T(p,q)$, where the maximizing hider picks *move probability* p and the minimizing searcher picks *move probability* q. We use the same game dynamics but consider the simpler rendezvous problem of minimizing the expected meeting time $f_m(p) = T(p,p)$, when both players move with probability p. Our analysis of the dynamics is a simpler version of [14].

We consider a reduced state game where the difference between the two nodes (mod m) is the state. At each stage the state moves according to the independent randomizations of the players, except that the state 0 is an absorbing state. The m-state Markov chain which describes the dynamics is given for $m > 4$ by the matrix $A = \{a_{ij}\}$, where

$$a_{ij} = \begin{cases} 1 & \text{if } i = j = 0, \\ 0 & \text{if } i = 0, j \neq 0, \\ p^2 & \text{if } j = i \pm 2 \pmod m, \\ 2p(1-2p) & \text{if } j = i \pm 1 \pmod m, \\ (1-2p)^2 + 2p^2 & \text{if } i = j \neq 0, \\ 0 & \text{otherwise.} \end{cases} \tag{6}$$

If $m = 4$ there is a small modification such that

$$a_{ij} = 2p^2 \quad \text{when } j = i \pm 2 \, (\operatorname{mod} 4). \tag{7}$$

The expected meeting time is the same as the expected number of periods that the state is in $\{1, 2, \ldots, m-1\}$. Let B denote the $(m-1) \times (m-1)$ submatrix of A corresponding to these states. Then it is easy to see that the expected meeting time is given by

$$f_m(p) = \left(\frac{1}{m}\right) \sum_{k=0}^{\infty} e' B^k e = \left(\frac{1}{m}\right) e'(I-B)^{-1} e, \tag{8}$$

where e and e' are the column and row vectors consisting of $m-1$ ones. The full justification of a more complicated version of this formula is given in [14], [15], where it is also shown that f_m is convex. Now it is easy to see that the minimum of f_m cannot be at either endpoint 0 or 1. At 0 the players don't move, so they don't meet unless they start together. At $p = 1/2$, the distance between the players changes by ± 2 in each period. So if they started an odd distance apart they can never meet. So we now determine the optimal *rendezvous speed* $\hat{p}_m$ and the corresponding rendezvous value (a restricted version) $R_m = f_m(\hat{p}_m)$ for $m > 3$. For $m = 4, 5$ we can calculate an exact value, but for $m \geq 6$ we present only numerical approximations.

5.1. Rendezvous on C_4 or C_5. When m is 4 we evaluate (8) using (6) with (7) to obtain the expected meeting time function

$$f_4(p) = \frac{5 - 9p}{4(2p - 6p^2 + 4p^3)}. \tag{9}$$

This function has a local minimum (which is global between 0 and 1/2) at

$$\hat{p}_4 = \frac{1}{3}, \qquad \hat{R}_4 = f_4(\hat{p}_4) = \frac{27}{8} \doteq 3.375. \tag{10}$$

Similarly, the expected meeting time formula for $m = 5$ is

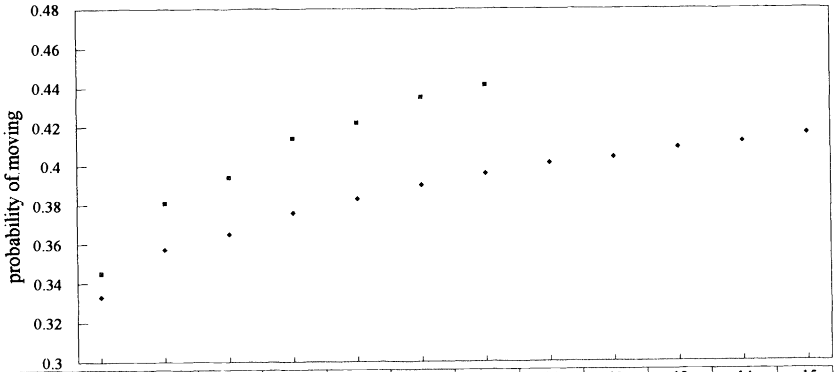
$$f_5(p) = \frac{4 - 6p}{4p - 10p^2 + 5p^3}, \tag{11}$$

which has a minimum at

$$\hat{p}_5 = \frac{2}{3}\left(1 - \frac{1}{\sqrt[3]{10}}\right), \qquad \hat{R}_5 = f_5(\hat{p}_5) \doteq 4.88. \tag{12}$$

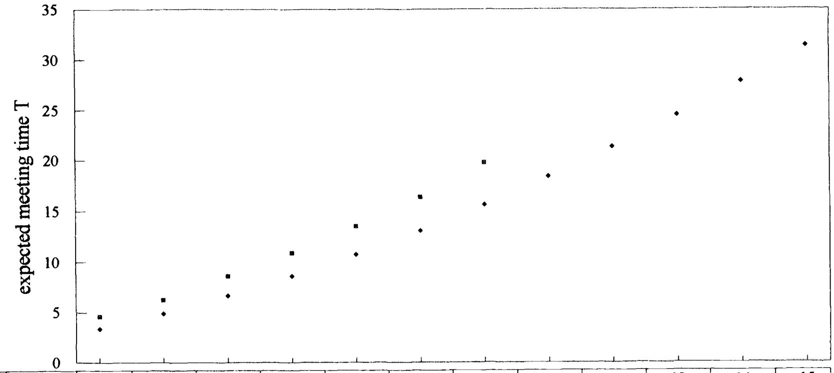
It seems likely that with some additional work a pattern for the rational functions f_m could be found, but we find it easier to proceed numerically for larger values of m.

5.2. Rendezvous vs. search game on C_m. Since Ruckle [14], [15] has given such a complete analysis of the search game on C_m, it is easy to compare our numerical results with his (which he gives for $m \leq 10$). In Fig. 1 we state and plot the optimal probabilities of moving $p = \hat{p}_m$ and $q = \hat{q}_m$ for our rendezvouser and Ruckle's searcher. Note that in each case the move probabilities are increasing in m (presumably to a limiting value of 1/2) and the probabilities for even m are slightly lower than the interpolated curve generated by odd values. There is a blip in the data at $m = 4$, which is not surprising given that the Markov chain has a different formula. The main observation is that it is necessary to move faster to catch an evader than to meet a rendezvouser, i.e., $q > p$. In Fig. 2 we present our numerical results for the



m=#(nodes)	4	5	6	7	8	9	10	11	12	13	14	15
▪ Search: q	0.345	0.381	0.394	0.414	0.422	0.435	0.441					
◆ Rendezvous: p	0.333	0.357	0.365	0.376	0.383	0.39	0.396	0.401	0.404	0.409	0.412	0.416

FIG. 1. *Optimal probability of moving.*



m = #(nodes)	4	5	6	7	8	9	10	11	12	13	14	15
▪ Search Val V	4.61	6.23	8.58	10.88	13.54	16.41	19.77					
◆ Rendezvous R	3.37	4.88	6.64	8.58	10.74	13.09	15.65	18.4	21.34	24.48	27.81	31.33

FIG. 2. *Rendezvous vs. search values.*

rendezvous value $R = R_m$ along with Ruckle's search values V_m for $m \leq 10$. Surprisingly, making the evader cooperate does not significantly reduce the expected meeting time, that is, R_m is not very much less than V_m.

The main reason for going through the analysis of this section is that it gives us, admittedly in a very restricted context, a direct comparison of the search game and rendezvous values, and of the respective optimal search strategies.

6. Rendezvous search on the line. Up to this point in the paper the search region X has been compact. We now consider the natural starting point for an investigation of noncompact regions, the real line. There has been much work by Beck and others [7], [8] on variations of the *linear search problem*, where a searcher starting at 0 and moving at unit speed attempts to find a stationary hider whose distribution is known. The problem posed in this section may therefore be called the *linear rendezvous search problem.*

The story behind this problem is as follows: Two friends have agreed to meet at noon on a certain street but have neglected to specify a specific point on the street. Assuming they know the distribution of their arrival points on the street (at noon), how should they move to meet in minimum expected time?

To formalize the problem, we first observe that everything is translation invariant, so we need to assume only a known cumulative distribution function F for the distance between them at time zero. Both players know F but neither knows the direction of the other player. So we assume that their mixed strategies are invariant with respect to reflection about their starting point. Let the rendezvous value $R^s(F)$ be the minimum expected meeting time achievable by players using the same reflection-invariant strategies. Both bounded and unbounded distributions will be of interest, but since we have nothing to say here about the latter, we will assume there is maximum initial distance $D = D_F = \min\{x : F(x) = 1\}$ and hence a finite mean initial distance $\mu = \mu_F$. Also, we may assume without loss of generality that the two players will not be placed at the same point or $F(0) = 0$.

One method the players may employ to keep a bound on their distance is to return periodically to their respective starting points. However, a less obvious and perhaps more effective way of achieving the same result is the following strategy, which I call 1F2B(x) for *one step forwards, two steps backwards*: pick either direction to call forward equiprobably; go a distance x forward at unit speed; then go a distance $2x$ backwards at unit speed; repeat the entire process indefinitely with independent randomization. A player following strategy 1F2B is globally following a random walk of step length x in time unit $3x$, so his position is certainly unbounded in time. However, the *relative* positions of *two* players following this strategy (until they meet) are maintained as described in the following proposition.

PROPOSITION 2. *Suppose two players are using strategy* $1F2B(x)$, *where* $x \geq D/2$, *and they have not met by time* $3x$. *Then the conditional distribution of their distance at time* $3x$ *is the same as their initial distance distribution* F. *In fact, their distance at time* $3x$ *is the same as their initial distance.*

Proof. It suffices to prove the last sentence. This is easily established by considering two cases: the players move in the same direction, or they move in opposite directions. In the latter case we either have $T \leq x$ (if they move toward each other) or $2x \leq T \leq 3x$ (if they move away from each other first). In the former case, the distance between them is preserved for all times $t \in [0, 3x]$, and they will not meet by time $3x$ (since we are assuming that $F(0) = 0$). Hence if the game has not ended by time $3x$ (that is, if $T > 3x$) then they must have gone in the same direction, and, therefore, their distance at time $3x$ equals their initial distance. □

PROPOSITION 3. *For any bounded distribution* F, $R^s(F) \leq 2D_F + \mu_F/2$.

Proof. We show that the bound is the expected meeting time $\hat{T}$ for two players using 1F2B$(D/2)$, where $D = D_F$ and $\mu = \mu_F$. First observe that $\hat{T}$ is certainly finite because the probability of meeting by time $3nD/2$ is at least 2^{-n}. To calculate $\hat{T}$ we consider separately the three cases where the players move in the same direction (probability 1/2), towards each other (probability 1/4), or away from each other (probability 1/4). In the first case the expected meeting time is $\hat{T} + 3D/2$ by the previous proposition. In the second case the cumulative distribution of meeting times is given by $G(t) = F(2t)$, so the expected meeting time is $\mu/2$. Finally, the third case reduces to the second case after time D, so the expected meeting time is $D + \mu/2$. Combining these observations, we obtain the equation

$$\hat{T} = \frac{1}{2}\left(\hat{T} + \frac{3D}{2}\right) + \frac{1}{4}\left(\frac{\mu}{2}\right) + \frac{1}{4}\left(D + \frac{\mu}{2}\right), \tag{13}$$

which has, as claimed, the unique solution

$$\hat{T} = 2D + \mu/2. \qquad \square \tag{14}$$

A particularly simple version of the rendezvous linear search problem is one in which the two players know the initial distance (say 1) between them, but neither knows the direction of the other. In this case $D = \mu = 1$, so the above estimate gives $R^s \leq 5/2$. We conjecture that the rendezvous value is in fact 5/2 for this problem and that 1F2B (1/2) is optimal. It is also likely to be optimal for the uniform distribution over [0,1], in which case that rendezvous value would be 9/4. If the initial distribution F is very sparse near D or unbounded, then of course 1F2B is not effective. However, it can still be used with finitely many repetitions with the step length successively increased as the conditional distribution changes.

The work done in this section is of course very preliminary. However, we are sure the linear rendezvous search problem defined here will prove to be a very productive sidekick to its cousin, the linear search problem.

7. Questions for further work. While we feel that the formalization of the rendezvous search problem given here is important, we are aware of the very preliminary and exploratory nature of the results. Even in cases (such as COHATU on the circle or 1F2B on the line) where we think our suggested search strategies are effective, we have not shown them to be optimal. In other cases (such as search on a network) even the suggested strategies are not very good and serve only to give preliminary upper bounds on the generic rendezvous value. However, we hope these first attacks on the problem of rendezvous search will stimulate further work in this area.

We conclude the paper with a number of questions on rendezvous search that would extend the scope of the theory.

- Many players: For simplicity, we have limited the discussion in this paper to the case of two players wishing to rendezvous. In general, when two players meet (out of $n > 2$) they may exchange information about where they have been (which is their only private information). There are two cases to consider: they may be restricted to stay together once they have met, or they may be allowed to agree on a joint strategy which may allow separate motion. The former is probably easier to analyze.
- Local vision: The formulation given in this paper assumes the players can see the whole search region. An alternative is to assume they know the search region but can perceive it only locally (within a given distance). For example, our formulation, for the unit interval $X = [0, 1]$, allows the strategy of *go to* 1/2 even if the isometry of reflecting about 1/2 is given. With a local vision restriction the players would only know where 1/2 was when they were close to 0 or 1, when they would know the whole interval (possibly up to reflection).
- Rendezvous in a maze: We may further limit the players' information by giving them local vision and not telling them the nature of the search region X. Perhaps they are only told, for example, that they will land in a network of total length 1. This is a two-sided version of maze search problems.
- Adversary-rendezvous games: Suppose the two searchers' initial position in X is chosen by another player who wishes to maximize their meeting time. This other may even choose X, possibly in a maze or local vision scenario. This becomes a two-person zero-sum game where the rendezvousers are a single player.
- Lower bounds on R: Upper bounds on the rendezvous value may be obtained by guessing good search strategies. How can lower bounds be obtained? (Of course half the expected initial distance is a lower bound, but hardly a good one.)
- Extremal theory: For regions which can be searched in, say, unit time, which have the smallest and largest rendezvous values?
- A calculus of rendezvous: Suppose $2X$ denotes the search region consisting of two copies of X, each with the same symmetries as the original. Suppose a player may move between (perhaps corresponding) points in the two copies in a given time L, but that two players so moving will not meet. Furthermore, suppose the original distribution is split equiprobably between the two copies. This defines a search game on the region $2X$. Similarly, $X + Y$ could be defined with allowed transport between any pair of points in the two regions. This seems to

be a good model of a couple trying to meet who have been invited to two large parties, where the only transport between them is via taxi. The transport time L is probably significant. We can either assume the taxi is free or add an additional cost other than time.

REFERENCES

[1] S. Alpern, *The search game with mobile hider on the circle*, in Differential Games and Control Theory, E. O. Roxin et al., eds., Marcel Dekker, New York, 1974, pp. 181–200.

[2] S. Alpern and M. Asic, *Ambush strategies in search games on graphs*, SIAM J. Control Optim., 24 (1986), pp. 66–75.

[3] ———, *The search value of a network*, Networks, 15 (1985), pp. 229–238.

[4] E. J. Anderson and M. Aramendia, *A linear programming approach to the search game on a network with mobile hider*, SIAM J. Control Optim., 30 (1992), pp. 675–694.

[5] E. J. Anderson and R. R. Weber, *The rendezvous problem on discrete locations*, J. Appl. Probab., 28 (1990), pp. 839–851.

[6] V. J. Baston and F. A. Bostock, *A continuous game of ambush*, Naval Res. Logis., 34 (1987), pp. 645–654.

[7] A. Beck, *On the linear search problem*, Israel J. Math., 2 (1964), pp. 221–228.

[8] A. Beck and D. J. Newman, *Yet more on the linear search problem*, Israel J. Math., 8 (1970), pp. 419–429.

[9] S. Gal, *Search Games*, Academic Press, New York, 1980.

[10] ———, *Search games with mobile and immobile hider*, SIAM J. Control Optim., 17 (1979), pp. 99–122.

[11] A. Y. Garnaey, *Search game in a rectangle*, J. Optim. Theory Appl., 69 (1991), pp. 531–542.

[12] R. Isaacs, *Differential Games*, John Wiley, New York, 1965.

[13] S. P. Lalley and H. E. Robbins, *Stochastic search in a convex region*, Probab. Theory Related Fields, 77 (1988), pp. 99–116.

[14] W. H. Ruckle, *Pursuit on a cyclic graph*, Internat. J. Game Theory, 10 (1983), 91–99.

[15] ———, *Geometric Games and their Applications*, Pitman, Boston, 1983.

[16] T. C. Schelling, *The Strategy of Conflict*, Harvard University Press, Cambridge, 1960.

[17] M. I. Zelikin, *On a differential game with incomplete information*, Soviet Math. Dokl., 13 (1971), pp. 228–231.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 33, No. 3, pp. 684–701, May 1995

© 1995 Society for Industrial and Applied Mathematics
002

H_∞ BOUNDARY CONTROL WITH STATE FEEDBACK: THE HYPERBOLIC CASE*

VIOREL BARBU†

Abstract. Necessary and sufficient conditions of existence for the H_∞ control problem associated with infinite dimensional boundary control systems of hyperbolic type are given.

Key words. boundary control systems, feedback controller, differential game, exponentially stable semigroup

AMS subject classifications. 93B50, 93C35, 49A40

1. Introduction and problem function. Let X, U, W, Z be real Hilbert spaces and A be the infinitesimal generator of C_0-semigroup e^{At} on X, $B_1 \in L(W,Z)$, $B_2 \in L(U,(D(A^*))')$, $C_1 \in L(X,Z)$, $D_{12} \in L(U,Z)$. Here A^* is the adjoint of A, $D(A^*)$ is the domain of A^* endowed with the graph norm, and $(D(A^*))'$ is the dual space to $D(A^*)$. Consider the input-output system defined by

$$x'(t) = Ax(t) + B_1w(t) + B_2u(t), \qquad t \in R^+ = [0,\infty[, \tag{1.1}$$

$$z(t) = C_1x(t) + D_{12}u(t), \qquad t \in R^+. \tag{1.2}$$

Here $x(t) \in X$ is the state of the system, $u(t) \in U$ is the control input, $w(t) \in W$ is an exogeneous input, and $z(t) \in Z$ is the controlled output.

We shall assume

(i) $A^{-1}B_2 \in L(U,X)$, and for every $T > 0$ the operator $B_2^*e^{A^*t}$ admits a continuous extension from X to $L^2(0,T;U)$, i.e.,

$$\int_0^T |B_2^*e^{A^*t}x|_U^2 \le C_T|x|^2 \quad \forall x \in X. \tag{1.3}$$

We denoted the adjoint of $B_2 \in L(U,(D(A^*))')$ by $B_2^* \in L(D(A^*),U)$. We shall denote the norms of X, Z, U, W by $|\cdot|_X$, $|\cdot|_Z$, $|\cdot|_U$, $|\cdot|_W$ and the corresponding scalar products by $(\cdot,\cdot)$, $(\cdot,\cdot)_Z$, $(\cdot,\cdot)_U$, $(\cdot,\cdot)_W$. By assumption (i) it follows that for every $T > 0$, system (1.1) with initial condition $x(0) = x_0 \in X$, and inputs $u \in L^2(0,T;U)$, $w \in L^2(0,T;W)$ has a mild solution $x \in C([0,T];X)$ given by

$$x(t) = e^{At}x_0 + \int_0^t e^{A(t-s)}(B_1w(s) + B_2u(s))ds \quad \forall t \in [0,T]. \tag{1.4}$$

More precisely, for every $T > 0$ the operator $u \to \int_0^t e^{A(t-s)}B_2u(s)ds$ is continuous from $L^2(0,T;U)$ to $C([0,T];X)$ (see, e.g., [5]).

This is the abstract formulation of a large class of boundary control systems with distributed disturbance input including the one governed by the wave equation with Dirichlet and Neumann boundary control, first-order hyperbolic systems, and Euler–Bernoulli equations (see [5], [8]). For instance, the input-output system

$$\begin{aligned} y_{tt} - \Delta y = w \quad &\text{in}\,\Omega \times R^+, \\ y = u \quad &\text{in}\,\partial\Omega \times R^+, \end{aligned} \tag{1.5}$$

[*] Received by the editors October 14, 1992; accepted for publication (in revised form) December 3, 1993.

[†] Department of Mathematics, University of Iaşi, 6600 Iaşi, Romania.

where $u \in L^2(R^+; L^2(\partial\Omega))$, $w \in L^2(R^+; L^2(\Omega))$ can be written in the form (1.1) when $X = L^2(\Omega) \times H^{-1}(\Omega)$, $U = L^2(\partial\Omega)$, $W = L^2(\Omega)$, and

$$A = \left\| \begin{matrix} 0 & I \\ \Delta & 0 \end{matrix} \right\|, \quad B_1 w = \left\| \begin{matrix} 0 \\ w \end{matrix} \right\|, \quad B_2 u = \left\| \begin{matrix} 0 \\ -\Delta D u \end{matrix} \right\|.$$

Here Δ is the Laplace operator with the domain $H_0^1(\Omega) \cap H^2(\Omega)$ (Ω is an open-bounded subset of R^n with a smooth boundary $\partial\Omega$) and $D \in L(L^2(\partial\Omega),\ L^2(\Omega))$ is the Dirichlet map, i.e., $\Delta D u = D$ in Ω, $Du = u$ in $\partial\Omega$ (see, e.g., [5]).

We shall denote by $\mathcal{F}$ the class of feedback controllers $F \in L(D(A), U)$ having the property that $A_F = A + B_2 F$ generates an exponentially stable C_0-semigroup $e^{A_F t}$ on X and $D(A_F) \subset D(F)$; $Fe^{A_F t} \in L(X, L^2(R^+; U))$. We note that the operator $A + B_2 F$ is continuous from $D(A)$ to $(D(A^*))'$, so A_F, which is its restriction to H, is well defined. Since $B_1 \in L(W, Z)$, we see that for $F \in \mathcal{F}$ the closed loop operator $S_F : L^2(R^+; W) \to L^2(R^+; Z)$, $(S_F w)(t) = (C_1 + D_{12} F) \int_0^t e^{A_F(t-s)} B_1 w(s) ds$; $t \geq 0$, is well defined.

As a matter of fact $x = Sw$ is the mild solution of (1.1) with feedback control $u = Fx$ and initial value condition $x(0) = 0$ (see [7] for a definition of mild solutions of an infinite dimensional Cauchy problem).

Let $\gamma > 0$. Following the standard H_∞ control theory (see [2]) we say that the feedback controller $F : X \to U$ is an H_∞ controller for system (1.1), (1.2) if $F \in \mathcal{F}$ and $\|S_F\| < \gamma$. Here $\|S_F\|$ is the norm of the operator $S_F \in L(L^2(R^+; W), L^2(R^+; Z))$.

In addition to assumption (i), we shall assume that

(ii) The pair (A, C_1) is exponentially detectable, i.e., there exists $K \in L(Z, X)$ such that $A + KC_1$ generates an exponentially stable semigroup;

(iii) $D_{12}^* |C_1, D_{12}| = |0, I|$.

Assumption (iii) simply says that

$$|C_1 x + D_{12} u|_Z^2 = |C_1 x|_Z^2 + |u|_U^2 \quad \forall (x, u) \in X \times U. \tag{1.6}$$

Theorem 1 below represents the main result of this paper.

THEOREM 1. *Let $\gamma > 0$ and suppose that assumptions* (i), (ii), *and* (iii) *hold. Then there exists an H_∞ controller $F \in \mathcal{F}$ such that $\|S_F\| < \gamma$ if and only if there exists $P \in L(X, X)$, $P = P^* \geq 0$ such that $A_P = A - B_2 B_2^* + \gamma^{-2} B_1 B_1^* P$ is the infinitesimal generator of an exponentially stable C_0-semigroup and*

$$B_2^* P \in L(D(A), U) \cap L(D(A_P), U), \qquad A^* P \in L(D(A_P), U), \tag{1.7}$$

$$(Ax, Py) + (Px, Ay) - (P(B_2 B_2^* - \gamma^{-2} B_1 B_1^*) Px, y) + (C_1 x, C_1 y) = 0 \tag{1.8}$$

for all $x, y \in D(A)$ or else for all $x, y \in D(A_P)$. Moreover, in this case the state feedback $F = -B_2^* P$ is an H_∞ control, $\|S_F\| < \gamma$, and the solution P of Riccati equation (1.8) *is unique in the class of $P \in L(X, X^*)$, $P = P^* \geq 0$, having the property that A_P is exponentially stable.*

We note that by (1.7) the operator $x \to Ax - B_2 B_2^* Px + \gamma^{-2} B_1 B_1^* P$ is continuous from $D(A)$ to $(D(A^*))'$. A_P is the restriction of this operator to H.

Theorem 1 resembles the standard finite dimensional results [2]. The case $B_2 \in L(U, X)$, i.e., the case of distributed input controllers, was previously studied in [9].

A standard approach to the suboptimal H_∞ controllers is to associate with system (1.1), (1.2) the differential game

$$\sup_{w \in L^2(R^+; W)} \inf_{u \in L^2(R^+; U)} \int_0^\infty (|z(t)|_Z^2 - \gamma^2 |w(t)|_W^2) dt \tag{1.9}$$

and write the closed loop strategies in terms of an algebraic Riccati equation [2], [9]. However, the extension of [9] to the present situation is not trivial and relies on some recent results [3] on synthesis of quadratic optimal control problems with infinite horizon. Related results were recently obtained for the differential game (1.9) by McMillan and Triggiani [6] under the additional hypothesis that either e^{At} is exponentially stable or the system $x' = A^*x + (C_1^*C_1)^*v$ is exactly controllable in finite time.

2. The sup–inf problem. Here we shall study the differential game (1.9) under assumptions (i), (ii), (iii), and

(j) the pair (A, B_2) is exponentially stabilizable, i.e., $\exists F \in L(D(A), U)$ such that $A + B_2F$ generates an exponentially stable semigroup $e^{(A+B_2F)t}$ and the following inequality holds:

$$\sup_{w\in L^2(R^+;W)} \inf\left\{\int_0^\infty (|C_1x|_Z^2 + |u|_U^2)dt;\ x' = Ax + B_2u + B_1w;\right.$$
$$\left. x(0) = 0,\ u \in L^2(R^+;U)\right\}\left(\int_0^\infty |w|_W^2\,dt\right)^{-1} < \gamma^2.$$

In particular, assumption (j) holds if there is an H_∞ control $F \in \mathcal{F}$ for system (1.1).

LEMMA 1. *If assumption* (j) *holds then*

(2.1)
$$\inf\left\{\int_0^\infty (|C_1y|_Z^2 + |u|_U^2)dt; y = Ay + B_2u + B_1w; y(0) = x_0,\ u \in L^2(R^+;U)\right\}$$
$$< (\gamma^2 - \varepsilon)\int_0^\infty |w|_W^2\,dt + C|x_0|^2 \quad \forall x_0 \in X \quad \forall w \in L^2(R^+;W)$$

for some $\varepsilon > 0$ *and* $C \in R$ *independent of* w.

Proof. If condition (j) holds, then $\exists\,\varepsilon > 0$ such that

$$\inf\left\{\int_0^\infty (|C_1x|_Z^2 + |u|_U^2)dt;\ x' = Ax + B_2u + B_1w, x(0) = 0,\ u \in L^2(R^+;U)\right\}$$
$$< (\gamma^2 - \varepsilon)\int_0^\infty |w|_W^2\,dt \quad \forall w \in L^2(R^+;W).$$

Under assumptions (i), (ii), (j), $\exists P_0 \in L(X,X)$, $P_0 = P_0^* \geq 0$ such that $A_{P_0} = A - B_2B_2^*P_0$ is exponentially stable and $B_2^*P_0e^{A_{P_0}t}x_0 \in L^2(R^+;U)$, $\forall x_0 \in X$ (see [3]). This implies that

$$\inf\left\{\int_0^\infty (|C_1x|_Z^2 + |u + B_2^*P_0e^{A_{P_0}t}x_0|_U^2)dt;\right.$$
$$\left. u \in L^2(R^+;U),\ x' = Ax + B_2u + B_1w,\ x(0) = 0\right\}$$
$$< (\gamma^2 - 2^{-1}\varepsilon)\int_0^\infty |w|_W^2\,dt + C|x_0|^2 \quad \forall w \in L^2(R^+;W),\ x_0 \in X.$$

Hence

$$\inf\left\{\int_0^\infty (|C_1x|_Z^2 + |u|_U^2)dt;\ x' = Ax + B_2(u - B_2^*P_0e^{A_{P_0}t}x_0) + B_1w;\right.$$
$$\left. x(0) = 0,\ u \in L^2(R^+;U)\right\} < (\gamma^2 - 2^{-1}\varepsilon)\int_0^\infty |w|_W^2\,dt + C|x_0|^2.$$

Since the solution y of (1.1) with the initial value condition $y(0) = x_0$ can be represented as $y(t) = x(t) + e^{A_{P_0}t}x_0$, where $x' = Ax + B_2(u - B_2^* P_0 e^{A_{P_0}t}x_0) + B_1 w$, $x(0) = 0$, the latter inequality implies (2.1) as desired.

Let $K_{x_0} : L^2(R^+; U) \times L^2(R^+; W) \to [-\infty, +\infty]$ be defined by

$$K_{x_0}(u,w) = \int_0^\infty (|C_1 x + D_{12}u|_Z^2 - \gamma^2 |w|_W^2)dt = \int_0^\infty (|C_1 x|_Z^2 + |u|_U^2 - \gamma^2|w|_W^2)dt,$$

where x is the mild solution of (1.1) with $x(0) = x_0$.

In terms of K_{x_0} we may equivalently write (1.9) as

$$\sup_{w \in \mathcal{W}} \inf_{u \in \mathcal{U}} K_{x_0}(u,w), \tag{2.2}$$

where $\mathcal{U} = L^2(R^+; U)$ and $\mathcal{W} = L^2(R^+; W)$. In this section we shall prove the following result.

PROPOSITION 1. *Problem* (2.2) *has a unique solution* (u^*, w^*) *characterized by*

$$u^*(t) = B_2^* p^*(t) \quad a.e.\, t > 0, \tag{2.3}$$

$$w^*(t) = -\gamma^{-2} B_1^* p^*(t) \quad a.e.\, t > 0, \tag{2.4}$$

where $p^* \in C(R^+; X)$ *is the solution of*

$$\begin{gathered}(p^*)' = -A^* p^* + C_1^* C_1 x^* \quad in\, R^+, \\ p^*(\infty) = 0,\end{gathered} \tag{2.5}$$

and x^* *is the solution of* (1.1) *with* $x^*(0) = x_0$, $u = u^*$, *and* $w = w^*$.

Equation (2.5) should be understood, of course, in the following mild sense:

$$p(t) = e^{A^*(T-t)}p(T) - \int_t^T e^{A^*(s-t)} C_1^* C_1 x^*(s) ds \tag{2.6}$$

for all $0 \le t \le T < \infty$.

We shall first study the minimization problem

$$\inf\{K_{x_0}(u,v);\ u \in \mathcal{U}\}, \tag{2.7}$$

where $w \in \mathcal{W}$ is arbitrary but fixed.

This is a linear quadratic optimal control problem on $R^+ = [0, \infty[$.

LEMMA 2. *For each* $w \in \mathcal{W}$, *problem* (2.7) *has a unique solution* $\bar{u} = \Gamma_{x_0} w$.

Proof. By assumption (j) the function $u \to K_{x_0}(u,w)$ is not identically $+\infty$ for each $w \in \mathcal{W}$.

Moreover, it is strictly convex and continuous, and

$$K_{x_0}(u,w) \ge \|u\|_U^2 + C \quad \forall\, u \in \mathcal{U}. \tag{2.8}$$

This implies that (2.7) has a unique solution $\bar{u}$ by the standard existence result.

LEMMA 3. *The function* $\bar{u}$ *is the solution of* (2.7) *if and only if there exists* $p \in C(R^+; X)$ *such that*

$$p' = -A^* p + C_1^* C_1 \bar{x} \quad in\, R^+;\ p(\infty) = 0, \tag{2.9}$$

$$B_2^* p(t) = \bar{u}(t) \quad a.e.\, t > 0. \tag{2.10}$$

Here $\bar{x}$ *is the solution of* (1.1) *with* $u = \bar{u} = \Gamma_{x_0}\bar{w}$.

Proof of Lemma 2. Let $\bar{u}$ be the solution of (2.7). Consider the family of approximating control problems

$$\varphi_n(x_0) = \inf\left\{\frac{1}{2}\int_0^\infty (|C_1x|_Z^2 + |u|_U^2)dt; x' = Ax + B_2u + B_1w;\ x(0) = x_0\right\}, \tag{2.11}$$

and denote the corresponding solution by (x_n, u_n). We have

$$\varphi_n(x_0) \leq \inf\left\{\frac{1}{2}\int_0^\infty (|C_1x|_Z^2 + |u|_U^2)dt;\ x' = Ax + B_2u + B_1w;\ x(0) = x_0\right\},$$

and since the function $(x, u) \to \int_0^\infty (|C_1x|_Z^2 + |u|_U^2)dt$ is weakly lower semicontinuous we get

$$\begin{aligned} u_n &\to \bar{u} \quad \text{strongly in } L^2(R^+; U), \\ x_n(t) &\to \bar{x} \quad \text{uniformly on compacta}, \\ C_1x_n &\to C_1\bar{x} \quad \text{strongly in } L^2(R^+; Z). \end{aligned}$$

On the other hand if we write the state system (1.1) as

$$x' = (A + KC_1)x - KC_1x + B_2u + B_1w, \tag{2.12}$$

where $K \in L(Z, X)$ is as in assumption (ii), we have

$$x_n(t) - \bar{x}(t) = \int_0^t e^{A_K(t-s)}((B_2(u_n(s) - \bar{u}(s)) - K(C_1x_n(s) - C_1\bar{x}(s))ds.$$

Next, by Lemma 5.2 in [3] we have

$$\int_0^\infty |B_2^* e^{A_K^* t}x_0|_U^2\, dt \leq C|x_0|^2 \quad \forall\, x_0 \in X.$$

This yields

$$x_n(t) \to \bar{x}(t) \quad \text{uniformly on } R^+.$$

Moreover, since $\bar{u} \in L^2(R^+; U)$ and $w \in L^2(R^+; W)$, by the same argument it follows that $\bar{x} \in L^2(R^+; X)$ and

$$\lim_{t\to\infty} \bar{x}(t) = 0 \quad \text{strongly in } X.$$

Moreover, according to a standard result in the theory of linear quadratic control problems $\exists p_n \in C([0, n]; X)$ such that

$$\begin{aligned} p_n' &= -A^*p_n + C_1^*C_1x_n \quad \text{in } [0, n], \\ p_n(n) &= 0, \end{aligned} \tag{2.13}$$

$$u_n(t) = B_2^*p_n(t) \quad \text{a.e. in } (0, n). \tag{2.14}$$

Equation (2.13) should be viewed in the mild sense, i.e.,

$$\text{(2.15)}\quad p_n(t) = e^{A^*(T-t)}p_n(T) - \int_t^T e^{A^*(\tau-t)}C_1^*C_1x_n(\tau)d\tau, \qquad 0 \le t \le T \le n.$$

We set

(2.16)

$$\varphi_n(t, x_0) = \inf\left\{\frac{1}{2}\int_t^n (|C_1x|_Z^2 + |u|_U^2)ds;\ x' = Ax + B_2u + B_1w \,\text{in}\, (t,n);\ x(t) = x_0\right\}.$$

The function $\varphi_n(t,\cdot)$ is convex (in fact it is quadratic) and differentiable. Moreover, we have

$$\text{(2.17)}\qquad p_n(t) = -\nabla\varphi_n(t, x_n(t)) \quad \forall\, t \in [0,n].$$

where ∇ stands for Fréchet differential. Indeed since (x_n, u_n) is optimal in (2.11), where $x_0 = x_n(t)$, we have

$$\begin{aligned}
&\varphi_n(t, x_n(t)) - \varphi_n(t, y_0)\\
&\quad = \frac{1}{2}\int_t^n (|C_1x_n|_Z^2 - |C_1y|_Z^2)d\tau + \frac{1}{2}\int_t^n (|u_n|_U^2 - |u|_U^2)d\tau\\
&\quad \le \int_t^n ((C_1x_n, C_1(x_n - y)) + (u_n, u_n - u)_U)d\tau,
\end{aligned}$$

where $y' = Ay + B_2u + B_1w;\ y(t) = y_0$.

Then, by (2.13), (2.14), and Lemma 7 in the appendix, we see that

$$\text{(2.18)}\qquad \varphi_n(t, x_n(t)) - \varphi_n(t, y_0) \le -(p_n(t), x_n(t) - y_0) \quad \forall\, y_0 \in X$$

as claimed.

If in (2.18) we take $y_0 = x_n(t) - \rho\theta$, where $|\theta| = 1$, we get

$$\text{(2.19)}\qquad |p_n(t)| \le |x_n(t)||p_n(t)| + \varphi_n(t, \rho\theta)$$

for all $t \in [0,n]$, $|\theta| = 1$, and $\rho < 0$. On the other hand, we have

$$\begin{aligned}
\varphi_n(t, \rho\theta) &\le \inf\left\{\frac{1}{2}\int_t^\infty (|C_1x|_Z^2 + |u|_U^2)ds;, x' = Ax + B_2u + B_1w;\ x(t) = \rho\theta\right\}\\
&\le \frac{1}{2}\int_t^\infty (|C_1\tilde{x}|_Z^2 + |B_2^*P_0\tilde{x}|_U^2)dt,
\end{aligned}$$

where $\tilde{x}' = (A - B_2B_2^*P_0)\tilde{x} + B_1w,\ \tilde{x}(t) = \rho\theta$, and $P_0 \in L(X, X)$ is such that $A_{P_0} = A - B_2B_2^*P_0$ generates an exponentially stable semigroup with $B_2^*P_0e^{A_{P_0}t}x_0 \in L^2(R^+; U)$, $\forall\, x_0 \in X$ (see [3]). Here $\tilde{x}$ is given by $\tilde{x}(s) = e^{A_{P_0}(s-t)}(\rho\theta) + \int_t^s e^{A_{P_0}(s-\tau)}B_1w(\tau)d\tau$, $s \ge t$, and, therefore,

$$\begin{aligned}
&|C_1\tilde{x}(s)|_Z^2 + |B_2^*P_0\tilde{x}(s)|_U^2\\
&\quad \le C\Bigg(e^{-\alpha(s-t)}\rho^2 + \left(\int_t^s e^{-\alpha(s-\tau)}|B_1w(\tau)|d\tau\right)^2\\
&\qquad + \left(\int_t^s \|B_2^*P_0e^{A_{P_0}(s-\tau)}\||B_1w(\tau)|d\tau\right)^2\Bigg),\\
&\qquad s \ge t \ge 0, \qquad \rho > 0,
\end{aligned}$$

for some $\alpha > 0$. Hence, by the Young formula we have

$$\varphi_n(t, \rho\theta) \leq C\left(\rho^2 + \int_t^\infty |w|_W^2\, d\tau\right) \quad \forall t, \rho > 0, \tag{2.20}$$

where C is independent of t and ρ.

In particular, we conclude by (2.19) and (2.20) that

$$|p_n(t)| \leq C(\rho - |x_n(t)|)^{-1}\left(\rho^2 + \int_t^\infty |w|_W^2\, d\tau\right) \quad \forall t \in [0, n]. \tag{2.21}$$

Since $x_n(t) \to \tilde{x}(t)$ uniformly in R^+ we infer that

$$\limsup_{n\to\infty} |p_n(t)| \leq C(\rho - |\tilde{x}(t)|)^{-1}\left(\rho^2 + \int_t^\infty |w|_W^2\, d\tau\right).$$

In particular, it follows that $\{p_n\}$ is bounded in $L^\infty(R^+; X)$, so on a subsequence $p_n \to p$ weak star in $L^\infty(R^+; X)$. By (2.13) we have

$$\begin{aligned}&\int_0^T (p_n(t), \psi(t))dt\\ &\quad= \int_0^T (p_n(T), e^{A(T-t)}\psi(t))dt - \int_0^T \left(\psi(t), \int_t^T e^{A^*(\tau-t)}C_1^*C_1x_n(\tau)d\tau\right)dt\\ &\quad\forall \psi \in C^1(0, T; X).\end{aligned}$$

Letting $n \to \infty$ we get

$$\begin{aligned}&\int_0^T (p(t), \psi(t))dt\\ &\quad= \left(p_T, \int_0^T e^{A(T-t)}\psi(t)dt\right) - \int_0^T \left(\psi(t), \int_t^T e^{A^*(\tau-t)}C_1^*C_1x(\tau)d\tau\right)dt\\ &\quad\forall \psi \in C^1(0, T; X),\end{aligned}$$

where $p_T = w - \lim_{n\to\infty} p_n(T)$ on a subsequence. Hence

$$p(t) = e^{A^*(T-t)}p_T - \int_t^T e^{A^*(\tau-t)}C_1^*C_1x(\tau)d\tau \quad \text{a.e. } t \in (0, T).$$

Extend p by continuity on $[0, T]$ so we get $p \in C(R^+; X)$, which satisfies (2.9), i.e.,

$$p(t) = e^{A^*(T-t)}p(T) - \int_t^T e^{A^*(\tau-t)}C_1^*C_1x(\tau)d\tau \quad \text{for all } 0 \leq t \leq T < \infty.$$

By (2.21) we have

$$|p(t)| \leq \lim_{n\to\infty} |p_n(t)| \leq C(\rho - |\bar{x}(t)|)^{-1}\left(\rho^2 + \int_t^\infty |w|_W^2\, d\tau\right) \quad \forall t \geq 0.$$

Since, as seen earlier, $\lim_{t\to\infty} \bar{x}(t) = 0$,

$$|p(t)| \leq C\left(\rho + \rho^{-1}\int_t^\infty |w(\tau)|_W^2\, d\tau\right) \quad \text{for } t \geq T(\rho).$$

For $\rho = (\int_t^\infty |w(\tau)|_W^2\, d\tau)^{1/2}$ we have, therefore, that

$$\lim_{t\to\infty} |p(t)| = 0.$$

On the other hand, by (2.14) we have

$$\begin{aligned} u_n(t) =& B_2^* p_n(t) = B_2^* e^{A^*(T-t)} p_n(T) - \int_t^T B_2^* e^{A^*(\tau-t)} C_1^* C_1 x_n \, d\tau \\ &\text{for } t \in [0,T] \subset [0,n]. \end{aligned}$$

Since the map $q \to B_2^* e^{A^*(T-t)} q$ is continuous from X to $L^2(0,T;U)$, and the map $v \to \int_t^T B_2^* e^{A^*(\tau-t)} v(\tau) d\tau$ is continuous from $L^1(0,T;X)$ to $L^2(0,T;U)$ (see [3]), we may pass to limit in the previous equation to get that

$$u^*(t) = B_2^* p(t) \quad \text{a.e. } t > 0$$

as claimed.

To prove that system (2.9), (2.10) is sufficient for the optimality of $\bar{u}$ in problem (2.7), consider a solution $(\bar{x}, \bar{u}, p)$ of (1.1), (2.9), (2.10) with $\bar{x}(0) = x_0$. We have

$$(2.22) \qquad |C_1 \bar{x}|_Z^2 \le |C_1 x|_Z^2 + 2(C_1^* C_1 \bar{x}, \bar{x} - x) \quad \forall x \in X.$$

Now consider an arbitrary solution (x,u) of (1.1); it follows by Lemma 7 that

$$(2.23) \qquad \begin{aligned} &\int_0^T (C_1^* C_1 \bar{x}, \bar{x} - x) dt \\ &\quad = (p(T), \bar{x}(T) - x(T)) - \int_0^T (\bar{u}, \bar{u} - u)_U \, dt \quad \forall T > 0. \end{aligned}$$

Now if we write system (1.1) in the form (2.12) and use assumption (ii), it follows as above that $\bar{x}$, $x \in L^2(R^+;X)$ for any solution x of (1.1) having the property that $Cx \in L^2(R^+;Z)$.

Since $\lim_{T\to\infty} p(T) = 0$ by (2.22) and (2.23), we see that $\bar{u}$ is optimal in problem (2.7) as claimed.

LEMMA 4. *The solution* $\bar{u} = \Gamma_{x_0}$ *of problem* (2.7) *can be represented as*

$$(2.24) \qquad \Gamma_{x_0} w = \Gamma_0 w + f_{x_0} \quad \forall w \in W,$$

where $\Gamma_0 \in L(\mathcal{W}, \mathcal{U})$ *and* $f_{x_0} = \arg\inf\{K_{x_0}(u,0);\ u \in \mathcal{U}\}$.

Proof. Let

$$\Gamma_0 w = \arg\inf\{K_0(u,w);\ u \in \mathcal{U}\}.$$

By Lemma 3 it follows that $\Gamma_0 w + f_{x_0} = B_2^* p$, where p is the solution of (2.9); this implies (2.24). According to the same lemma, Γ_0 is linear from $\mathcal{W}$ to $\mathcal{U}$. It is also readily seen that Γ_0 is bounded.

Proof of Proposition 1. Consider the function $\psi : \mathcal{W} \to R$ defined by

$$\psi(w) = -K_{x_0}(\Gamma_{x_0} w, w), \qquad w \in \mathcal{W}.$$

By Lemma 4, we may write ψ as

$$\psi(w) = \|Dw\|_W^2 + (Dw, f) + \alpha \quad \forall w \in \mathcal{W},$$

where $D \in L(\mathcal{W}, \mathcal{W})$, $f \in \mathcal{W}$, and $\alpha \in R$. On the other hand, by assumption (j) it follows that

$$\psi(w) \geq \omega \|w\|_{\mathcal{W}}^2 + \beta \quad \forall w \in \mathcal{W},$$

where $\omega > 0$ and $\alpha \in R$.

This implies that D^*D is positive definite, so ψ attains its infimum on $\mathcal{W} = L^2(R^+; W)$ in a unique point w^*. Clearly $(u^* = \Gamma_{x_0} w^*, w^*)$ is the unique solution to sup–inf problem (1.9). To conclude the proof it remains to show that

$$w^*(t) = -\gamma^{-2} B_1^* p^*(t) \quad \text{a.e. } t > 0. \tag{2.25}$$

We have

$$\begin{aligned}&\int_0^\infty (\gamma^2 |w^*|_W^2 - |\Gamma_0 w^* + f_{x_0}|_U^2 - |C_1 x^*|_Z^2) dt \\ &\quad \leq \int_0^\infty (\gamma^2 |w^* + \lambda w|_W^2 - |\Gamma_0(w^* + \lambda w) + f_{x_0}|_U^2 - |C_1(x^* + \lambda \tilde{x})|_Z^2) dt \\ &\quad \forall w \in \mathcal{W},\ \lambda \in R,\end{aligned}$$

where

$$\tilde{x}' = A\tilde{x} + B_2 \Gamma_0 w + B_1 w \qquad t \geq 0, \quad \tilde{x}(0) = 0. \tag{2.26}$$

This yields

$$\int_0^\infty (\gamma^2 (w^*, w)_W - (\Gamma_0^*(\Gamma_0 w^* + f_{x_0}), w)_W - (C_1^* C_1 x^*, \tilde{x})) dt = 0 \quad \forall w \in \mathcal{W}. \tag{2.27}$$

We note that

$$\tilde{x} = \text{arg inf} \left\{ \int_0^\infty (|C_1 x|_Z^2 + |u|_U^2) dt;\ x' = Ax + B_2 u + B_1 w;\ x(0) = 0 \right\},$$

so $C_1 \tilde{x} \in L^2(R^+; X)$, $\Gamma_0 w \in L^2(R^+; U)$.

Next, by Lemma 7 we have

$$\begin{aligned}&\int_0^T (C_1^* C_1 x^*, \tilde{x}) dt \\ &\quad = (p^*(T), \tilde{x}(T)) - \int_0^T (\Gamma_0 w, B_2^* p)_U \, dt + \int_0^T (B_1 w, p^*)_W \, dt \\ &\quad = (p^*(T), \tilde{x}(T)) - \int_0^T (w, \Gamma_0^*(\Gamma_0 w^* + f_{x_0}))_W \, dt + \int_0^T (B_1^* p^*, w)_W \, dt.\end{aligned}$$

Let $w \in L^2(R^+; W)$ be arbitrary but with compact support in R^+. Since (A, C_1) is detectable and $C_1 \tilde{x} \in L^2(R^+; X)$, we infer that $\lim_{t \to \infty} \tilde{x}(t) = 0$. Then letting $T \to \infty$ it follows that

$$\int_0^\infty (C_1^* C_1 x^*, \tilde{x}) dt = -\int_0^\infty ((B_1^* p^*, w)_W + (w, u^*)_U) dt,$$

so by (2.27) we get

$$\int_0^\infty (B_1^* p^* - \gamma^2 w^*, w)_W \, dt = 0.$$

This implies (2.25) as desired, and the proof of (2.4) is complete.

Let us now prove that (2.3) – (2.5) are also sufficient for the optimality of (u^*, w^*) in problem (2.2). Indeed, by Lemma 3, if (u^*, w^*, p^*) satisfy system (2.3)–(2.5) then $u^* = \Gamma w^*$. Then, since $B_1^* p^* = -\gamma^2 w^*$ by Lemma 7, it follows equality (2.27) for all $(\tilde{x}, w)$ satisfying (2.26). Hence,

$$\int_0^\infty (\gamma^2 |w^*|_W^2 - |u^*|_U^2 - |C_1 x^*|_Z^2) dt$$
$$\leq \int_0^\infty (\gamma^2 |w^* + \lambda w|_W^2 - |\Gamma_0(w^* + \lambda w) + f_{x_0}|_U^2 - |C_1 x^* + \lambda \tilde{x}|_Z^2) dt + 0(\lambda),$$
$$\lambda > 0.$$

Equivalently,

$$\psi(w^*) \leq \psi(w^* + \lambda w) + 0(\lambda) \quad \forall w \in W, \ \lambda > 0.$$

Since ψ is convex, this implies that

$$w^* = \text{arg inf}\, \psi(w) = \text{arg sup}\, K_{x_0}(\Gamma_{x_0} w, w)$$

as claimed. This completes the proof of Proposition 1.

3. Proof of Theorem 1. Assume first that the H_∞ problem has a solution, i.e., hypothesis (j) holds. Then, as seen in Proposition 1, problem (2.1) has a unique solution $(u^*, w^*) \in \mathcal{U} \times \mathcal{W}$ given by (2.2) – (2.4). We set

$$Px_0 = -p^*(0), \tag{3.1}$$

where p^* is a solution of (2.5).

LEMMA 5. *We have*

$$P \in L(X, X), \quad P = P^*, \quad P \geq 0; \tag{3.2}$$

$$(Px_0, x_0) = K_{x_0}(u^*, w^*); \tag{3.3}$$

$$p^*(t) = -Px^*(t) \quad \forall t \geq 0. \tag{3.4}$$

Proof. Let us first prove that

$$-(p^*(0), x_0) = K_{x_0}(u^*, w^*). \tag{3.5}$$

Indeed, by using the mild form of (1.1) and (2.5), we get by Lemma 7 that

$$(p(T), x^*(T)) - (p(0), x^*(0))$$
$$= \int_0^T |C_1 x^*(t)|_Z^2 dt + \int_0^T (|B_2^* p(t)|_U^2 - \gamma^{-2} |B_2^* p(t)|_W^2) dt$$
$$= \int_0^T |C_1 x^*(t)|_Z^2 + |u^*|_U^2) - \gamma^2 |w^*|_W^2) dt \quad \forall T > 0.$$

Letting T approach $+\infty$ and recalling that $\lim_{T \to \infty} p(T) = \lim_{T \to \infty} x^*(T) = 0$ we get (3.5). Next, by (3.1) and (3.5) we see that P is single valued and linear. Once again using system (2.5) we see that $(Px_0, y_0) = (Py_0, x_0)$ for all $x_0, y_0 \in X$, while by (3.3) we have

$$(Px_0, x_0) \geq \inf\{K_{x_0}(u, 0);\ u \in \mathcal{U}\} \geq 0 \quad \forall x_0 \in X, \quad \text{i.e., } P = P^* \geq 0.$$

As a supremum of lower semicontinuous convex functions $x_0 \to K_{x_0}(\Gamma_{x_0}w, w)$, the function $x_0 \to (Px_0, x_0)$ is itself lower semicontinuous. Since it is convex and everywhere finite we infer that it is continuous. Hence $P \in L(X, X)$. Note also that by inequality (2.1) it follows that

$$(Px_0, x_0) + \varepsilon \int_0^\infty |w^*|_W^2\, dt \le C|x_0|^2 \tag{3.6}$$

for some $\varepsilon > 0$ and $C > 0$ independent of x_0.

Then, again using the detectability hypothesis along with (2.12), where $x = x^*$, $u = u^*$, and $w = w^*$, we see that

$$|(x^*(t), h)| \le |e^{A_K t}x_0||h| + \int_0^t |u^*(s)|_U |B_2^* e^{A_K^*(t-s)}h|_U ds$$
$$+|h| \int_0^t \|e^{A_K(t-s)}\|_{L(X,X)}(|KC_1x^*(s)| + |B_1w^*(s)|)ds \quad \forall\, h \in X.$$

Hence

$$|x^*(t)| \le Ce^{-\omega t}|x_0| + \int_0^t e^{-\omega(t-s)}(|u^*(s)|_U + |C_1x^*(s)|_Z + |w(s)|_W)ds,$$

and by (3.6) we get

$$|x^*|_{L^2(R^+;X)} + |x^*|_{L^\infty(R^+;X)} \le C|x_0|, \tag{3.7}$$

where C is independent of x_0.

Note that for every $t \ge 0$, (u^*, w^*) is the solution of the following problem:

$$\sup_{w\in L^2(R_t^+;W)} \inf_{u\in L^2(R_t^+;U)} \left\{ \int_t^\infty (|C_1x|_Z^2 + |u|_U^2 - \gamma^2|w|_W^2)ds; \right.$$
$$\left. x' = Ax + B_2u + B_1w \,\text{in}\, R_t^+ = (t, \infty);\ x(t) = x^*(t) \right\}. \tag{3.8}$$

We may prove (3.8) by a dynamic programming argument, but we may also use Proposition 1 observing that u^*, w^*, x^*, p^* satisfy on $[t, \infty)$ system (1.1), (2.3)–(2.5) with the initial condition $x(t) = x^*(t)$, which is sufficient for optimality. Then (3.4) follows by (3.1).

Denote by $S_P(t) : X \to X$ the family of linear operators

$$S_P(t)x_0 = x^*(t) \quad \forall\, t \ge 0,$$

where x^* is the optimal state in problem (2.2).

Since $x \to x^*(t + s)$ is the solution to problem (3.8), we infer that $S_P(t + s)x_0 = S_P(t)S_P(s)x_0$, $\forall\, t, s \ge 0$.

Next, by (3.7) we see that $S_P(t) \in L(X, X)$, $\forall\, t \ge 0$. Hence $S_P(t)$ is a C_0-semigroup on X.

Let A_P be the infinitesimal generator of S_P. In Lemma 6 below we collect some properties of A_P and B_2^*P for later use.

LEMMA 6. *We have*

$$D(A_P) \subset D(B_2^*P), \tag{3.9}$$

$$B_2^* P \in L(D(A_P), U) \cap L(D(A), U), \tag{3.10}$$

$$A_P x_0 = (A - B_2 B_2^* P + \gamma^{-2} B_1 B_1^* P) x_0 \quad \forall x_0 \in D(A_P). \tag{3.11}$$

Proof. This proof is essentially the same as that of Lemma 4.5 in [3]; however, we sketch it for the reader's convenience.

By (2.6) and (3.1) we have

$$P x_0 = -e^{A^* T} P e^{A_P T} x_0 - \int_0^T e^{A^* t} C_1^* C_1 e^{A_P t} x_0 \, dt \quad \forall T > 0. \tag{3.12}$$

If $x_0 \in D(A_P)$ this yields

$$\begin{aligned} B_2^* P x_0 =& B_2^* e^{A^* T} P e^{A_P T} x_0 - B_2^* (A^*)^{-1} (e^{A^* T} C_1^* C_1 e^{A_P T} x_0 - C_1^* C_1 x_0) \\ & - \int_0^T e^{A^* t} C_1^* C_1 e^{A_P t} A_P x_0 \, dt. \end{aligned}$$

Since $B_2^* e^{A^* t} P e^{A_P t} x_0 \in L^2_{\text{loc}}(R^+; U)$, the latter makes sense for almost all $T > 0$, so $B_2^* P x_0 \in X$. Hence $D(A_P) \subset D(B_2^* P)$. In particular, this implies that $B_2^* P$ is densely defined and continuous from $D(A_P)$ to U.

On the other hand, for all $x_0 \in H$ and $z \in D(A^*)$ we have

$$\begin{aligned} &\frac{d}{dt}(S_P(t) x_0, z) \\ &\quad = (S_P(t) x_0, A^* z) + (B_2 u^*(t) + B_1 w^*(t), z) \\ &\quad = (S_P(t) x_0, A^* z) - (B_2 B_2^* P S_P(t) x_0 - \gamma^{-2} B_1 B_1^* P S_P(t) x_0, z) \quad \text{a.e. } t > 0. \end{aligned}$$

Since $B_2^* P \in L(D(A_P), U)$ and $A^{-1} B_2 \in L(X, X)$, we have

$$\begin{aligned} (A_P x_0, z) &= (x_0, A^* z) - (B_2 B_2^* P x_0 - \gamma^{-2} B_1 B_1^* P x_0, z) \\ & \quad \forall x_0 \in D(A_P) \quad \forall z \in D(A^*). \end{aligned} \tag{3.13}$$

By (3.10) the operator $A - B_2 B_2^* P + \gamma^{-2} B_1 B_1^* P$ is well defined from $D(A_P)$ to $(D(A^*))'$. Then, (3.13) follows by (3.11).

Next, arguing as in the proof of Lemma 4.1 in [3], we see that

$$A^* P \in L(D(A_P), X), \qquad A_P^* P \in L(D(A), X). \tag{3.14}$$

For $x_0 \in D(A_P)$ and $z \in X$ we have

$$\begin{aligned} (A_P x_0, P z) &= (P(A - B_2 B_2^* P + \gamma^{-2} B_1 B_1^* P) x_0, z) \\ &= (P A x_0, z) - (B_2^* P x_0, B_2^* P z) + \gamma^{-2} (B_1^* P x_0, B_1^* P z)_W. \end{aligned}$$

Hence

$$(B_2^* P x_0, B_2^* P z)_U = -(A_P x_0, P z) + (P A x_0, z) + \gamma^{-2} (B_1^* P x_0, B_1^* P z)_W,$$

and by virtue of (3.14) we get

$$|(B_2^* P x_0, B_2^* P z)_U| \le C |x_0|_{D(A_P)} |z|_{D(A)} \quad \forall x_0 \in D(A_P), \ z \in D(A),$$

and

$$|(B_2^*Px_0, B_2^*Pz)_U| \le C|x_0|_{D(A)}|z|_{D(A_P)} \quad \forall x_0 \in D(A),\ z \in D(A_P).$$

In particular, this yields $B_2^*P \in L(D(A),U) \cap L(D(A_P),U)$ as claimed.

Proof of Theorem 1 (continued). To prove that P is a solution to Riccati equation (1.8), we note first that by (3.3), (3.4), and (3.8) we have

$$(Px^*(t), x^*(t)) = \int_t^\infty (|C_1x^*(s)|_Z^2 + |u^*(s)|_U^2 - \gamma^2|w^*(s)|_W^2)ds, \qquad t \ge 0.$$

For $x_0 \in D(A_P)$ this yields

$$2(Px^*(t), A_Px^*(t)) + |C_1x^*(t)|_Z^2 - |B_2^*Px^*(t)|_U^2 = -\gamma^{-2}|B_1^*Px^*(t)|_W^2 \quad \text{a.e. } t > 0,$$

and since $B_2^*P \in L(D(A_P), X)$, we get

(3.15) $$2(Px_0, A_Px_0) + |C_1x_0|_Z^2 - |B_2^*Px_0|_U^2 + \gamma^{-2}|B_1^*Px_0|_W^2 = 0 \quad \forall x_0 \in D(A_P).$$

By (3.10) and (3.14), the previous equality extends to all $x_0 \in D(A)$, and in virtue of (3.11) we may write it as

$$2(Px_0, Ax_0) + |C_1x_0|_Z^2 - |B_2^*Px_0|_U^2 + \gamma^{-2}|B_1^*Px_0|_W^2 = 0 \quad \forall x_0 \in D(A).$$

If we differentiate the latter equation in the space $D(A)$ we get (1.8).

To prove that the semigroup e^{A_Pt} is exponentially stable we use detectability assumption (ii). Let $K \in L(X,Z)$ as in this assumption. Then we have

$$\begin{aligned} x^*(t) = e^{(A+KC_1)}x_0 &+ \int_0^t e^{(A+KC_1)(t-s)}(B_2u^*(s) + B_1w^*(s))ds \\ &- \int_0^t e^{(A+KC_1)(t-s)}KC_1x^*(s)ds \quad \forall t \ge 0. \end{aligned}$$

Since $B_1w^*,\ KC_1x^* \in L^2(R^+; X)$, it remains to show that

$$\int_0^t e^{(A+KC_1)(t-s)}B_2u^*(s)ds \in L^2(R^+; X).$$

The latter follows by assumption (i) and Lemma 5.2 in [3].

Clearly $A - B_2B_2^*P = A_P - \gamma^{-2}B_1B_1^*P$ is the infinitesimal generator of a C_0-semigroup on X. Let $x_0 \in D(A_P)$. Then, multiplying the equation

(3.16) $$x' = (A - B_2B_2^*P)x, \qquad x(0) = x_0$$

by Px and using (3.15), we get

$$\frac{d}{dt}(Px(t), x(t)) = -|Cx(t)|_Z^2 - |B_2^*Px(t)|_U^2 - \gamma^{-2}|B_1^*Px(t)|_W^2, \qquad t > 0.$$

Hence

$$\int_0^\infty (|C_1x(t)|_Z^2 + |B_2^*Px(t)|_U^2 + |B_1^*Px(t)|_W^2)dt < \infty.$$

If we write (3.16) as

$$x' = A_P x - \gamma^{-2} B_1^* P x, \qquad x(0) = 0,$$

and use the facts that $B_1^* P x \in L^2(R^+; W)$ and $e^{A_P t}$ is exponentially stable, we conclude that $x \in L^2(R^+; X)$. Hence, by Datko's theorem, $e^{(A - B_2 B_2^* P)t}$ is exponentially stable.

Let $w \in L^2(R^+; W) \cap C^1(R^+; W)$ be arbitrary but fixed, and let $y \in C^1(R^+; X)$ be the solution of the following equation:

$$y' = (A - B_2 B_2^* P) y + B_1 w, \qquad y(0) = x_0 \in D(A_P). \tag{3.17}$$

This yields

$$\begin{aligned}\frac{d}{dt}&(Py(t), y(t)) \\ &= -|C_1 y(t)|_Z^2 - |B_2^* P y(t)|_U^2 - \gamma^{-2} |B_1^* P x(t)|_W^2 + 2(B_1^* P y(t), w(t))_W \\ &= -|C_1 y(t)|_Z^2 - |B_2^* P y(t)|_U^2 - |\gamma w - \gamma^{-1} B_1^* P y(t)|_W^2 + \gamma^2 |w|_W^2 \qquad \forall t \geq 0.\end{aligned}$$

Hence

$$\begin{aligned}\int_0^\infty &(|C_1 y(t)|_Z^2 + |B_2^* P y(t)|_U^2 - \gamma^2 |w(t)|_W^2) dt \\ &= -\gamma^2 \int_0^\infty |\bar{w}(t)|_W^2 \, dt + (P x_0, x_0),\end{aligned} \tag{3.18}$$

where $\bar{w} = w - \gamma^{-2} B_1^* P y$.

On the other hand, we have for $x_0 = 0$,

$$\|w\|_{L^2(R^+;W)} \leq \|\bar{w}\|_{L^2(R^+;W)} + \beta \|y\|_{L^2(R^+;X)} \leq \alpha \|\bar{w}\|_{L^2(R^+;W)},$$

because $y' = A_P y + B_1 \bar{w}$.

Substituting the latter in (3.18) we get

$$\int_0^\infty (|C_1 y|_Z^2 + |B_2^* P y|_U^2 - \gamma^2 |w|_W^2) dt \leq -\delta \|w\|_{L^2(R^+;W)}^2 \quad \forall w \in L^2(R^+; W),$$

where $\delta > 0$ is independent of w. Hence $\|S_{-B_2^* P}\| < \gamma$, which completes the "only if" part of the proof.

We now complete the "if" part of the proof. Suppose that $P = P^* \geq 0$ is a solution of (1.8) satisfying (1.7) such that $A_P = A - B_2 B_2^* P + \gamma^{-2} B_1 B_1^* P$ is exponentially stable. Then, arguing as above, we infer that $A - B_2 B_2^* P = A_P - \gamma^{-2} B_1 B_1^* P$ is exponentially stable. Then if y is the solution of Cauchy problem (3.17), we get as above (see (3.18)) that

$$B_2^* P e^{(A - B_2 B_2^* P)t} \in L(X, L^2(R^+; U)),$$

i.e., $-B_2^* P \in \mathcal{F}$. Moreover, by (3.18) we infer as above that $\|S_{-B_2^* P}\| < \gamma$.

Let us show now that the solution P of Riccati equation (1.8) having the properties stated in Theorem 1 is unique. Indeed, if P_1 and P_2 are two such solutions of (1.8) we have

$$(A_{P_1} x, (P_1 - P_2) y) + ((P_1 - P_2) x, A_{P_2} y) = 0 \quad \forall x, y \in D(A_{P_1}) \cap D(A_{P_2}),$$

and this yields

$$\frac{d}{dt}(e^{A_{P_1} t} x, (P_1 - P_2) e^{A_{P_2} t} y) = 0 \quad \forall t \geq 0.$$

Since $e^{A_{P_i} t}$, $i = 1, 2$, are exponentially stable, we get

$$((P_1 - P_2) y, x) = 0 \quad \forall x, y \in D(A_{P_1}) \cap D(A_{P_2}),$$

and by density we infer that $P_1 = P_2$. This completes the proof of Theorem 1.

4. H_∞ control with boundary disturbance input. We shall consider here system (1.1), where assumption (i) is replaced by

(k) $B_2 \in L(U, X)$, $A^{-1}B_1 \in L(U, X)$, and

$$\int_0^T |B_1^* e^{A^* t} x|_W^2 \, dt \le C_T |x|^2 \quad \forall x \in X,\ T > 0. \tag{4.1}$$

Then, for every $w \in L^2(R^+; W)$ and $u \in L^2(R^+; U)$, system (1.1) has a unique mild solution x given by the variation of constant formula (1.4). Given $\gamma > 0$, an H_∞ controller is by definition a stabilizing feedback controller that guarantees the inequality $\|S_F\| < \gamma$, where $S_F \in L(L^2(R^+; W),\ L^2(R^+; Z))$ is the closed loop operator $S_F(w) = C_1 x + D_{12} F x$. It is not clear whether Theorem 1 remains valid in this case. Proposition 2 below represents only a partial result (see [6] for other recent results on this case).

PROPOSITION 2. *Let $\gamma > 0$ and suppose that assumptions* (k), (ii), (iii), *and* (j) *hold. Then there exists $P \in L(X, X)$ such that $P = P^* \ge 0$, $B_1^* P \in L(D(A), U) \cap L(D(A_P), U)$, P satisfies Riccati equation* (1.8), *$A_P = A - B_2 B_2^* P + \gamma^{-2} B_1 B_1^* P$, and $A - B_2 B_2^* P$ generates exponentially stable semigroups. Conversely, if a such a solution P to* (1.8) *exists, then $F = -B_2^* P$ is an exponentially stabilizing feedback controller and guarantees $\|S_F\| \le \gamma$.*

Proof. Since the proof is essentially the same as that of Theorem 1, it will only be sketched. We note first that Proposition 1 remains valid under present assumptions. Indeed, the proof of Lemma 1 remains unchanged (with some simplifications because B_2 is bounded) and (2.4) follows by the same argument; note that by assumption (k) and (2.6), $B_1^* P \in L^2(0, T; W)$ for all $T > 0$. Then, we define $P \in L(X, X)$ by (3.1), and Lemma 5 remains valid. Define

$$\tilde{S}_P(t) x_0 = x^*(t) \quad \forall t \ge 0; \tag{4.2}$$

then $\tilde{S}_P$ is a C_0-semigroup on X. Denote by $\tilde{A}_P$ its infinitesimal generator. Arguing as in the proof of Lemma 6, and using assumption (k), it follows that

$$\begin{aligned}
&D(\tilde{A}_P) \subset D(B_1^* P),\\
&\quad B_1^* P \subset L(D(\tilde{A}_P), U) \cap L(D(A), U),\\
&\quad \tilde{A}_P x_0 = (A - B_2 B_2^* P + \gamma^{-2} B_1 B_1^* P) x_0 \quad \forall x_0 \in D(A_P).
\end{aligned}$$

Then, as in the proof of Theorem 1, it follows that P satisfies (1.8), and the semigroups generated by $\tilde{A}_P$ and $A - B_2 B_2^* P$ are exponentially stable.

Now assume the existence of $P \in L(X, X)$ satisfying Riccati equation (1.8) and the conditions stated in Proposition 2. Then, for $w \in C^1(R^+; W) \cap L^2(R^+; W)$ and $x_0 = 0$, system (3.17) has a solution $y \in C^1(R^+; D(A))$ and y is weakly differentiable. (This follows by assumption (4.1).) Hence $B_1^* P y \in L^\infty_{\mathrm{loc}}(R^+; W)$, and using (1.8) we get

$$\begin{aligned}
\frac{d}{dt}(Py(t), y(t)) = -|C_1 y(t)|_Z^2 - |B_2^* P y(t)|_U^2 - \gamma w(t) - \gamma^{-1} B_1^* P y(t)|_W^2\\
+ \gamma^2 |w(t)|_W^2 \quad \text{a.e. } t > 0.
\end{aligned}$$

This yields

$$\int_0^T (|C_1 y(t)|_Z^2 - |B_2^* P y(t)|_U^2 - \gamma^2 |w(t)|_W^2) dt \le 0 \quad \forall T > 0.$$

Hence $\|S_{-B_2^* P}\| \le \gamma$ as claimed.

5. Some remarks on H_∞ control with dynamic feedback. Consider again the input-output system

$$x' = Ax + B_2u + B_1w, \quad z = C_1x + D_{12}u \tag{5.1}$$

with the observation

$$y = C_2x + D_{12}\eta. \tag{5.2}$$

Here $C_2 \in L(X,Y)$, $D_{12} \in L(W_0,Y)$, η is a perturbation modelling the measurement error, and Y, W_0 are Hilbert spaces. The aim is to find a dynamic feedback controller of the form

$$\begin{gathered} u = Lp, \\ p' = (A+M)p + Ny, \qquad p(0) = 0 \end{gathered} \tag{5.3}$$

that exponentially stabilizes the system and reduces the influence of w on z. The linear maps $M : X \to X$, $L : X \to U$, $N : Y \to X$, define a feedback controller $K = (M,L,N)$, which substituted into system (5.1) yields

$$\begin{gathered} x' = Ax + B_2Lp + B_1w, \qquad t \in R^+, \\ p' = NC_2x + (A+M)p + ND_{12}\eta, \\ x(0) = 0, \qquad p(0) = 0. \end{gathered} \tag{5.4}$$

Equivalently,

$$\begin{gathered} \begin{pmatrix} x \\ p \end{pmatrix}' = \mathcal{A}\begin{pmatrix} x \\ p \end{pmatrix} + \left\| \begin{matrix} B_1w \\ ND_{21} \end{matrix} \right\|, \qquad t \in R^+, \\ \begin{pmatrix} x \\ p \end{pmatrix}(0) = \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \end{gathered} \tag{5.4$'$}$$

where

$$\mathcal{A} = \left\| \begin{matrix} A & B_2L \\ NC_2 & A+M \end{matrix} \right\|. \tag{5.5}$$

Here we shall restrict ourselves to feedback controllers $K = (M,L,N)$ with the following properties:

(a) $M \in L(D(A),(D(A^*))'), L \in L(D(A),U)$, $N \in L(Y,X)$.

(b) The operator $\mathcal{A} : X \times X \to X \times X$ is the infinitesimal generator of an exponentially stable C_0-semigroup on $X \times X$.

(c) $D(\mathcal{A}) \subset D(L) \times X$ and $\left\| \begin{smallmatrix} 0 & L \\ 0 & 0 \end{smallmatrix} \right\| e^{\mathcal{A}t}$ is continuous from X to $L^2(R^+; X \times X)$.

For such a feedback controller K, define $S_K : L^2(R^+;W) \to L^2(R^+;Z)$, $S_K(w) = z = C_1x + D_{12}Lp$, where (x,p) is the solution of (5.4).

The feedback controller K is called an H_∞ control if it satisfies (a), (b), (c), and

$$\|S_K\| < \gamma. \tag{5.6}$$

In addition to (i), (ii), and (iii), we shall assume that

(l) the pair (B_1^*, A^*) is exponentially detectable;

(ll) $D_{12}|B_1^*, D_{12}^*| = |0, I|$.

Proposition 3 below extends only partially the standard results of H_∞ control theory with dynamic measurement feedback [2], [4].

PROPOSITION 3. *Let $\gamma > 0$. If there exists an H_∞ controller satisfying* (5.6) *then there exist $P, Q \in L(X,X)$, $P = P^* \geq 0$, $Q = Q^* \geq 0$, which satisfy the following conditions*:

(m) $B_2^*P \in L(D(A),U)$, $A + (\gamma^{-2}B_1B_1^* - B_2B_2^*)P$ *generates an exponentially stable semigroup and*

$$(Ax,Py) + (Px,Ay) - (P(B_2B_2^* - \gamma^{-2}B_1B_1^*)Px,y) + (C_1^*C_1x,y) = 0 \quad \forall\, x,y \in D(A). \tag{5.7}$$

(mm) $A^* + (\gamma^{-2}C_1^*C_1 - C_2^*C_2)Q$ *generates an exponentially stable semigroup and*

$$(Ax,Qy) + (Qx,Ay) - ((C_2^*C_2 - \gamma^{-2}C_1^*C_1)Qx,Qy) + (B_1^*B_1x,y) = 0 \quad \forall\, x,y \in D(A). \tag{5.8}$$

(mmm) $(I - \gamma^{-2}PQ)^{-1} \in L(X,X)$ *and*

$$Q(I - \gamma^{-2}PQ)^{-1} \geq 0. \tag{5.9}$$

If these conditions hold then the feedback controller $K = (M,L,N)$ given by

$$\begin{aligned} M &= (\gamma^{-2}B_1B_1^* - B_2B_2^*)P - Q(I - \gamma^{-2}PQ)^{-1}C_2^*C_2, \\ L &= B_2^*P, \\ N &= -Q(I - \gamma^{-2}PQ)^{-1}C_2^*, \end{aligned} \tag{5.10}$$

is an H_∞ controller guaranteeing the closed loop inequality $\|S_K\| \leq \gamma$.

We omit the proof, which follows by Theorem 1 and Proposition 2 by the duality arguments developed in [2], [9], and [1].

In fact, the direct approach of [1] can be used mutatis mutandis in the present situation, but we do not give details.

6. Appendix. We shall prove here the following lemma.

LEMMA 7. *Let $(x,p) \in C([a,b];X) \times C([a,b];X)$ be a mild solution to the system*

$$x' = Ax + B_2u + f, \qquad t \in [a,b], \tag{6.1}$$

$$p' = -A^*p + g, \tag{6.2}$$

where $u \in L^2(a,b;U)$ and $f,g \in L^1(0,T;X)$. Then

$$\begin{aligned} &\int_a^b ((u(t),B_2^*p(t))_U + (f(t),p(t)) + (g(t),x(t)))dt \\ &\quad = (x(b),p(b)) - (x(a),p(a)). \end{aligned} \tag{6.3}$$

Proof. For simplicity we take $a = 0$ and $b = T$. By virtue of assumption (i) and Fubini's theorem we may write

$$\begin{aligned} &\int_0^T (g(t),x(t))dt \\ &= \int_0^T \left(g(t), e^{At}x_0 + \int_0^t e^{A(t-s)}f(s)ds\right) dt \\ &\quad + \int_0^T \left(g(t), \int_0^t e^{A(t-s)}B_2u(s)ds\right) dt \\ &= \int_0^T \left(g(t), e^{At}x_0 + \int_0^t e^{A(t-s)}f(s)ds\right) dt \\ &\quad + \int_0^T \left(u(s), \int_s^T B_2^*e^{A^*(t-s)}g(t)dt\right) ds. \end{aligned} \tag{6.4}$$

Recalling that

$$p(t) = e^{A^*(T-t)}p(T) - \int_t^T e^{A^*(s-t)}g(s)ds, \qquad t \in [0,T],$$

and substituting the latter in (7.4) we get (7.3) as claimed. (We recall that, by virtue of assumption (i), the operator

$$g \to \int_s^T B_2^* e^{A^*(t-s)}g(t)dt$$

is continuous from $L^1(0,T;X)$ to $L^2(0,T;U)$.)

Acknowledgments. The author is indebted to R. Triggiani for useful remarks and for pointing out a gap in the proof of Lemma 3 in the first draft of the manuscript. The author also thanks the anonymous referee for numerous suggestions that helped improve the presentation of this work.

REFERENCES

[1] A. Bensoussan and P. Bernhard, *Remarks on the theory of robust control*, Optimization, Optimal Control and Partial Differential Equations, V. Barbu et al., eds., Birkhäuser-Verlag, Basel, Boston, Berlin, 1992, pp. 149–166.

[2] J. Doyle, K. Glover, P. Khargonekar, and B. Francis, *State-space solutions to standard H_2 and H_∞-control problems*, IEEE Trans. Automat. Control. AC-34 (1989), pp. 831–847.

[3] F. Flandoli, I. Lasiecka, and R. Triggiani, *Algebraic Riccati equation with non-smoothing observation arising in Euler–Bernoulli boundary control problems*, Ann. Mat. Pura Appl. TCLIII (1988), pp. 307–383.

[4] A. Ichikawa, *H_∞-control and min–max problems in Hilbert spaces*

[5] I. Lasiecka and R. Triggiani, *Differential and Algebraic Riccati Equations with Applications to Boundary/Point Control Problems: Continuous Theory and Approximation Theory*, in Lecture Notes in Control and Information Sciences 164, Springer-Verlag, Berlin, New York, 1991.

[6] C. McMillan and R. Triggiani, *Min-max game theory and algebraic Riccati equations for boundary control problems with continuous input-output map*, Appl. Math. Optim., 29 (1994), pp. 1–66.

[7] A. Pazy, *Semigroups of Linear Operators and Applications to Partial Differential Equations*, Springer-Verlag, New York, Berlin, Heidelberg, 1983.

[8] A. J. Pritchard and D. Salmon, *A semi-group theoretic approach for systems with unbounded input and output operators*, SIAM J. Control Optim., 25 (1987), pp. 121–144.

[9] B. van Keulen, M. Peters, and R. Curtain, *H_∞-control with state feedback; the infinite dimensional case*, preprint.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 33, No. 3, pp. 702–717, May 1995

© 1995 Society for Industrial and Applied Mathematics
003

ALMOST SURE STABILIZABILITY AND RICCATI'S EQUATION OF LINEAR SYSTEMS WITH RANDOM PARAMETERS*

PHILIPPE BOUGEROL†

Abstract. Linear systems with stationary time-varying parameters are considered. The notions of almost sure stabilizability and almost sure detectability are explored. Under these properties the asymptotic behavior of the associated Riccati equation is described. In particular, the stability of the Kalman filter and the convergence of linear observers are proved.

Key words. Riccati equation, Kalman filtering, observers, stochastic parameters, stabilizability, detectability, linear discrete-time systems

AMS subject classifications. 93C05, 93E11, 60G35, 34F05

1. Introduction.

Consider either the filtering linear system

$$\begin{aligned} X_{n+1} &= A_n X_n + B_n \varepsilon_n, \\ Y_{n+1} &= C_n X_n + \eta_n, \end{aligned} \tag{1}$$

or the linear control system

$$X_{n+1} = A_n X_n + B_n U_n, \qquad Y_n = C_n X_n, \tag{2}$$

where $X_n \in \mathbb{R}^d$ is the state vector, $Y_n \in \mathbb{R}^q$ is an observation, $(\varepsilon_n, \eta_n) \in \mathbb{R}^p \times \mathbb{R}^q$ is a Gaussian white noise, and $U_n \in \mathbb{R}^p$ is a control. We suppose that $\{(A_n, B_n, C_n),\ n \in \mathbb{Z}\}$ is a stationary ergodic sequence of matrices of appropriate dimension. Under weak conditions we show that the error system of the Kalman filter associated with (1) is almost surely exponentially stable, and the linear system (2) can be almost surely stabilized.

When the parameters are not random, analogous results are well known under strong hypotheses (either uniform stabilizability and detectability or uniform controllability and observability); see, e.g., Jazwinski [10], Anderson and Moore [2], and the recent survey in the introduction of De Nicolao [9]. These hypotheses are too stringent and usually do not hold for systems with random parameters. We shall see that they can be significantly weakened when the parameters are stationary by making use of ideas from ergodic theory.

Systems with stationary parameters are used in several models and are worth studying in view of the applications. Stationary processes can be either deterministic (e.g., almost periodic sequences) or nondeterministic (e.g., identically independent random variables, functions of ergodic Markov chains, autoregressive moving average (ARMA) sequences, etc.). They occur, for instance, in the descriptions of random (or multirate deterministic) sampling of the control of failure-prone production plants, and in adaptive stochastic control. Some of these systems are out of order at random times; this prevents them from verifying uniform assumptions. These examples and references to previous works are described in detail in [5]. Inspired by Snyder and Fishman [14], filtering for systems with random parameters is studied, in particular, by Viano [15] under a strong stability hypothesis. Here we shall suppose that the parameters are observed. The opposite situation is studied, for instance, by Wonham [17] and De Koning [8].

Let us describe our main results. It is convenient to introduce the following, perhaps unusual, definitions. The random matrices are defined on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$.

* Received by the editors April 8, 1992; accepted for publication (in revised form) December 22, 1993.

† Laboratoire de Probabilités, Université Pierre et Marie Curie, 4, Place Jussieu, 75252 Paris, France (bougerol@ccr.jussieu.fr).

DEFINITION 1.1. *A sequence $\{A_n, n \in \mathbb{Z}\}$ of $d \times d$ random matrices is called almost surely weakly stable if*

$$\lim_{n \to +\infty} \|A_n A_{n-1} \dots A_1\| = 0 \quad a.s.$$

It is called almost surely exponentially stable if there exists $\gamma > 0$ with the following property: for any $\varepsilon > 0$ and almost all $\omega \in \Omega$, there is a $C(\omega) > 0$ such that

$$\|A_{n-1}(\omega)A_{n-2}(\omega)\dots A_{n-m}(\omega)\| \leq C(\omega)e^{-m\gamma}e^{(|n|+m)\varepsilon}$$

for all $n \in \mathbb{Z}$ and all $m \geq 1$.

Note that when (A_n) is almost surely exponentially stable, the solution of the linear equation $X_{n+1} = A_n X_n$ converges almost surely to 0 exponentially fast as time progresses. It will be proven that these definitions are exactly what is needed here, where we meet random sequences without moment. Their first properties are explored in §2. We say that the linear systems above are almost surely weakly or exponentially stabilizable if there is a feedback control such that the associated closed loop system is almost surely weakly or exponentially stable, respectively (see Definition 2.4 for precision). In §3, we show using [5] that a sufficient condition for almost sure exponential stabilizability is the existence of an integer $n \geq 1$ such that

$$\mathbb{P}(B_n B_n^* + A_n B_{n-1} B_{n-1}^* A_n^* + \dots + (A_n \dots A_2) B_1 B_1^* (A_2^* \dots A_n^*) \text{ is invertible}) \neq 0.$$

This is called weak controllability. For instance, when $d = 1$ this condition is always fulfilled unless $B_n = 0$ almost surely for all n. In §4 we prove that an almost surely weakly stabilizable system is actually almost surely exponentially stabilizable. Similar results hold for the corresponding notions of detectability and weak observability defined by duality.

In §5 we suppose that the linear system is almost surely weakly stabilizable and weakly detectable. We study the Riccati difference equation

$$P_{n+1} = B_n B_n^* + A_n P_n (I + C_n^* C_n P_n)^{-1} A_n^*, \tag{3}$$

where $P_n, n \in \mathbb{N}$, are nonnegative symmetric matrices. This equation is usually written as follows:

$$\begin{aligned} P_{n+1} &= B_n B_n^* + (A_n - K_n C_n) P_n A_n^*, \\ K_n &= A_n P_n C_n^* (I + C_n P_n C_n^*)^{-1}, \end{aligned} \tag{4}$$

where K_n is called the gain matrix associated with P_n. Let us recall how it occurs in the filtering of (1). We suppose that the Gaussian white noise (ε_n, η_n) has a covariance matrix equal to the identity and is independent of X_0 and the sigma-algebra $\mathcal{F}_0$ generated by the parameters. The random variable X_0 is Gaussian and independent of $\mathcal{F}_0$. For any $n \in \mathbb{N}$, let $\mathcal{F}_n = \sigma(\mathcal{F}_0, Y_1, \dots, Y_n)$. This represents the observation at time n. Let

$$\hat{X}_n = \mathbb{E}(X_n/\mathcal{F}_n), \qquad P_n = \mathbb{E}\left((X_n - \hat{X}_n)(X_n - \hat{X}_n)^*/\mathcal{F}_n\right).$$

Then the Kalman recursive equations are given by (4) and

$$\hat{X}_{n+1} = A_n \hat{X}_n + K_n (Y_{n+1} - C_n \hat{X}_n)$$

(see, e.g., Whittle [16, p. 260 and Thm. 5.7.1]). The main result of this section is the existence of a unique stationary solution $(\bar{P}_n)$ of the Riccati difference equation (Theorem 5.1). It is analogous to the "moving equilibrium" solution defined in Kalman [11]. This solution is attractive in the sense that if (P_n) is another solution of this equation, then $\bar{P}_n - P_n$ converges almost surely to 0 exponentially fast as n tends to $+\infty$ (Theorem 5.3). Moreover, the filter is almost surely exponentially stable and thus efficient (Theorem 5.6). The initial conditions are rapidly forgotten (this is useful since they are often chosen arbitrarily). For instance, let us suppose only that $\mathbb{E}(\log\|A_0\|) < 0$, or more generally that the Lyapunov exponents of the sequence (A_n) are negative. This is the condition under which (X_n), given by (1), is itself a stationary process (see, e.g., [6]). Then the system is almost surely weakly stabilizable and weakly detectable. Thus the results above hold, although the system does not fulfill the "uniform" conditions considered in the literature on time-varying parameter systems. In §6 we briefly consider applications to control theory. An efficient observer for linear systems with stationary coefficients is given. We also show that the quadratic cost converges almost surely when the horizon increases to infinity (this latter result is of limited practical interest for nondeterministic systems since the optimal controls depend on the parameters indexed by the future).

There are some connections between this work and multiplicative ergodic theory that are explained in [3]. The properties of $\bar{P}_n$ are related to the hyperbolicity of associated Hamiltonian matrices (see Remark 4.4), but are not only consequences of this fact. An important technical difficulty arises from the fact that, generally, $\log^+\|\bar{P}_n\|$ is not integrable. Therefore, the sequence $(A_n - K_n C_n)$ does not have Lyapunov exponents. This explains why some proofs are rather involved (see also Remark 5.7).

Let us make some technical comments on our filtering linear model (1). Several variants are often used; for instance, one could replace the second equation by $Y_n = C_n X_n + \eta_n$. Our choice is justified by the fact that duality is more transparent here. Anyway, all the models are essentially equivalent and it is easy to transpose our results to other ones. The hypotheses on the random parameters model can also be weakened. It is, in fact, sufficient that there is a sigma-algebra $\mathcal{F}_0$ such that, if $\mathcal{F}_n = \sigma(\mathcal{F}_0, Y_1, \dots, Y_n)$, then (A_n, B_n, C_n) is $\mathcal{F}_n$-measurable and (ε_n, η_n) is independent of $\mathcal{F}_n$ and X_n. In that so-called conditionally Gaussian case, the filter is also given by the previous equations (see Whittle [16], Chen, Kumar, and van Schuppen [7]), and, therefore, our results hold true. One can also eliminate the assumption that the white noise (ε_n, η_n) has a unit covariance matrix at the price of some complications. Finally, we remark that the well-known equivalence between (4) and (3) follows directly from the relation $(C_n^* C_n P_n + I)^{-1} = I - C_n^*(I + C_n P_n C_n^*)^{-1} C_n P_n$ or from Whittle [16, Thm. 5.7.1].

This paper is a logical continuation of [5]; its results were announced in [4].

2. Preliminaries. In this section we explore the notions of stability introduced in Definition 1.1. We shall also use them for nonrandom sequences (i.e., when Ω has only one element), in which case we do not write "almost surely." The definition of exponential stability given there is weaker than the usual one (in which $\varepsilon = 0$ is allowed); it thus deserves some comments. Let us first note the following result, the proof of which is straightforward. Note that the assertion is no more true with "lim sup" instead of "lim" (for instance, take $a_{2n} = \frac{1}{3}2^{|n|}, a_{2n-1} = 2^{-|n|}$).

LEMMA 2.1. *Let $\{a_n, n \in \mathbb{Z}\}$ be a sequence of positive real numbers. If for some $\alpha, \beta > 0$,*

$$\lim_{n\to+\infty} \frac{1}{n} \log(a_n \dots a_1) = -\alpha, \qquad \lim_{n\to+\infty} \frac{1}{n} \log(a_{-1} \dots a_{-n}) = -\beta,$$

then the sequence is exponentially stable.

In this paper, we will mainly be interested in stationary (in the strict sense) sequences of matrices. If (A_n) is such a sequence and $\log^+ \|A_0\|$ is integrable, then its largest Lyapunov exponent γ is defined as

$$\gamma = \lim_{n\to+\infty} \frac{1}{n}\,\mathbb{E}(\log \|A_n \dots A_1\|).$$

Throughout the paper we choose the operator norm associated with the euclidean structure for a matrix (usually, this choice is not important). In this case, we also have $\gamma = \inf_{n>0} \frac{1}{n}\mathbb{E}(\log \|A_n \dots A_1\|)$.

PROPOSITION 2.2. *Let $\{A_n, n \in \mathbb{Z}\}$ be a stationary ergodic sequence of matrices such that* $\log^+ \|A_0\|$ *is integrable. Then the following three conditions are equivalent*:

(a) *the sequence (A_n) is weakly stable almost surely*;

(b) *the Lyapunov exponents of the sequence (A_n) are negative*;

(c) *the sequence (A_n) is almost surely exponentially stable.*

We need the following well-known lemma.

LEMMA 2.3. *If $\{a_n, n \in \mathbb{Z}\}$ is a stationary sequence of real random variables such that a_0 is integrable, then, almost surely,* $\lim_{|n|\to+\infty} \frac{1}{n} a_n = 0$.

Proof. The proof follows from the Borel–Cantelli lemma since, for any $\varepsilon > 0$,

$$\sum_{k=-\infty}^{+\infty} \mathbb{P}(|a_k| > |k|\varepsilon) = \sum_{k=-\infty}^{+\infty} \mathbb{P}(|a_0| > |k|\varepsilon)$$
$$\leq \mathbb{P}(|a_0| > 0) + 2\sum_{k=1}^{+\infty} \mathbb{P}\left(\frac{|a_0|}{\varepsilon} > k\right) \leq 1 + 2\mathbb{E}\left(\frac{|a_0|}{\varepsilon}\right) < +\infty. \qquad \square$$

Proof of Proposition 2.2. (a) $\Rightarrow$ (b) is proved in [6, Lem. 3.4]. Let us assume that (b) holds. Then there exists $k > 0$ such that $\mathbb{E}(\log \|A_k \dots A_1\|)$ is negative. Let $M_n = A_{(n+1)k} \dots A_{nk+1}$. It follows from Lemma 2.1 and from Birkhoff's ergodic theorem that the sequence $\{\|M_n\|, n \in \mathbb{Z}\}$ is almost surely exponentially stable. This easily implies that (c) holds true by using the fact that if $m < qk < pk \leq n$, then

$$\|A_n A_{n-1} \dots A_{m+1}\| \leq \|A_n \dots A_{pk+1}\| \cdot \|M_{p-1} \dots M_q\| \cdot \|A_{qk} \dots A_{m+1}\|,$$

and Lemma 2.1 applied to the sequence $a_n = \log^+ \|A_n\|$. Finally, $\frac{1}{n} \log \|A_n \dots A_1\|$ converges almost surely to the largest Lyapunov exponent as $n \to +\infty$ (see, e.g., Kingman [12]). Thus (c) $\Rightarrow$ (a). $\square$

This proposition shows that "almost surely exponentially stable" is a kind of substitute for "with negative Lyapunov exponents" for stationary sequences of matrices, the norms of which are log integrable. It will appear in the following sections that non-log-integrable sequences occur naturally in the control/filtering setting, and are even unavoidable. This explains the choice of Definition 1.1. From a technical point of view the robustness properties exhibited by Lemmas 4.1 and 5.5 will be crucial for us.

Let us give a simple natural example where only our notion of stability is fulfilled. We consider a sequence $a_n, n \in \mathbb{Z}$, of independent and identically distributed positive random variables such that $\mathbb{P}(a_0 > M) > 0$ for any $M > 0$, and $\mathbb{E}(\log a_0) < 0$. This sequence is almost surely exponentially stable. On the other hand, for any $M > 0$,

$$\sum_{n=-\infty}^{+\infty} \mathbb{P}(a_n > M) = \sum_{n=-\infty}^{+\infty} \mathbb{P}(a_0 > M) = +\infty,$$

which implies by the Borel–Cantelli lemma that almost surely there is no $C > 0$ such that, for all $n \in \mathbb{Z}$, $|a_n| \leq C$. Therefore, the condition given in the definition of exponential stability does not hold when $\varepsilon = 0$. Let us introduce some definitions.

DEFINITION 2.4. *The linear system* (1), (2), *or the sequence* $\{(A_n, B_n), n \in \mathbb{Z}\}$, *is said to be almost surely weakly stabilizable if there is a sequence* $F_n, n \in \mathbb{Z}$, *of* $p \times d$ *random matrices such that, almost surely, the sequence* $\{A_n + B_nF_n, n \in \mathbb{Z}\}$ *is weakly stable. It is called almost surely exponentially stabilizable if, moreover,* $\{A_n + B_nF_n, n \in \mathbb{Z}\}$ *is almost surely exponentially stable and* $\lim_{|n|\to+\infty} \frac{1}{n} \log^+ \|F_n\| = 0$.

These systems, or the sequence $\{(A_n, C_n), n \in \mathbb{Z}\}$, *are said to be almost surely weakly detectable if there is a sequence* (G_n) *of* $d \times q$ *random matrices such that, almost surely,* $\{A_n + G_nC_n, n \in \mathbb{Z}\}$ *is weakly stable. It is called almost surely exponentially detectable if, moreover,* $\{A_n + G_nC_n, n \in \mathbb{Z}\}$ *is almost surely exponentially stable and* $\lim_{|n|\to+\infty} \frac{1}{n} \log^+ \|G_n\| = 0$.

For example, it follows from Proposition 2.2 that if the Lyapunov exponents of the sequence $\{A_n, n \in \mathbb{Z}\}$ are negative, then the system is almost surely exponentially stabilizable and detectable. In the definition of exponential stabilizability, the condition on $\frac{1}{n} \log^+ \|F_n\|$ indicates that these matrices are, in a sense, not too large. It replaces the boundedness assumption of classical systems and can be technically useful (see the proof of Theorem 5.1, for instance). Note that this condition is not required for weak stabilizability. The dual of linear system (2) is

$$x_{n+1} = A^*_{-n}x_n + C^*_{-n}v_n, \qquad y_{n+1} = B^*_{-n}x_n, \tag{5}$$

where $x_n \in \mathbb{R}^d$, $y_n \in \mathbb{R}^p$, and $v_n \in \mathbb{R}^q$. Thus a system is almost surely weakly or exponentially detectable if and only if its dual is almost surely weakly or exponentially stabilizable.

In the following sections the random variables are defined on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$. Usually, we will suppose that the following hypothesis holds (where $\log^+ x = \max(0, \log x)$).

Hypothesis (H).

(i) The matrices A_n are invertible.

(ii) $\log^+ \|A_0\|, \log^+ \|A_0^{-1}\|, \log^+ \|B_0\|$, and $\log^+ \|C_0\|$ are integrable.

(iii) There exists a bijective ergodic transformation $\theta : \Omega \to \Omega$ which preserves $\mathbb{P}$ such that, for any $n \in \mathbb{Z}$ and $\omega \in \Omega$,

$$A_n(\omega) = A_0(\theta^n(\omega)), \quad B_n(\omega) = B_0(\theta^n(\omega)), \quad C_n(\omega) = C_0(\theta^n(\omega)).$$

Under condition (iii) above the sequence $\{(A_n, B_n, C_n), n \in \mathbb{Z}\}$ is stationary and ergodic. Conversely, if this latter property holds there is no loss of generality in supposing that (iii) is true.

3. Weak controllability implies, almost surely, stabilizability. We first define weak controllability and weak observability. Similar notions were already introduced in [5] for slightly different systems.

DEFINITION 3.1. *Let* $R_n = C_n^*C_n$ *and* $S_n = B_nB_n^*$. *The linear system* (1), *or the sequence* $\{(A_n, B_n), n \in \mathbb{Z}\}$, *is said to be weakly controllable if for some* $n \geq 1$,

$$\mathbb{P}(\det\{S_n + A_nS_{n-1}A_n^* + \cdots + (A_n \ldots A_2)S_1(A_2^* \ldots A_n^*)\} \neq 0) \neq 0.$$

This system, or the sequence $\{(A_n, C_n), n \in \mathbb{Z}\}$, *is said to be weakly observable if, for some* $n \geq 1$,

$$\mathbb{P}(\det\{R_1 + A_1^*R_2A_1 + \cdots + (A_1^* \ldots A_{n-1}^*)R_n(A_{n-1} \ldots A_1)\} \neq 0) \neq 0.$$

The dual of a weakly controllable system is weakly observable, and the converse is true as well. The purpose of this section is to prove that a weakly controllable or observable linear system with stationary coefficients is almost surely exponentially stabilizable or detectable, respectively. This result will be obtained as a consequence of a more precise result that will be very useful later. The strategy of the proof is the same as in the classical situation: we shall associate with the system a filtering problem that admits an almost surely exponentially stable closed loop system.

Notation. In all of the following sections $\mathfrak{P}$ denotes the set of $d \times d$ symmetric nonnegative definite matrices. It is equipped with its natural order (i.e., $P \geq Q$ when $P - Q \in \mathfrak{P}$).

For any $n \in \mathbb{Z}$ define $\phi_n : \mathfrak{P} \to \mathfrak{P}$ by

$$(6) \qquad \phi_n(P) = B_n B_n^* + A_n P (I + C_n^* C_n P)^{-1} A_n^*, \qquad P \in \mathfrak{P}.$$

It follows from (3) that Riccati's difference equation can be written as

$$\phi_n(P_n) = P_{n+1}.$$

We shall need the three ensuing lemmas. The first one is well known (see, e.g., Whittle [16, Chap. 5, §2, form. (11)]).

LEMMA 3.2. *Let (P_n) be a solution of the Riccati equation. For any $d \times q$ matrix K,*

$$P_{n+1} \leq (A_n - KC_n) P_n (A_n - KC_n)^* + KK^* + B_n B_n^*,$$

and equality holds when K is equal to the associated gain matrix K_n.

Let us prove the following variant of Lyapunov's lemma.

LEMMA 3.3. *Let $\{(A_n, B_n), n \in \mathbb{Z}\}$ be a weakly controllable sequence. We suppose that there is a sequence $\{Q_n, n \in \mathbb{Z}\}$ of invertible matrices in $\mathfrak{P}$ with the three following properties*:

(a) *for all $n \in \mathbb{Z}$, $Q_{n+1} = A_n Q_n A_n^* + B_n B_n^*$*;

(b) *the sequence $\{(A_n, B_n, Q_n), n \in \mathbb{Z}\}$ is stationary and ergodic*;

(c) *almost surely, $\lim_{|n| \to +\infty} \frac{1}{n} \log^+ \|Q_n\| = \lim_{|n| \to +\infty} \frac{1}{n} \log^+ \|Q_n^{-1}\| = 0$.*

Then the sequence $\{A_n, n \in \mathbb{Z}\}$ is almost surely exponentially stable.

Proof. Let us introduce the following notation: for $n \in \mathbb{Z}$ *and* $r \in \mathbb{N}$,

$$M_n = Q_{n+1}^{-\frac{1}{2}} A_n Q_n^{\frac{1}{2}}, \qquad T_n = Q_{n+1}^{-\frac{1}{2}} B_n B_n^* Q_{n+1}^{-\frac{1}{2}},$$

$$M_{n,r} = M_n M_{n-1} \ldots M_{n-r+1}, \qquad T_{n,r} = \sum_{i=0}^{r} M_{n,i} T_{n-i} M_{n,i}^*.$$

Then $I = M_n M_n^* + T_n$ and, more generally, $I = M_{n,r} M_{n,r}^* + T_{n,r}$, where I is the identity matrix. Hence $\|M_{n,r}\| \leq 1$. On the other hand, since $\{(A_n, B_n), n \in \mathbb{Z}\}$ is weakly controllable, there exists some $k > 0$ such that $\mathbb{P}(T_{n,k}$ is invertible$) \neq 0$ and thus $\mathbb{P}(\|M_{n,k}\| < 1) \neq 0$. These relations imply that the Lyapunov exponents of the stationary sequence (M_n) are negative. Therefore, this sequence is almost surely exponentially stable by Proposition 2.2. Since

$$\begin{aligned} \|A_n \ldots A_k\| &= \|Q_{n+1}^{\frac{1}{2}} M_n \ldots M_k Q_k^{-\frac{1}{2}}\| \\ &\leq \|Q_{n+1}^{\frac{1}{2}}\| \cdot \|M_n \ldots M_k\| \cdot \|Q_k^{-\frac{1}{2}}\|, \end{aligned}$$

it then follows from (c) that (A_n) is almost surely exponentially stable. □

The following lemma is from Wos [18]. To obtain this formulation we apply his Theorem 2 to $f = g \circ \theta - g$.

LEMMA 3.4 (Wos). *For any measurable function* $g\colon \Omega \to \mathbb{R}^+$, *if* $\limsup_{n\to+\infty} \frac{1}{n} g\circ\theta^n = 0$ *almost surely then* $\limsup_{n\to+\infty} \frac{1}{n} g\circ\theta^{-n} = 0$ almost surely.

We now derive a slight improvement of [5]. It will be generalized in §5.

THEOREM 3.5. *We consider a weakly controllable and weakly observable linear system such that Hypothesis* (H) *holds. Then there exists a unique solution* $\{\bar{P}_n, n\in\mathbb{Z}\}$ *of the Riccati difference equation* (4) *with values in* $\mathfrak{P}$ *such that* $\{(A_n, B_n, C_n, \bar{P}_n), n\in\mathbb{Z}\}$ *is a stationary and ergodic sequence. Moreover, almost surely* $\bar{P}_n$ *is invertible,*

$$\lim_{|n|\to\infty} \frac{1}{n} \log^+ \|\bar{P}_n\| = \lim_{|n|\to\infty} \frac{1}{n} \log^+ \|\bar{P}_n^{-1}\| = 0, \tag{7}$$

and the sequence $\{A_n - \bar{K}_n C_n, n\in\mathbb{Z}\}$ *is almost surely exponentially stable, where* $\bar{K}_n$ *is the gain matrix associated with* $\bar{P}_n$.

Proof. Let $S_n = B_n B_n^*$ and $R_n = C_n^* C_n$. We consider the equation

$$Q_{n+1} = (A_n Q_n A_n^* + S_n)(I + R_{n+1} A_n Q_n A_n^* + R_{n+1} S_n)^{-1}, \tag{8}$$

where $Q_n \in \mathfrak{P}$ (notice the occurrence of R_{n+1}). It follows from [3, Thm. 2.4] that this equation has a unique stationary solution, say $\{\bar{Q}_n, n\in\mathbb{Z}\}$. By construction, $\bar{Q}_n = \bar{Q}_0 \circ \theta^n$, which implies that the sequence $\{(A_n, B_n, C_n, \bar{Q}_n), n\in\mathbb{Z}\}$ is stationary and ergodic. Moreover, by [3, Prop. 3.4 and form. (24)], almost surely,

$$\lim_{n\to+\infty} \frac{1}{n} \log^+ \|\bar{Q}_n\| = \lim_{n\to+\infty} \frac{1}{n} \log^+ \|\bar{Q}_n^{-1}\| = 0.$$

It follows from Lemma 3.4 that, almost surely,

$$\lim_{n\to+\infty} \frac{1}{n} \log^+ \|\bar{Q}_{-n}\| = \lim_{n\to+\infty} \frac{1}{n} \log^+ \|\bar{Q}_{-n}^{-1}\| = 0.$$

Now, we remark that (Q_n) is a solution of (8) if and only if $P_n = A_{n-1} Q_{n-1} A_{n-1}^* + S_{n-1}$ is a solution of the Riccati recursion (4). Therefore, $\bar{P}_n = A_{n-1}\bar{Q}_{n-1}A_{n-1}^* + S_{n-1}$ is the unique stationary solution of (4). We check that (7) holds true with the relations above and Lemma 2.3. The sequence $\{(A_n, B_n), n\in\mathbb{Z}\}$ is weakly controllable, thus the sequence $\{(A_n, (B_n\, \bar{K}_n)), n\in\mathbb{Z}\}$ is also weakly controllable. By making use of the relation

$$A_n - \bar{K}_n C_n = A_n - (B_n\, \bar{K}_n)(0\, C_n^*)^*,$$

it follows from Anderson and Moore [2, Lem. 3.1] that $\{(A_n - \bar{K}_n C_n, (B_n\, \bar{K}_n)), n\in\mathbb{Z}\}$ is also weakly controllable. Since, by Lemma 3.2,

$$\bar{P}_{n+1} = (A_n - \bar{K}_n C_n)\bar{P}_n(A_n - \bar{K}_n C_n)^* + \bar{K}_n \bar{K}_n^* + B_n B_n^*,$$

the almost sure exponential stability of $(A_n - \bar{K}_n C_n)$ is a consequence of Lemma 3.3 above. □

THEOREM 3.6. *Under Hypothesis* (H), *a weakly controllable sequence is almost surely exponentially stabilizable, and a weakly observable sequence is almost surely exponentially detectable.*

Proof. Let us suppose that the sequence $\{(A_n, C_n), n\in\mathbb{Z}\}$ is weakly observable and show that this sequence is almost surely detectable. Let B_n be equal to the identity matrix of order d. Then the system associated with (A_n, B_n, C_n) is weakly observable and weakly controllable. Thus, by Theorem 3.5, $\{A_n - \bar{K}_n C_n, n\in\mathbb{Z}\}$ is almost surely exponentially stable, where $\bar{K}_n$ is the gain matrix associated with $\bar{P}_n$. On the other hand, $\|(I+C_n\bar{P}_nC_n^*)^{-1}\| \leq 1$ so that

$$\log\|\bar{K}_n\| \leq \log\|\bar{P}_n\| + \log\|C_n\| + \log\|A_n\|.$$

Using (7) and Lemma 2.3, we see that almost surely $\lim_{|n|\to\infty} \frac{1}{n}\log^+\|\bar{K}_n\| = 0$. The almost sure exponential detectability is thus proved. The other statement follows by duality. □

4. Almost sure weak stabilizability implies almost sure exponential stabilizability.

LEMMA 4.1. *Consider a sequence $\{M_n, n\in\mathbb{Z}\}$ of $d\times d$ matrices written in block form as*

$$M_n = \begin{pmatrix} A_n & B_n \\ 0 & C_n \end{pmatrix},$$

where the respective sizes of the matrices A_n, B_n, C_n are $p\times p$, $p\times q$, $q\times q$ (and $p+q=d$). We suppose that the sequences $\{A_n, n\in\mathbb{Z}\}$ and $\{C_n, n\in\mathbb{Z}\}$ are exponentially stable and that $\lim_{|n|\to+\infty}\frac{1}{n}\log^+\|B_n\| = 0$. Then $\{M_n, n\in\mathbb{Z}\}$ is exponentially stable.

Proof. For all $k,n\in\mathbb{Z}$ such that $k<n$, we consider the matrices $A_{n,k}$, $B_{n,k}$, and $C_{n,k}$, defined by the formula

$$\begin{pmatrix} A_{n,k} & B_{n,k} \\ 0 & C_{n,k} \end{pmatrix} = M_nM_{n-1}\dots M_{k+1}.$$

Since the sequences (A_n) and (C_n) are exponentially stable, there exists $\gamma>0$ such that, for any $\varepsilon>0$ and some $\alpha>0$,

$$\|A_{n,k}\| \leq \alpha e^{(k-n)\gamma+(|n|+|k|)\varepsilon}, \qquad \|C_{n,k}\| \leq \alpha e^{(k-n)\gamma+(|n|+|k|)\varepsilon},$$

when $k<n$. Besides, there exists $\beta>0$ such that $\|B_n\|\leq\beta e^{|n|\varepsilon}$ for all $n\in\mathbb{Z}$. Let $\mu>0$ be a constant such that $r\leq\mu e^{r\varepsilon}$ for any $r\in\mathbb{N}$. It is readily seen that $B_{n,k} = \sum_{m=k+1}^{n} A_{n,m}B_mC_{m-1,k}$, therefore,

$$\begin{aligned}\|B_{n,k}\| &\leq \alpha^2\beta e^{(k-n)\gamma}\sum_{m=k+1}^{n} e^{\gamma+(|m-1|+2|m|+|k|+|n|)\varepsilon} \\ &\leq \alpha^2\beta|n-k|e^{(k+1-n)\gamma}e^{4(|k|+|n|)\varepsilon} \\ &\leq \alpha^2\beta\mu e^{(k+1-n)\gamma}e^{5(|k|+|n|)\varepsilon}.\end{aligned}$$

Since all the norms on a finite-dimensional space are equivalent, there exists a universal constant λ such that

$$\|M_nM_{n-1}\dots M_{k+1}\| \leq \lambda\max\{\|A_{n,k}\|, \|B_{n,k}\|, \|C_{n,k}\|\}.$$

Thus, the previous inequalities imply that (M_n) is exponentially stable. □

The next proposition is inspired by the canonical structure theorem of linear systems.

PROPOSITION 4.2. *Consider a linear system* (1) *for which Hypothesis* (H) *holds true. Then there exists a random orthogonal matrix Γ_0 such that, if $\Gamma_n = \Gamma_0\circ\theta^n$ for all $n\in\mathbb{Z}$, we can write the ensuing decompositions in blocks as follows*:

$$\Gamma_n^{-1}A_n\Gamma_{n-1} = \begin{pmatrix} A_n^{1,1} & 0 & A_n^{1,3} \\ A_n^{2,1} & A_n^{2,2} & A_n^{2,3} \\ 0 & 0 & A_n^{3,3} \end{pmatrix}, \qquad \Gamma_n^{-1}B_n = \begin{pmatrix} B_n^{1,1} \\ B_n^{2,1} \\ 0 \end{pmatrix},$$

$$C_n\Gamma_{n-1} = (C_n^{1,1} \quad 0 \quad C_n^{1,3}),$$

where the following three properties hold:

(a) *The linear system associated with the sequence* $\{(\tilde{A}_n, \tilde{B}_n, \tilde{C}_n), n \in \mathbb{Z}\}$ *is weakly controllable*, *where*

$$\tilde{A}_n = \begin{pmatrix} A_n^{1,1} & 0 \\ A_n^{2,1} & A_n^{2,2} \end{pmatrix}, \quad \tilde{B}_n = \begin{pmatrix} B_n^{1,1} \\ B_n^{2,1} \end{pmatrix}, \quad \tilde{C}_n = (C_n^{1,1}\, 0).$$

(b) *The linear system associated with* $\{(A_n^{1,1}, B_n^{1,1}, C_n^{1,1}), n \in \mathbb{Z}\}$ *is weakly controllable and weakly observable.*

(c) *If the linear system associated with* $\{(A_n, B_n, C_n), n \in \mathbb{Z}$ *is almost surely weakly stabilizable or almost surely weakly detectable, then* $\{A_n^{3,3}, n \in \mathbb{Z}\}$ *or* $\{A_n^{2,2}, n \in \mathbb{Z}\}$ *is almost surely exponentially stable*, *respectively*.

Proof. For each $\omega \in \Omega$ and each $n \in \mathbb{Z}$, let $W_n(\omega)$ be the linear subspace of $\mathbb{R}^d$ spanned by

$$\{A_n(\omega)A_{n-1}(\omega)\ldots A_{n-k+1}(\omega)B_{n-k}(\omega)x; x \in \mathbb{R}^d, k \in \mathbb{N}\}.$$

Then $W_n = W_0 \circ \theta^n$, where θ is the transformation that appears in assumption (iii) of Hypothesis (H). We remark that $A_1 W_0$ is contained in W_1. Since A_1 is invertible, this implies that dim $W_0 \leq$ dim W_1. Thus, by ergodicity there exists an integer $r \in [0, d]$ such that dim $W_0 = r$ almost surely. Let V be the linear subspace of $\mathbb{R}^d$ spanned by the first r vectors of the canonical basis. We choose a random orthogonal matrix Q_0 such that $Q_0 V = W_0$, and set $Q_n = Q_0 \circ \theta^n$ for each $n \in \mathbb{Z}$. Then $Q_n V = W_n$, and thus $Q_n^{-1} A_n Q_{n-1} V$ is contained in V. Moreover, the image of $Q_n^{-1} B_n$ is contained in V, since the image of B_n is contained in W_n. Therefore, we can write the following in block form:

$$Q_n^{-1} A_n Q_{n-1} = \begin{pmatrix} A_n^1 & A_n^2 \\ 0 & A_n^3 \end{pmatrix}, \quad Q_n^{-1} B_n = \begin{pmatrix} B_n^1 \\ 0 \end{pmatrix}, \quad C_n Q_{n-1} = (C_n^1\, C_n^2),$$

where, for example, the respective size of the matrices A_n^1, B_n^1, and C_n^1 is $r \times r$, $r \times p$, and $q \times r$. Since V is spanned by the ranges of the matrices $A_0^1(\omega)\ldots A_{-k+1}^1(\omega)B_{-k}^1(\omega)$, $k \in \mathbb{N}$, we see that $\{(A_n^1, B_n^1, C_n^1)\}$ is weakly controllable.

Let us now suppose that the system associated with (A_n, B_n, C_n) is almost surely weakly stabilizable. Then, by definition, there is a sequence $\{F_n, n \in \mathbb{Z}\}$ such that $\{A_n + B_n F_n, n \in \mathbb{Z}\}$ is almost surely weakly stable. Of course, $\{Q_n^{-1}(A_n + B_n F_n) Q_{n-1}, n \in \mathbb{Z}\}$ is then also almost surely weakly stable. If we write in block form $F_n Q_{n-1} = (F_n^1\, F_n^2)$, where F_n^1 and F_n^2 are $p \times r$ and $(d - p) \times r$ matrices, respectively, then

$$Q_n^{-1}(A_n + B_n F_n) Q_{n-1} = \begin{pmatrix} A_n^1 + B_n^1 F_n^1 & A_n^2 + B_n^1 F_n^2 \\ 0 & A_n^3 \end{pmatrix}.$$

This means that the sequence (A_n^3) is almost surely weakly stable. Since $\log^+ \|A_0^3\|$ is integrable, this sequence is also almost surely exponentially stable by Proposition 2.2.

If we apply the previous construction to the dual of the system associated with $\{(A_n^1, B_n^1, C_n^1), n \in \mathbb{Z}\}$, we find that there exists a sequence $\Lambda_n, n \in \mathbb{Z}$, of $p \times p$ orthogonal matrices such that $\Lambda_n = \Lambda_0 \circ \theta^n$, and we can write

$$\Lambda_n^{-1} A_n^1 \Lambda_{n-1} = \begin{pmatrix} A_n^{1,1} & 0 \\ A_n^{2,1} & A_n^{2,2} \end{pmatrix}, \quad \Lambda_n^{-1} B_n^1 = \begin{pmatrix} B_n^{1,1} \\ B_n^{2,1} \end{pmatrix}, \quad C_n^1 \Lambda_{n-1} = (C_n^{1,1}\, 0),$$

where the system $\{(A_n^{1,1}, B_n^{1,1}, C_n^{1,1})\}$ is weakly observable. Let

$$\Gamma_n = Q_n \begin{pmatrix} \Lambda_n & 0 \\ 0 & I \end{pmatrix}.$$

Then the decompositions in blocks hold. The proof is completed by simple computation. □

THEOREM 4.3. *Under Hypothesis* (H), *an almost surely weakly stabilizable or detectable system is almost surely exponentially stabilizable or detectable, respectively.*

Proof. We consider a system associated with (A_n, B_n, C_n) and use the decomposition introduced in Proposition 4.2. Since $\{(\tilde{A}_n, \tilde{B}_n, \tilde{C}_n), n \in \mathbb{Z}\}$ is weakly controllable, it follows from Theorem 3.6 that this sequence is almost surely exponentially stabilizable. Thus, there is a sequence $\{\tilde{F}_n, n \in \mathbb{Z}\}$ such that $\lim_{|n|\to\infty} \frac{1}{n} \log^+ \|\tilde{F}_n\| = 0$ and $\{\tilde{A}_n + \tilde{B}_n\tilde{F}_n, n \in \mathbb{Z}\}$ is almost surely exponentially stable. Now we suppose that the original system is weakly stabilizable. Then $(A_n^{3,3})$ is also almost surely exponentially stable. Let

$$D_n = \begin{pmatrix} A_n^{1,3} \\ A_n^{2,3} \end{pmatrix}, \qquad F_n = (\tilde{F}_n\, 0)\Gamma_{n-1}^{-1}.$$

Since $\lim_{|n|\to\infty} \frac{1}{n} \log^+ \|D_n\| = 0$ by Lemma 2.3, and

$$A_n + B_nF_n = \Gamma_n \begin{pmatrix} \tilde{A}_n + \tilde{B}_n\tilde{F}_n & D_n \\ 0 & A_n^{3,3} \end{pmatrix} \Gamma_{n-1}^{-1},$$

we deduce from Lemma 4.1 that $\{A_n + B_nF_n, n \in \mathbb{Z}\}$ is an almost surely exponentially stable sequence. This proves that the system is almost surely exponentially stabilizable. The other statement is obtained by duality.

Remark 4.4. Let us consider an almost surely weakly stabilizable and detectable system. It is possible to prove that all the Lyapunov exponents of the associated Hamiltonian matrices are nonzero by using Proposition 4.2 and [3]. We will not use this fact and refer to [3] for the details of this statement.

5. Asymptotic behavior of the Riccati difference equation. In this section we establish the main properties of the Riccati equation under the assumption that the system is weakly stabilizable and detectable almost surely. For each $\omega \in \Omega$, these conditions on the parameters $(A_n(\omega), B_n(\omega), C_n(\omega))$ are very weak. Therefore, the following theorem cannot be deduced from existing results on time-varying systems. Below $\bar{K}_n$ is the gain matrix associated with $\bar{P}_n$.

THEOREM 5.1. *We consider linear system* (1). *We suppose that Hypothesis* (H) *holds and that this system is weakly stabilizable and weakly detectable almost surely. Then there exists a unique stationary process* $\{\bar{P}_n, n \in \mathbb{Z}\}$ *with values in* $\mathfrak{P}$ *that is a solution of Riccati's difference equation* (4). *The process* $\{(A_n, B_n, C_n, \bar{P}_n), n \in \mathbb{Z}\}$ *is stationary ergodic, and* $\{A_n - \bar{K}_nC_n, n \in \mathbb{Z}\}$ *is almost surely exponentially stable.*

Proof. It is well known that the mappings $\phi_n : \mathfrak{P} \to \mathfrak{P}$ defined by (6) are increasing for the usual order on $\mathfrak{P}$ (this follows, for instance, from Lemma 3.2). Therefore, for each fixed $n \in \mathbb{Z}$, the sequence of nonnegative symmetric matrices

$$P_{n,k} = (\phi_{n-1} \circ \phi_{n-2} \circ \cdots \circ \phi_{n-k})(0), \qquad k \geq 1,$$

is increasing in $\mathfrak{P}$. Since, by Theorem 4.3, the system is almost surely exponentially detectable, there exists a sequence $\{G_n, n \in \mathbb{Z}\}$ of $d \times q$ matrices such that $\{A_n - G_nC_n, n \in \mathbb{Z}\}$ is almost surely exponentially stable and $\lim_{|n|\to+\infty} \frac{1}{n}\log^+ \|G_n\| = 0$. Let $M_n = A_n - G_nC_n$ and $T_n = G_nG_n^* + B_nB_n^*$. Lemma 3.2 gives

$$P_{n,k} \leq T_{n-1} + M_{n-1}P_{n-1,k-1}M_{n-1}^*,$$

which leads, by induction, to

$$P_{n,k} \leq T_{n-1} + M_{n-1}T_{n-2}M_{n-1}^* + \cdots + (M_{n-1}\dots M_{n-k+1})T_{n-k}(M_{n-k+1}^*\dots M_{n-1}^*).$$

Since (M_n) is almost surely exponentially stable, there exists $\gamma > 0$ such that, almost surely, for any $\varepsilon > 0$ and some $\alpha \geq 1$,

$$\|M_{n-1}\dots M_{n-m}\| \leq \alpha\, e^{-\gamma m}\, e^{(|n|+m)\varepsilon}$$

for all $n \in \mathbb{Z}$ and $m \geq 1$. On the other hand, it follows from integrability condition (ii) of Hypothesis (H) and from Lemma 2.3 that, almost surely, for some $\beta > 0$, $\|T_n\| \leq \beta e^{\varepsilon|n|}$ for all $n \in \mathbb{Z}$. These inequalities yield that, almost surely,

$$\|P_{n,k}\| \leq \alpha^2\beta \sum_{m=0}^{k-1} e^{-2\gamma m} e^{3(|n|+m+1)\varepsilon},$$

and, if $3\varepsilon < 2\gamma$,

$$\|P_{n,k}\| \leq \alpha^2\beta\, e^{3(|n|+1)\varepsilon}\{1 - e^{(3\varepsilon-2\gamma)}\}^{-1}. \tag{9}$$

This shows that the increasing sequence $P_{n,k}, k \in \mathbb{N}$, is bounded. Therefore, this sequence converges almost surely when $k \to +\infty$ to a limit in $\mathfrak{P}$, denoted by $\bar{P}_n$. Since $\phi_n(\bar{P}_n) = \bar{P}_{n+1}$, $\{\bar{P}_n, n \in \mathbb{Z}\}$ is a solution of Riccati's difference equation. It follows from condition (iii) of Hypothesis (H) that, for each $n \in \mathbb{Z}$, $\bar{P}_n = \bar{P}_0 \circ \theta^n$. This implies that $\{(A_n, B_n, C_n, \bar{P}_n), n \in \mathbb{Z}\}$ is a stationary ergodic process. Inequality (9) is also satisfied by $\bar{P}_n$, and therefore, almost surely,

$$\lim_{|n|\to+\infty} \frac{1}{n}\log^+ \|\bar{P}_n\| = 0. \tag{10}$$

Let us now prove that $\{A_n - \bar{K}_nC_n, n \in \mathbb{Z}\}$ is almost surely exponentially stable. We will use Proposition 4.2 and its notation. First, we remark that if (P_n) and (K_n) are solutions of the Riccati equation associated with the sequence $\{(A_n, B_n, C_n), n \in \mathbb{Z}\}$, then $(\Gamma_{n-1}P_n\Gamma_{n-1}^{-1})$ and $(\Gamma_n^{-1}K_n)$ are solutions of the Riccati equation associated with $\{(\Gamma_n^{-1}A_n\Gamma_{n-1}, \Gamma_n^{-1}B_n, C_n\Gamma_{n-1}), n \in \mathbb{Z}\}$. Making use of this remark and writing the matrices $\Gamma_{n-1}P_{n,k}\Gamma_{n-1}^{-1}$ in block form, it is easily seen that we can write

$$\Gamma_{n-1}\bar{P}_n\Gamma_{n-1}^{-1} = \begin{pmatrix} \bar{P}_n^{1,1} & \bar{P}_n^{1,2} & 0 \\ \bar{P}_n^{2,1} & \bar{P}_n^{2,2} & 0 \\ 0 & 0 & 0 \end{pmatrix}, \qquad \Gamma_n^{-1}\bar{K}_n = \begin{pmatrix} \bar{K}_n^{1,1} \\ \bar{K}_n^{2,1} \\ 0 \end{pmatrix}, \tag{11}$$

where

$$\bar{P}_n^1 = \begin{pmatrix} \bar{P}_n^{1,1} & \bar{P}_n^{1,2} \\ \bar{P}_n^{2,1} & \bar{P}_n^{2,2} \end{pmatrix}$$

is a solution of the Riccati equation associated with $\{(\tilde{A}_n, \tilde{B}_n, \tilde{C}_n), n \in \mathbb{Z}\}$. Moreover, $(\bar{P}_n^{1,1})$ is a stationary solution of the Riccati equation associated with the sequence $(A_n^{1,1}, B_n^{1,1}, C_n^{1,1})$, and $\bar{K}_n^{1,1}$ is the associated gain matrix. Since $\{(A_n^{1,1}, B_n^{1,1}, C_n^{1,1}), n \in \mathbb{Z}\}$ is weakly controllable and weakly observable (see Proposition 4.2), it follows from Theorem 3.5 that $\{A_n^{1,1} - \bar{K}_n^{1,1}C_n^{1,1}, n \in \mathbb{Z}\}$ is almost surely exponentially stable. We note that

$$\Gamma_n^{-1}(A_n - \bar{K}_nC_n)\Gamma_{n-1} = \begin{pmatrix} A_n^{1,1} - \bar{K}_n^{1,1}C_n^{1,1} & 0 & A_n^{1,3} - \bar{K}_n^{1,1}C_n^{1,3} \\ A_n^{2,1} - \bar{K}_n^{2,1}C_n^{1,1} & A_n^{2,2} & A_n^{2,3} - \bar{K}_n^{2,1}C_n^{1,3} \\ 0 & 0 & A_n^{3,3} \end{pmatrix}.$$

The sequences $(A_n^{2,2})$ and $(A_n^{3,3})$ are almost surely exponentially stable by Proposition 4.2. On the other hand, it is easy to deduce from Lemma 2.3 and (10) that, almost surely,

$$\lim_{|n|\to+\infty} \frac{1}{n} \log^+ \|A_n - \bar{K}_nC_n\| = 0.$$

Thus it follows from Lemma 4.1 that the sequence

$$\begin{pmatrix} A_n^{1,1} - \bar{K}_n^{1,1}C_n^{1,1} & 0 \\ A_n^{2,1} - \bar{K}_n^{2,1}C_n^{1,1} & A_n^{2,2} \end{pmatrix}$$

is also almost surely exponentially stable. This in turn implies $\Gamma_n^{-1}(A_n - \bar{K}_nC_n)\Gamma_{n-1}$, and thus $A_n - \bar{K}_nC_n$ is almost surely exponentially stable by another application of Lemma 4.1. The matrices Γ_n are orthogonal, thus $\{A_n - \bar{K}_nC_n, n \in \mathbb{Z}\}$ is also almost surely exponentially stable.

Finally, let us prove that $(\bar{P}_n)$ is the unique stationary solution of Riccati's equation. Let (P_n) be another stationary solution. It follows from Lemma 3.2 that

$$(12) \qquad P_{n+1} - \bar{P}_{n+1} \leq (A_n - \bar{K}_nC_n)(P_n - \bar{P}_n)(A_n - \bar{K}_nC_n)^*.$$

This implies that, for $n \geq 0$, if $D_n = (A_{-1} - \bar{K}_{-1}C_{-1})\dots(A_{-n} - \bar{K}_{-n}C_{-n})$, then

$$P_0 - \bar{P}_0 \leq D_n(P_{-n} - \bar{P}_{-n})D_n^*.$$

By stationarity, $P_{-n} - \bar{P}_{-n}$ is bounded in probability. Since D_n converges to 0 almost surely, the right-hand side above converges to 0 in probability as $n \to +\infty$. This proves that $P_0 \leq \bar{P}_0$. On the other hand, for all $k \geq 1$, since $P_{-k} \geq 0$,

$$P_0 = (\phi_{-1} \circ \cdots \circ \phi_{-k})(P_{-k}) \geq (\phi_{-1} \circ \cdots \circ \phi_{-k})(0) = P_{0,k},$$

hence $P_0 \geq \bar{P}_0$ since $P_{0,k}$ converges to $\bar{P}_0$ as $k \to +\infty$. This proves that $P_0 = \bar{P}_0$. □

LEMMA 5.2. *Consider a weakly controllable system. Let $P_n, n \geq 0$, be the solution of the Riccati equation* (4) *such that $P_0 = 0$. Then, almost surely, P_n is positive definite for all $n \geq 0$ large enough.*

Proof. The sequence $Q_n = P_n(I + R_nP_n)^{-1}, n \in \mathbb{N}$, is a solution of the difference equation (8). Using the weak controllability assumption, we deduce from [3, Prop. 1.5 (ii) and Lem. 2.2], that, almost surely, Q_n is positive definite for all n large enough, which implies Lemma 5.2. □

The following demonstration is inspired by a proof presented in Anderson and Moore [1, § 4.4] for systems with constant coefficients.

THEOREM 5.3. *Under the assumptions of Theorem* 5.1, *there exists $\gamma > 0$ such that, for any $P \in \mathfrak{P}$ almost surely,*

$$(13) \qquad \overline{\lim_{n\to+\infty}} \frac{1}{n} \log \|\bar{P}_n - P_n\| \leq -\gamma,$$

where $P_n, n \in \mathbb{N}$, is the solution of the Riccati equation for which $P_0 = P$.

Proof. We will need the following two classical relations:

$$P_{n+1} \geq (A_n - K_n C_n) P_n (A_n - K_n C_n)^*, \tag{14}$$

$$P_{n+1} - \bar{P}_{n+1} = (A_n - K_n C_n)(P_n - \bar{P}_n)(A_n - \bar{K}_n C_n)^*. \tag{15}$$

Relation (14) follows immediately from Lemma 3.2. On the other hand, since P_n and $\bar{P}_n$ are solutions of (4),

$$P_{n+1} - \bar{P}_{n+1} = (A_n - K_n C_n) P_n A_n^* - A_n \bar{P}_n (A_n - \bar{K}_n C_n)^*,$$
$$(A_n - K_n C_n) P_n C_n^* \bar{K}_n^* = K_n \bar{K}_n^* = K_n C_n \bar{P}_n (A_n - \bar{K}_n C_n)^*,$$

which yields (15). Since the sequence $(A_n - \bar{K}_n C_n)$ is almost surely exponentially stable, we deduce from (10) and (12) that, almost surely,

$$\lim_{n \to +\infty} \frac{1}{n} \log^+ \|P_n\| = 0.$$

Let us first suppose that for some (maybe random) index $k \geq 0$, P_k is positive definite. Then, by (14) the previous relation leads to

$$\lim_{n \to +\infty} \frac{1}{n} \log^+ \|(A_n - K_n C_n)(A_{n-1} - K_{n-1} C_{n-1}) \dots (A_k - K_k C_k)\| = 0.$$

In this case, the conclusion of the theorem follows immediately from (15) and the exponential stability of the sequence $(A_n - \bar{K}_n C_n)$. In particular, it holds true when P is nondegenerate.

Now, let us suppose that $P = 0$. By making use of Proposition 4.2 and its notation (see also (11)), we can write the block decomposition

$$\Gamma_{n-1} P_n \Gamma_{n-1}^{-1} = \begin{pmatrix} P_n^1 & 0 \\ 0 & 0 \end{pmatrix}, \qquad \Gamma_{n-1} \bar{P}_n \Gamma_{n-1}^{-1} = \begin{pmatrix} \bar{P}_n^1 & 0 \\ 0 & 0 \end{pmatrix},$$

where $\{P_n^1, n \geq 0\}$ is the solution of the Riccati equation associated with the weakly controllable sequence $(\tilde{A}_n, \tilde{B}_n, \tilde{C}_n)$ that satisfies $P_0^1 = 0$, and where $(\bar{P}_n^1)$ is its stationary solution. Therefore, in order to prove that (13) holds when $P = 0$, it suffices to consider weakly controllable systems. In this case, it follows from Lemma 5.2 that, almost surely, P_k is nondegenerate for some $k > 0$ large enough. Thus we deduce from the first part of the proof that the theorem also holds when $P = 0$.

Finally, let P be an arbitrary matrix in $\mathfrak{P}$. Let P_n^0, P_n, or P_n', be the solutions of the Riccati equation such that $P_0^0 = 0$, $P_0 = P$, or $P_0' = P + I$, respectively. Then the relation $P_0^0 \leq P_0 \leq P_0'$ and the monotonicity of the transformations ϕ_n imply that $P_n^0 \leq P_n \leq P_n'$. Thus

$$\|P_n - \bar{P}_n\| \leq \max(\|P_n^0 - \bar{P}_n\|, \|P_n' - \bar{P}_n\|).$$

Since $P_0^0 = 0$ and P_0' is nondegenerate, we have just proved that (13) holds for (P_n^0) and for (P_n'). It follows from this inequality that it also holds for (P_n). □

We now consider the filtering setup. Let us suppose that X_0 is a Gaussian vector with mean $\hat{X}_0$ and covariance matrix P_0, and let $\hat{X}_n = \mathbb{E}(X_n / \mathcal{F}_n)$. We know that, conditionally on $\mathcal{F}_n$, the random vector $X_n - \hat{X}_n$ has a Gaussian law with mean 0 and covariance matrix equal to P_n. Therefore, for any $y \in \mathbb{R}^d$,

$$\mathbb{E}(\exp(i\langle y, X_n - \hat{X}_n \rangle)) = \mathbb{E}(\mathbb{E}(\exp(i\langle y, X_n - \hat{X}_n \rangle)/\mathcal{F}_n))$$
$$= \mathbb{E}\left(\exp\left(-\frac{1}{2}\langle y, P_n y \rangle\right)\right).$$

Thus the following corollary follows from the previous theorem.

COROLLARY 5.4. *Under the assumptions of Theorem* 5.1, $X_n - \hat{X}_n$ *converges in distribution to the probability measure* μ *on* $\mathbb{R}^d$ *with Fourier transform given by the following: for any* $y \in \mathbb{R}^d$,

$$\int e^{i\langle y,x\rangle}\, d\mu(x) = \mathbb{E}\left(\exp\left(-\frac{1}{2}\langle y, \bar{P}_0 y\rangle\right)\right).$$

The adaptation of Definition 1.1 to sequences indexed by $\mathbb{N}$ is obvious. Of course the following lemma also holds for sequences indexed by $\mathbb{Z}$, but we shall need this setup.

LEMMA 5.5. *Let* $\{A_n, n \in \mathbb{N}\}$ *and* $\{B_n, n \in \mathbb{N}\}$ *be two sequences of* $d \times d$ *matrices. If* $\{A_n, n \in \mathbb{N}\}$ *is exponentially stable and* $\limsup_{n\to+\infty} \frac{1}{n}\log\|B_n\| < 0$, *then* $\{A_n + B_n, n \in \mathbb{N}\}$ *is also exponentially stable.*

Proof. There exists $\gamma > 0$ such that, for any $\varepsilon > 0$ for some $C > 0$,

$$\|A_p \dots A_{q+1}\| \le Ce^{p\varepsilon}e^{(q-p)\gamma},$$

for all $0 \le q < p$. When $0 \le q < p-1$, let $M_{p,q} = B_pA_{p-1}\dots A_{q+2}A_{q+1}$ and $M_{p,p-1} = B_p$. Then, $\|M_{p,q}\| \le \xi_p e^{(q-p+1)\gamma}$, where $\xi_p = Ce^{p\varepsilon}\|B_p\|$. If $0 \le k < n$, then

$$\begin{aligned}(A_n + {}&B_n)\dots(A_{k+1} + B_{k+1})\\ &= \sum_{m=0}^{n-k}\sum_{k<r_1<\cdots<r_m\le n} A_n \dots A_{r_m+1}M_{r_m,r_{m-1}}M_{r_{m-1},r_{m-2}}\dots M_{r_1,k}.\end{aligned}$$

Thus we have

$$\begin{aligned}&\|(A_n + B_n)\dots(A_{k+1} + B_{k+1})\|\\ &\le \sum_{m=0}^{n-k}\sum_{k<r_1<\cdots<r_m\le n} Ce^{n\varepsilon}e^{(r_m-n)\gamma}\xi_{r_m}e^{(r_{m-1}-r_m+1)\gamma}\dots\xi_{r_1}e^{(k-r_1+1)\gamma}\\ &= Ce^{n\varepsilon}(e^{-\gamma} + \xi_n)\dots(e^{-\gamma} + \xi_{k+1})\\ &\le Ce^{n\varepsilon}e^{(k-n)\gamma}\prod_{r=1}^{+\infty}(1 + e^{\gamma}\xi_r).\end{aligned}$$

When ε is small enough the infinite product converges, since $\|B_n\|$ tends to 0 exponentially fast. This proves that the sequence $(A_n + B_n)$ is exponentially stable. □

THEOREM 5.6. *We suppose that Hypothesis* (H) *holds and the system is weakly stabilizable and weakly detectable almost surely. Then, for any* $P \in \mathfrak{P}$, *the sequence* $\{A_n - K_nC_n, n \in \mathbb{N}\}$ *is almost surely exponentially stable, where* $K_n, n \in \mathbb{N}$, *are the gain matrices associated with the Riccati equation* (4) *for which* $P_0 = P$.

Proof. Riccati's equation can be written as

$$P_{n+1} - B_nB_n^* = (A_n - K_nC_n)P_nA_n^*, \qquad K_n = (A_n - K_nC_n)P_nC_n^*.$$

Therefore, $K_n = (P_{n+1} - B_nB_n^*)A_n^{*-1}C_n^*$. This yields that, if $\bar{K}_n$ is the gain matrix associated with the stationary solution $\bar{P}_n$, then

$$\|(A_n - \bar{K}_nC_n) - (A_n - K_nC_n)\| \le \|P_{n+1} - \bar{P}_{n+1}\|\|A_n^{*-1}\|\|C_n^*C_n\|.$$

Thus, it follows from Theorem 5.3 and Lemma 2.3 that

$$\limsup_{n\to+\infty}\frac{1}{n}\log\|(A_n - \bar{K}_nC_n) - (A_n - K_nC_n)\| < 0.$$

The proposition is now a consequence of Lemma 5.5 since we have seen in Theorem 5.1 that the sequence $(A_n - \bar{K}_n C_n)$ is almost surely exponentially stable. □

Remark 5.7. In [4] we introduced the following strong detectability condition: there is a sequence $\{F_{-n}, n \in \mathbb{N}\}$ of random $d \times q$ matrices such that $\sup_{n>0} \mathbb{E}\|F_{-n}\|^\beta$ is finite for some $\beta > 0$ and

$$\limsup_{n\to+\infty}\{\mathbb{E}\|(A_{-1}-F_{-1}C_{-1})(A_{-2}-F_{-2}C_{-2})\dots(A_{-n}-F_{-n}C_{-n})\|^\beta\}^{\frac{1}{n}} < 1.$$

There we gave an easy direct proof of the fact that, under this condition, if for some $\alpha > 0$, $\|A_n\|^\alpha$, $\|B_n\|^\alpha$, $\|C_n\|^\alpha$ are integrable and the system is almost surely exponentially stabilizable, then the results of this section hold true. Moreover, $\mathbb{E}\|\bar{P}_n\|^r$ is finite when $r = \min(\alpha, \beta)/4$.

6. Application to control. Let us first provide an efficient observer for systems with stationary coefficients. We consider the linear control system

$$X_{n+1} = A_n X_n + V_n, \qquad Y_n = C_n X_n.$$

We suppose that $\{(A_n, C_n), n \in \mathbb{Z}\}$, is a stationary ergodic process. The sequence (V_n) is arbitrary. These parameters are known at time n. The question is whether it is possible to approach the X_n's if only the Y_n's are available. An observer is a linear system of the form

$$Z_{n+1} = D_n Z_n + F_n Y_n + W_n$$

such that, for any initial condition, $\|X_n - Z_n\| \to 0$ as $n \to +\infty$ (see, e.g., Luenberger [13] or, recently, Yaz [19], who studies systems with random parameters). It thus provides an approximation of the unknown states X_n. We consider the Riccati equation (4) in which we replace $B_n B_n^*$ by the identity matrix of order d. For any $P \in \mathfrak{P}$, let $K_n, n \in \mathbb{N}$, be the gain matrices obtained by this Riccati equation when $P_0 = P$. The following proposition provides an explicit observer.

PROPOSITION 6.1. *We suppose that Hypothesis* (H) *holds and the linear system is almost surely weakly detectable. For* $n \in \mathbb{N}$, *let* $D_n = A_n - K_n C_n, F_n = K_n$, *and* $W_n = V_n$. *Then, there exists* $\gamma > 0$ *such that for any initial condition, almost surely,*

$$\limsup_{n\to+\infty} \frac{1}{n} \log \|X_n - Z_n\| < -\gamma.$$

Proof. The proof readily follows from Theorem 5.6, since, with the chosen coefficients, $X_{n+1} - Z_{n+1} = (A_n - K_n C_n)(X_n - Z_n)$. □

Finally, we consider the quadratic control problem. For any $P \in \mathfrak{P}$ and $n \in \mathbb{N}$, we define matrices $P^{(n)}$ by the formula

$$P^{(n)} = (\phi_{-1} \circ \phi_{-2} \circ \cdots \circ \phi_{-n})(P).$$

It is straightforward to adapt the proof of uniqueness in Theorem 5.1 to see that, under the hypotheses of this theorem, almost surely,

$$\lim_{n\to+\infty} P^{(n)} = \bar{P}_0. \tag{16}$$

We define transformations ϕ_n^* of $\mathfrak{P}$ by the formula

$$\phi_n^*(P) = C_n^* C_n + A_n^* P (I + B_n B_n^* P)^{-1} A_n, \qquad P \in \mathfrak{P}. \tag{17}$$

It is well known that

$$\min\left\{X_n^*PX_n + \sum_{k=0}^{n-1}(U_k^*U_k + Y_k^*Y_k); U_0, \ldots, U_{n-1} \in \mathbb{R}^p\right\} = X_0^*Q^{(n)}X_0,$$

where $Q^{(n)} = (\phi_0^* \circ \phi_1^* \circ \cdots \circ \phi_{n-1}^*)(P)$ (see, e.g., Whittle [16]). Since the mappings (ϕ_{-n}^*) are dual to the mappings (ϕ_n), it follows from (16) applied to the dual system that $Q^{(n)}$ converges almost surely to a random matrix $\bar{Q}$ that does not depend on P as n tends to $+\infty$. In the same way it can be seen that the optimal controls converge almost surely. They depend on the future and can usually be directly implemented only for some deterministic systems, such as almost periodic ones.

Acknowledgments. We thank the referees for suggestions that improved the exposition of this paper. We are also grateful to Yves Derriennic for bringing reference [18] to our attention.

REFERENCES

[1] B. D. O. Anderson and J. B. Moore, *Optimal Filtering*, Prentice–Hall, Englewood Cliffs, NJ, 1979.

[2] ———, *Detectability and stabilizability of time-varying discrete time linear systems*, SIAM J. Control Optim., 19 (1981), pp. 20–32.

[3] P. Bougerol, *Filtre de Kalman Bucy et exposants de Lyapounov*, in Lyapunov Exponents, Proceedings, Oberwolfach 1990, L. Arnold, H. Crauel, and J. P. Eckmannn, eds., Lecture Notes in Math. 1486, Springer-Verlag, New York, Berlin, Heidelberg, 1991, pp. 112–122.

[4] ———, *Some results on the filtering Riccati equation with random parameters*, in Applied Stochastic Analysis, I. Karatzas and D. Ocone, eds., Lecture Notes in Control and Inform. Sci. 177, Springer-Verlag, New York, Berlin, Heidelberg, 1992, pp. 30–37.

[5] ———, *Kalman filtering with random coefficients and contractions*, SIAM J. Control Optim., 31 (1993), pp. 942–959.

[6] P. Bougerol and N. Picard, *Strict stationarity of generalized autoregressive processes*, Ann. Probab., 20 (1992), pp. 1714–1730.

[7] H. F. Chen, P. R. Kumar, and J. H. van Schuppen, *On the Kalman filtering for conditionally Gaussian systems with random matrices*, Systems Control Lett., 13 (1989), pp. 397–404.

[8] W. L. De Koning, *Optimal estimation of linear discrete-time systems with stochastic parameters*, Automatica J. IFAC, 20 (1984), pp. 113–115.

[9] G. De Nicolao, *On the time-varying Riccati difference equation of optimal filtering*, SIAM J. Control Optim., 30 (1992), pp. 1251–1269.

[10] A. H. Jazwinski, *Stochastic Processes and Filtering Theory*, Academic Press, New York, 1970.

[11] R. E. Kalman, *New methods in Wiener filtering theory*, in Proc. 1st Sympos. on Engr. Appl. of Random Function Theory and Probability, J. Bogdanoff and F. Kozin, eds., John Wiley, New York, 1963, pp. 270–388.

[12] J. F. C. Kingman, *Subadditive ergodic theory*, Ann. Probab., 1 (1973), pp. 883–909.

[13] D. G. Luenberger, *Observers for multivariable systems*, IEEE Trans. Automat. Control, AC-11 (1966), pp. 190–199.

[14] D. L. Snyder and P. M. Fishman, *How to track a swarm of fireflies by observing their flashes*, IEEE, Trans. Inform. Theory, IT-21 (1975), pp. 692–695.

[15] M. C. Viano, *Random iterations and Kalman filtering*, Theory Probab. Appl., 35 (1990), pp. 737–745.

[16] P. Whittle, *Optimization Over Time*, Vol. 1, John Wiley, New York, 1982.

[17] W. M. Wonham, *Random differential equations in control theory*, in Probabilistic Methods in Applied Mathematics, A. T. Bharucha-Reid, ed., Academic Press, New York, 1970.

[18] J. Wos, *Approximate convergence in ergodic theory*, Proc. London Math. Soc. (3), 53 (1986), pp. 65–84.

[19] E. Yaz, *On the almost sure and mean-square exponential convergence of some stochastic observers*, IEEE Trans. Automat. Control, AC-35 (1991), pp. 935–936.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 33, No. 3, pp. 718–724, May 1995

© 1995 Society for Industrial and Applied Mathematics
004

ON EXTREMAL SOLUTIONS OF CONTROLLED NONLINEAR FILTERING EQUATIONS*

VIVEK S. BORKAR† AND SUNIL KUMAR‡

Abstract. Controlled nonlinear filtering equations with "relaxed controls" arising out of the separated control problem for partially observed diffusions are considered. An equivalence relation is defined on the attainable laws of the joint state and control process by identifying any two of the latter when their one-dimensional marginals agree almost everywhere. Extreme points of the set of such equivalence classes (for fixed initial laws) are shown to correspond to Markov processes.

Key words. separated control problem, nonlinear filtering, partially observed diffusions, Markov controls, extremal measures

AMS subject classifications. Primary, 93E20; Secondary, 93E11, 60H16

1. Introduction. Consider a controlled, possibly degenerate diffusion controlled through its drift by a nonanticipative relaxed control with the initial law fixed. The set of all possible laws of the joint state and control process can be shown to be compact in the Prohorov topology. Identifying two elements of this set if their one-dimensional marginals agree a.e., the set of corresponding equivalence classes is convex compact in the quotient topology. The extremal elements of the latter were shown to correspond to Markov solutions in [2]. The aim of this paper is to extend this result to controlled nonlinear filtering equations arising out of the separated control problem for a controlled nondegenerate diffusion with partial observations [3, Chap. V], [6]. We begin with some notation.

For a compact metric "control" space V, let

(i) $m(.,.) = [m_1(.,.),\dots,m_d(.,.)]^T : \mathcal{R}^d \times U \to \mathcal{R}^d$, where $U = \mathbf{P}(V)$ (here and later, $\mathbf{P}(S)$ for a Polish space S will be the space of probability measures on S with the Prohorov topology) and is of the form

$$m_i(x,u) = \int_V \bar{m}_i(x,y)u(dy), \qquad 1 \le i \le d, \quad x \in \mathcal{R}^d, \quad u \in U$$

for $\bar{m}_i \in C_b(\mathcal{R}^d \times V)$, which are Lipschitz in the first argument uniformly with respect to the second (this is the relaxed control paradigm);

(ii) $\sigma(.) = [[\sigma_{ij}(.)]]_{i,j=1,\dots,d} : \mathcal{R}^d \to \mathcal{R}^{d\times d}$ is bounded Lipschitz with the least eigenvalue of $\sigma\sigma^T(.)$ bounded uniformly away from zero;

(iii) $l(.) = [(l_1(.),\dots,l_m(.)]^T : \mathcal{R}^d \to \mathcal{R}^m$ is bounded and twice continuously differentiable with bounded first and second derivatives.

For $\mu \in \mathbf{P}(\mathcal{R}^d)$, let $\mu(f) = \int f\,d\mu$ for $f \in C_b(\mathcal{R}^d)$. Also, for$f \in C_0(\mathcal{R}^d) = \{$bounded continuous functions $\mathcal{R}^d \to \mathcal{R}$ with bounded continuous first and second partial derivatives, all vanishing at infinity$\}$, let

$$L_u f(x) = \sum_i m_i(x,u)\frac{\partial f}{\partial x_i} + \frac{1}{2}\sum_{i,j,k}\sigma_{ik}(x)\sigma_{jk}(x)\frac{\partial^2 f}{\partial x_i \partial x_j}$$

* Received by the editors August 10, 1992; accepted for publication (in revised form) January 12, 1994.

† Department of Electrical Engineering, Indian Institute of Science, Bangalore 560012, India.

‡ Department of Electrical and Computer Engineering, and the Coordinated Science Laboratory, University of Illinois, 1308 W. Main Street, Urbana, Illinois 61801.

for $x = [x_1, \ldots, x_d]^T \in \mathcal{R}^d$. Then the controlled nonlinear filtering equation describes a $\mathbf{P}(\mathcal{R}^d)$-valued process $\pi(.)$ by

$$(1)\quad \pi(t)(f) = \pi_0(f) + \int_0^t \pi(s)(L_{u(s)}f)ds + \int_0^t \langle \pi(s)(lf) - \pi(s)(f)\pi(s)(l),\, d\hat{Y}(s)\rangle,$$

where $\pi(s)(lf)$, $\pi(s)(f)$, and $\pi(s)(l)$ imply componentwise multiplication and integration. $\hat{Y}(.) = [\hat{Y}_1(.), \ldots, \hat{Y}_m(.)]^T$ is the "innovations process," which is a standard Wiener process, and $u(.)$ is a U-valued measurable "control" process satisfying the wide-sense admissibility conditions of [6]. See [3, §V.1] for a detailed derivation of the nonlinear filtering equation (1). We shall call $u(.)$ a Markov control if $u(t) = v(\pi(t), t)$, $t \geq 0$ for a measurable $v : \mathbf{P}(\mathcal{R}^d) \times \mathcal{R}^+ \to U$.

Remark. Under our hypotheses, the solution $\pi(.)$ of (1) is a.s. unique when both $u(.)$ and $\hat{Y}(.)$ are prescribed processes [3, §V.1]. When $u(.)$ is Markov, however, neither existence nor uniqueness is guaranteed for a given choice of v above.

Topologize the path space of $u(.)$ as follows. (This topology is the same as that introduced in [7].) For $T > 0$, denote by B_T the closed unit ball of $L_\infty[0, T]$ with the weak topology of $L_2[0, T]$ relativized to it. B_T is compact metrizable and therefore Polish. Let B denote the closed unit ball in $L_\infty[0, \infty)$ with the coarsest topology to render continuous the maps $B \to B_T$, $T \geq 0$, that map $f(.) \in B$ to $f([0, T]) \in B_T$. Let $\{f_i\}$ be a countable dense set in the unit ball of $C(V)$. It is then a convergence-determining class for U. Let $\alpha_i(.) = \int f_i\, du(.)$, $i \geq 1$. Then $\alpha_i(.) \in B$ for each i and $\alpha(.) = [\alpha_1(.), \alpha_2(.), \ldots] \in B^\infty$. B, B^∞ are compact Polish spaces as well. The map $\phi : \mu \in U \to [\int f_1\, d\mu, \int f_2\, d\mu, \ldots] \in [-1, 1]^\infty$ is a homeomorphism between U and $\phi(U)$ by virtue of being a continuous injection with a compact domain. We identify $u(.)$, $\alpha(.)$ via this homeomorphism and use the notation to denote either, depending on the context. Similarly, $u([0, t])$ may be viewed as a B_t^∞-valued random variable for $t \geq 0$.

The path space of $\pi(.)$ is $\hat{C} = C([0, \infty); \mathbf{P}(\mathcal{R}^d))$. Let $\hat{C}_T = C([0, T]; \mathbf{P}(\mathcal{R}^d))$. Let $\mathcal{L}(\ldots)$ denote the "law of...." Define $\Gamma'_\eta = \{\mathcal{L}(\pi(.), u(.)) | u(.)$ is wide-sense admissible, $\mathcal{L}(\pi(0)) = \eta\}$, viewed as a subset of $\mathbf{P}(\hat{C} \times B^\infty)$.

LEMMA 1. *Γ'_η is compact convex.*

The proof of Lemma 1 is postponed to the end of this section. Define an equivalence relation on Γ'_η as follows: $\mathcal{L}(\pi(.), u(.))$, $\mathcal{L}(\pi'(.), u'(.)) \in \Gamma'_\eta$ (or, equivalently, $(\pi(.), u(.))$, $(\pi'(.), u'(.))$ themselves) are said to be marginally equivalent if $\mathcal{L}(\pi(.), u(.)) = \mathcal{L}(\pi'(.), u'(.))$ a.e. It is easy to see that this is indeed an equivalence relation. The corresponding equivalence classes are said to be marginal classes. Let Γ_η denote the set of marginal classes with the quotient topology inherited from Γ'_η. Then it is compact convex as well. Let $\langle \pi(.), u(.)\rangle^\sim$ (or equivalently $\langle \mathcal{L}(\pi(.), u(.))\rangle^\sim$) denote the marginal class containing $(\pi(.), u(.))$ (resp., $\mathcal{L}(\pi(.), u(.))$). Our main result is the following theorem.

THEOREM 1. *Every representative of an extremal element of* Γ_η *is a Markov process.*

Theorem 1 will be proved in the next section. The rest of this section is devoted to a proof of Lemma 1 and another result which will be used later. Let $C^* = C([0, \infty); \mathcal{R}^m)$, $C_T^* = C([0, T]; \mathcal{R}^m)$.

Proof of Lemma 1. It suffices to prove that $\tilde{\Gamma}_\eta = \{\mathcal{L}(\pi(.), \hat{Y}(.), u(.)) | u(.)$ is wide-sense admissible and $\mathcal{L}(\pi(0)) = \eta\} \subset \mathbf{P}(\hat{C} \times C^* \times B^\infty)$ is convex compact. Let $\mathcal{F}_t$, $t \geq 0$, be the natural filtration of $(\pi(.), \hat{Y}(.), u(.))$. For $f \in C_o^2(\mathcal{R}^d)$, define

$$M(t, f) = \pi(t)(f) - \int_0^t \pi(s)(L_{u(s)}f)ds, \qquad t \geq 0.$$

To prove the convexity of $\tilde{\Gamma}_\eta$, observe that (1) is equivalent (up to a change of probability space) to the following "martingale formulation": for $h \in C_b(\mathbf{P}(\mathcal{R}^d))$,

$$E[h(\pi(0))] = \int h\, d\eta, \tag{2}$$

and the following are $\{\mathcal{F}_t\}$-martingales:

$$M(t, f),$$

$$M(t,f)\hat{Y}_i(t) - \int_0^t (\pi(s)(fl_i) - \pi(s)(f)\pi(s)(l_i))ds, \qquad 1 \leq i \leq m,$$

$$(M(t,f) - \pi_0(f))^2 - \int_0^t \|\pi(s)(fl) - \pi(s)(f)\pi(s)(l)\|^2\, ds$$

(this equivalence can be proved as in [3, pp. 116–118]; see also [4, p. 1051]). The latter is in turn equivalent to the following: for $t \geq s \geq 0$, $g \in C_b(\hat{C}_s \times C_s^* \times B_s^\infty)$,

$$E[(Z(t) - Z(s))g(\pi([0,s]), \hat{Y}([0,s]), u([0,s]))] = 0, \tag{3}$$

where $Z(.)$ is any of the above processes claimed to be martingale. (Thus (3) is a family of equations and not a single equation.) If (2), (3) hold under two elements of $\mathbf{P}(\hat{C} \times C^* \times B^\infty)$, they do so under any convex combination thereof. The convexity of $\tilde{\Gamma}_\eta$ follows. Equations (2), (3) are preserved under convergence in $\mathbf{P}(\hat{C} \times C^* \times B^\infty)$ and hence $\tilde{\Gamma}_\eta$ is closed. Thus, to prove its compactness, it suffices to prove that it is tight. Since $\mathcal{L}(\hat{Y}(.))$ is constant (i.e., the Wiener measure) and $\mathbf{P}(B^\infty)$ is compact, we only need to show the tightness of $\{\mathcal{L}(\pi(.))|u(.)$ is wide-sense admissible, $\mathcal{L}(\pi(0)) = \eta\}$. This is proved as in [3, Lem. 3.7, pp. 128–129] (see also [1], [4]). □

Note that the well-posedness of (1) for prescribed $(\hat{Y}(.), u(.))$ (see [3, §V.1]) implies that $\pi(.)$ is adapted to the natural filtration of $(\hat{Y}(.), u(.))$ and, therefore, $\{\mathcal{F}_t\}$ is in fact the natural filtration of $(\hat{Y}(.), u(.))$. The last result of this section is a technical lemma needed later. This is quite standard; see, for example, [4, p. 1043].

LEMMA 2. *For $(\pi(.), \hat{Y}(.), u(.))$ as in* (1), *$t \geq 0$, and $\mathcal{G}_t$ a sub-σ-field of $\mathcal{F}_t$ containing $\sigma(\pi(t))$, the regular conditional law of $(\pi(t+.), \hat{Y}(t+.) - \hat{Y}(t), u(t+.))$ is a.s. the law of a triplet $(\pi'(.), \hat{Y}'(.), u'(.))$ satisfying* (1) *with $\pi'(0) = \pi(t)$ and $u'(.)$ wide-sense admissible.*

2. Proof of Theorem 1. We shall prove this theorem through a sequence of lemmas. Let $\{h_i\}$ be a countable subset of $C_b(\mathbf{P}(\mathcal{R}^d) \times U)$ that separates points of $\mathbf{P}(\mathbf{P}(\mathcal{R}^d) \times U)$. For $i \geq 1$, $\alpha \in (0, \infty)$, let $F_{\alpha i} : \Gamma'_\eta \to \mathcal{R}$, $\eta \in \mathbf{P}(\mathbf{P}(\mathcal{R}^d))$ be the map

$$\mathcal{L}(\pi(.), u(.)) \in \Gamma'_\eta \to E\left[\int_0^\infty e^{-\alpha t}\, h_i(\pi(t), u(t))dt\right].$$

This map is constant on marginal classes and therefore can be viewed as a map $\Gamma_\eta \to \mathcal{R}$.

LEMMA 3. *If $\mu_1, \mu_2 \in \Gamma_\eta$ satisfy $F_{\alpha i}(\mu_1) = F_{\alpha i}(\mu_2)$ for $i \geq 1$, and rational $\alpha \in (0, \infty)$, then $\mu_1 = \mu_2$.*

This follows easily from the injectivity of the Laplace transform on $\mathcal{R}^+$ and our choice of $\{h_i\}$. Let δ_π denote the Dirac measure at $\pi \in \mathbf{P}(\mathcal{R}^d)$. For a given $\mathcal{L}(\pi(.), u(.)) \in \Gamma'_\eta$, let $p(\pi, dy)$ denote a representative of the regular conditional law of $(\pi(.), u(.))$ given $\pi(0) = \pi$. From Lemma 2, one may suppose that $p(\pi, dy) \in \Gamma_{\delta_\pi}$ for each π.

LEMMA 4. *Suppose that for π in a set of strictly positive η-measures, $\langle p(\pi, dy)\rangle^{\sim}$ is not an extreme point of Γ_{δ_π}. Then there exists relatively compact $A \subset \boldsymbol{P}(\mathcal{R}^d)$ with $\eta(A) > 0$, $i \geq 1$, α rational in $(0,1)$, $\varepsilon > 0$, and for any $\pi \in A$, there exists $\nu_{1\pi}, \nu_{2\pi} \in \Gamma_{\delta_\pi}$ for which*

$$\langle p(\pi, dy)\rangle^{\sim} = \frac{(\nu_{1\pi} + \nu_{2\pi})}{2} \tag{4}$$

and

$$|F_{\alpha i}(\nu_{1\pi}) - F_{\alpha i}(\nu_{2\pi})| \geq \varepsilon. \tag{5}$$

Proof. Let $F : \Gamma_{\delta_\pi} \times \Gamma_{\delta_\pi} \to \Gamma_{\delta_\pi}$ denote the map $(\nu_1, \nu_2) \to (\nu_1+\nu_2)/2$. Let $\bar{A}(i, n, \alpha) = \{(\nu_1, \nu_2) \in \Gamma_{\delta_\pi} \times \Gamma_{\delta_\pi} \mid |F_{\alpha i}(\nu_1) - F_{\alpha i}(\nu_2)| \geq 1/n\}$,

$$\tilde{A} = \bigcup_{i,n,\alpha} \bar{A}(i, n, \alpha).$$

Suppose $\eta(\{\pi \mid \langle p(\pi, dy)\rangle^{\sim} \in F(\tilde{A})\}) = 0$. Then for π-a.s. x, the following holds: for all $i, n \geq 1$, and all rational $\alpha > 0$,

$$|F_{\alpha i}(\nu_1) - F_{\alpha i}(\nu_2)| \leq \frac{1}{n},$$

whenever $\nu_1, \nu_2 \in \Gamma_{\delta_\pi}$ satisfy $\langle p(\pi, dy)\rangle^{\sim} = (\nu_1 + \nu_2)/2$. By Lemma 3, $\nu_1 = \nu_2$, which contradicts the hypothesis that $\langle p(\pi, dy)\rangle^{\sim}$ is not an extreme point of Γ_{δ_π}. Therefore $\eta(\{\pi|\langle p(\pi, dy)\rangle^{\sim} \in F(\tilde{A})\}) > 0$. Hence for some $i, n \geq 1$, rational $\alpha > 0$,

$$\eta(\{\pi|\langle p(\pi, dy)\rangle^{\sim} \in F(\bar{A}(i, n, \alpha))\}) > 0.$$

Clearly, $\bar{A}(i, n, \alpha)$ is closed. Let $\hat{A} = \{\pi|\langle p(\pi, dy)\rangle^{\sim} \in F(\bar{A}(i, n, \alpha))\}$. Since η is a probability measure on a Polish space $\mathbf{P}(\mathcal{R}^d)$, $\eta(\cup_n K_n) = 1$ for some compact $K_n \in \mathbf{P}(\mathcal{R}^d)$. By replacing $\hat{A}$ with $A =$ its intersection with an appropriate finite union of K_n's, we may suppose that A is compact with $\eta(A) > 0$. This A satisfies the claim. □

Let $\Delta' \subset \mathbf{P}(\hat{C} \times B^\infty)$ be defined by

$$\Delta' = \overline{\bigcup_{\pi \in A} \Gamma'_{\delta_\pi}}.$$

The arguments of the proof of Lemma 1 can be adapated to show that Δ' is compact. We define marginal equivalence on the whole of $\mathbf{P}(\hat{C} \times B^\infty)$ by deeming two elements thereof to be marginally equivalent whenever the $\mathbf{P}(\mathcal{R}^d) \times U$-valued canonical processes have the same one-dimensional marginals a.e. Let $S^{\sim}$ indicate the space of corresponding equivalence classes with the quotient topology and $\Delta \in S^{\sim}$ indicate the set corresponding to Δ'. Then Δ is compact. Define $F : \Delta \times \Delta \to S^{\sim}$ by $(\nu_1, \nu_2) \to (\nu_1 + \nu_2)/2$ and let

$$G' = \{(\nu_1, \nu_2) \in S^{\sim} \times S^{\sim}||F_{\alpha i}(\nu_1) - F_{\alpha i}(\nu_2)| \geq \varepsilon\}$$

for α, i, ε as in (5).

LEMMA 5. *If $\langle \pi(.), u(.)\rangle^{\sim}$ is an extreme point of Γ_η, then for η-a.s. π, $\langle p(\pi, dy)\rangle^{\sim}$ is an extreme point of Γ_{δ_π}.*

Proof. Suppose Lemma 5 does not hold. Let $A, \varepsilon, \alpha, i, F_{\alpha i}$ be as above. For $\pi \in A$, let

$$K_\pi = \{(\nu_1, \nu_2) \in \Gamma_{\delta_\pi} \times \Gamma_{\delta_\pi}|(4), \text{ (5) hold with } (\nu_1, \nu_2) \text{ in place of } (\nu_{1\pi}, \nu_{2\pi})\}.$$

By Lemma 4, $K_\pi \neq \emptyset$. It is also compact because (4), (5) are preserved under convergence in $\Gamma_{\delta_\pi} \times \Gamma_{\delta_\pi}$, which is compact. Let $G \subset \Delta \times \Delta$ be closed and therefore compact. Note that

$$(6) \qquad \{\pi \in A | K_\pi \cap G \neq \emptyset\} = \{\pi \in A | \langle p(\pi, dy)\rangle^\sim \in F(G \cap G')\}.$$

Since G' is closed, $F(G \cap G')$ is compact. The map $\pi \to p(\pi, dy)$ and, therefore, the map $\pi \to \langle p(\pi, dy)\rangle^\sim$ is measurable. Thus the set in (6) is measurable. We conclude that the map $\pi \in A \to K_\pi \subset \Delta \times \Delta$ is measurable and, therefore, weakly measurable in the sense of [8, p. 862], in view of the remarks in paragraph 5 of [8, p. 862]. By Theorem 4.1 of [8, p. 867], there exists a measurable map $\pi \to (\nu'_\pi, \nu''_\pi) : A \to \Delta \times \Delta$ such that (4), (5) hold with ν'_π, ν''_π in place of $\nu_{1\pi}, \nu_{2\pi}$. Define

$$A^+ = \{\pi \in A | F_{\alpha i}(\nu'_\pi) - F_{\alpha i}(\nu''_\pi) \geq \varepsilon\}$$

and A^- analogously with $-\varepsilon, \leq$ in place of $\varepsilon, \geq$. Since $\eta(A) > 0$ and $A = A^+ \cup A^-$, at least one of $\eta(A^+)$, $\eta(A^-)$ is strictly positive. Suppose $\eta(A^+) > 0$. (If not, replace A^+ by A^-.) Define

$$\begin{aligned} \bar{\nu}_\pi &= \nu'_\pi & \pi \in A^+ \\ &= \langle p(\pi, dy)\rangle^\sim & \text{otherwise,} \\ \bar{\bar{\nu}}_\pi &= \quad \nu''_\pi & \pi \in A^+ \\ &= \langle p(\pi, dy)\rangle^\sim & \text{otherwise,} \\ \mu_1(d\pi, dy) &= \eta(d\pi)\bar{\nu}_\pi(dy), \\ \mu_2(d\pi, dy) &= \eta(d\pi)\bar{\bar{\nu}}_\pi(dy). \end{aligned}$$

Since $\eta(A^+) > 0$, $\langle \mu_1\rangle^\sim \neq \langle \mu_2\rangle^\sim$. Clearly, $\langle \pi(.), u(.)\rangle^\sim = (\langle \mu_1\rangle^\sim + \langle \mu_2\rangle^\sim)/2$. Since $\bar{\nu}_\pi, \bar{\bar{\nu}}_\pi \in \Gamma_{\delta_\pi}$ for each π, $\langle \mu_i\rangle^\sim \in \Gamma_\eta$, $i = 1, 2$. Thus $\langle \pi(.), u(.)\rangle^\sim$ is not an extreme point of Γ_η. This contradiction establishes the claim. □

The following is a special case of [4, Thm. 3.2, p. 1044]. Let $(\pi_i(.), u_i(.))$, $i = 1, 2$ be two pairs that satisfy (1).

LEMMA 6. *If* $\mathcal{L}(\pi_1(T), u_1(T)) = \mathcal{L}(\pi_2(0), u_2(0))$ *for some* $T > 0$, *then there exists a* $(\pi(.), u(.))$ *satisfying* (1) *such that*

$$\begin{aligned} \mathcal{L}(\pi(t), u(t)) &= \mathcal{L}(\pi_1(t), u_1(t)) \quad \text{for } t \in [0, T] \quad \text{and} \\ &= \mathcal{L}(\pi_2(t - T), u_2(t - T)) \quad \text{for } t \geq T. \end{aligned}$$

LEMMA 7. *Let* $\langle \pi(.), u(.)\rangle^\sim$ *be an extreme point of* Γ_η *with* $\mathcal{L}(\pi(T)) = \beta$ *for some* $T > 0$. *Then* $\langle \pi(T + .), u(T + .)\rangle^\sim$ *is an extreme point of* Γ_β.

This follows easily from Lemma 6 (cf. [2]). Henceforth, let $\langle \pi(.), u(.)\rangle^\sim$ be an extreme point of Γ_η. Fix $T > 0$. Let $\pi_T(.) = \pi([0, T])$ and $\mu_0 = \mathcal{L}(\pi_T(.)) \in \mathbf{P}(\hat{C}_T)$. Let $f : \hat{C}_T \to \mathbf{P}(\mathcal{R}^d) \times \hat{C}_T$ denote the map $\pi_T(.) \to (\pi(T), \pi_T(.))$. Let $\bar{\mu}$ denote the image of μ_0 under f. Let Q denote the set of measurable maps $\mathbf{P}(\mathcal{R}^d) \to \hat{C}_T$ such that $\phi \in Q$ implies that for all $\nu \in \mathbf{P}(\mathcal{R}^d)$, $\phi(\nu)$ evaluated at T coincides with ν. Let $\nu_0 = \mathcal{L}(\pi(T))$ and $\mathcal{M} \subset \mathbf{P}(\hat{C}_T)$ be the set of probability measures obtainable as the image of ν_0 under some map belonging to Q.

LEMMA 8. $\bar{\mu}$ (*resp.*, μ_0) *is the barycenter of some probability measure supported on*

$$f_*(\mathcal{M}) = \{\mu | \mu \textit{ it is the image of some element of } \mathcal{M} \textit{ under } f\} \quad (\textit{resp.}, \mathcal{M}).$$

Proof. In the setup of Lemma 2.2 in [2], let $S_1 = \mathbf{P}(\mathcal{R}^d)$, $S_2 = \hat{C}_T$, and $\mathcal{G}_1$, $\mathcal{G}_2$ equal their respective Borel σ-fields. By Lemmas 2.2 and 2.4 of [2], it follows that $\bar{\mu}$ is the barycenter of a probability measure ξ on

$$A = \{\mu \in \mathbf{P}(\mathbf{P}(\mathcal{R}^d) \times \hat{C}_T) | \mu(dx, dy) = \nu_0(dx)v(x, dy), \text{ where } v(x, .) \text{ is Dirac for } \nu_0\text{-a.s. } x\}.$$

Let

$$A' = f_*(\mathbf{P}(\hat{C}_T)) = \{\mu \in \mathbf{P}(\mathbf{P}(\mathcal{R}^d)\times\hat{C}_T) | \mu \text{ is the image of some element of } \mathbf{P}(\hat{C}_T) \text{ under } f\}.$$

If $\xi(A') < 1$, $\xi(A'^C) > 0$. This means that

$$\bar{\mu}((\mathbf{P}(\mathcal{R}^d) \times \hat{C}_T)\backslash f(\hat{C}_T)) > 0,$$

which is a contradiction. Hence ξ is supported on A' and therefore on $A \cap A'$, which is easily seen to be identical to $f_*(\mathcal{M})$. This proves the first claim. Since f is a bijection between $\hat{C}_T$ and its image under f, the map f_* that maps elements of $\mathbf{P}(\hat{C}_T)$ to their images under f is a bijection between $\mathbf{P}(\hat{C}_T)$ and $f_*(\mathbf{P}(\hat{C}_T))$. Let ξ_0 be the image of ξ under f_*^{-1}. Then ξ_0 is supported on $\mathcal{M}$ and μ_0 is the barycenter of ξ_0, proving the second claim. □

Let $(x, y) \in \mathbf{P}(\mathcal{R}^d) \times \hat{C}_T \to q((x, y), dz) \in \mathbf{P}(\hat{C} \times B^\infty)$ by any version of the regular conditional law of $(\pi(T + .), u(T + .))$ given $(\pi(T), \pi_T(.))$. Thus the law of $(\pi(T), \pi_T(.), (\pi(T + .), u(T + .)))$ is $\phi \in \mathbf{P}(\mathbf{P}(\mathcal{R}^d) \times \hat{C}_T \times \hat{C} \times B^\infty) = \tilde{Q}$, given by

$$\phi(dx,\, dy,\, dz) = \bar{\mu}(dx,\, dy)q((x, y),\, dz).$$

Let $H \subset \tilde{Q}$ be the set of measures $\mathcal{V}$ of the form

$$\mathcal{V}(dx,\, dy,\, dz) = \nu_0(dx)\delta_{\varphi(x)}(dx)q((x, y), dz)$$

for some $\varphi \in Q$, $S_{\varphi(x)}$ being the Dirac measure at $\varphi(x)$. By the above lemma, ϕ is the barycenter of a probability measure ξ_1 on H.

LEMMA 9. *With ξ_1-probability one, the probability measure $\beta \in \mathbf{P}(\hat{C} \times B^\infty)$ defined by*

$$\int f(z)\beta(dz) = \iint \nu_0(dx)q((x, \varphi(x)), dz)f(z)$$

for $f \in C_b(\mathbf{P}(\hat{C}) \times B^\infty)$ is in Γ'_{ν_0}. Furthermore, $q((x, \varphi(x)), dz)$ can be chosen to be in Γ'_{δ_x} by choosing an appropriate version.

This is a straightforward consequence of Lemma 2. Denote φ, β above as $\varphi_{\mathcal{V}}$, $\beta_{\mathcal{V}}$ to make explicit the $\mathcal{V}$ dependence. Recall that ϕ is the barycenter of a probability measure ξ_1 on H.

LEMMA 10. *For x outside a set of zero ν_0-measure, $\langle q((x, \varphi_{\mathcal{V}}(x)), dz)\rangle^\sim$ is the same for ξ_1-a.s. $\mathcal{V}$.*

Proof. Note that $\bar{\beta} = \mathcal{L}(\pi(T + .), u(T + .))$ given by

$$\int f(z)\bar{\beta}(dz) = \iint \bar{\mu}(dx, dy)q((x, y), dz)f(z)$$

for $f \in C_b(\hat{C} \times B^\infty)$ is an extreme point of Γ_{ν_0} by Lemma 7. Disintegrate $\bar{\beta}$ as $\nu_0(dx)p(x, dz)$, where $p(x, dz)$ is a representative of the regular conditional law of $(\pi(T + .), u(T + .))$ given that $\pi(T) = x$. By Lemma 5, $\langle p(x, dz)\rangle^\sim$ is an extreme point of Γ_{δ_x} for ν_0-a.s. x. Now ϕ is the barycenter of a probability measure ξ_1 on H. Thus $\bar{\beta}$ is the barycenter of a probability measure on $\{\beta_{\mathcal{V}}, \mathcal{V} \in H\}$. By Lemma 2.3 of [2], for ν_0-a.s. x, $p(x, dz)$ is the barycenter of a probability measure on $\{q((x, \varphi_{\mathcal{V}}(x)), dz),\ \mathcal{V} \in H\}$ and, in turn, $\langle p(x, dz)\rangle^\sim$ is the barycenter of a probability measure on $\{\langle q((x, \varphi_{\mathcal{V}}(x)), dz)\rangle^\sim\}$. For x outside a set of zero ν_0-measure, outside of which the foregoing holds and $\langle p(x, dz)\rangle^\sim$ is extremal in Γ_{δ_x}, we must have

$$\langle p(x, dz)\rangle^\sim = \langle q((x, \varphi_{\mathcal{V}}(x)), dz)\rangle^\sim$$

for ξ_1-a.s. $\mathcal{V}$, thus proving the claim. □

Proof of Theorem 1. Fix $t > 0$ and let $\hat{p}(x, dz)$, $\hat{q}((x, y), dz) \in \mathbf{P}(\mathbf{P}(\mathcal{R}^d))$ denote the images of $p(x, dz)$ and $q((x, y), dz)$, resp., under the map $(x(.), y(.)) \in \hat{C} \times B^{\infty} \to x(t) \in \mathbf{P}(\mathcal{R}^d)$. Then the law of $(\pi(T), \pi_T(.), \pi(T + t))$ is

$$\begin{aligned}\mu(dx, dy)\hat{q}((x, y), dz) &= \int \xi_1(d\mathcal{V})\nu_0(dx)\delta_{\varphi_{\mathcal{V}}(x)}(dy)q((x, \varphi_{\mathcal{V}}(x)), dz)\\ &= \nu_0(dx)\eta(x, dy)\hat{p}(x, dz)\end{aligned}$$

by the above lemma, where

$$\eta(x, dy) = \int \xi_1(d\mathcal{V})\delta_{\varphi_{\mathcal{V}}(x)}(dy).$$

Thus $\pi(T + t)$, $\pi_T(.)$ are conditionally independent given $\pi(T)$. Given the arbitrary choice of T, t, the claim follows. □

REFERENCES

[1] V. S. Borkar, *Existence of optimal controls for partially observed diffusions*, Stochastics Stochastics Rep., 13 (1983), pp. 103–142.

[2] ———, *On extremal solutions to stochastic control problems*, Appl. Math. Optim., 24 (1991), pp. 317–330.

[3] ———, *Optimal Control of Diffusion Processes*, in Pitman Research Notes in Mathematics 203, Longman Scientific and Technical, Harlow, U.K., 1989.

[4] N. El Karoui, D. Huu Nguyen and M. Jeanblanc-Picque, *Existence of an optimal Markovian filter for the control under partial observations*, SIAM J. Control Optim., 26 (1988), pp. 1025–1061.

[5] ———, *Martingale measures and partially observable diffusions*, Stochastic Anal. Appl., 9 (1991), pp. 147–176.

[6] W. H. Fleming and E. Pardoux, *Optimal control of partially observed diffusions*, SIAM J. Control Optim., 20 (1982), pp. 261–285.

[7] J. Jacod and J. Memin, *Sur un type de convergence intermédiaire entre la convergence en loi et la convergence en probabilité*, in Seminaire de Probabilités XV, 1979–80, J. Azéma, M. Yor, eds., Lecture Notes in Mathematics 850, Springer-Verlag, Berlin, Heidelberg, 1981, pp. 529–546.

[8] D. Wagner, *Survey of measurable selection theorems*, SIAM J. Control Optim., 15 (1977), pp. 859–903.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 33, No. 3, pp. 725–749, May 1995

© 1995 Society for Industrial and Applied Mathematics
005

STRONG STABILITY IN VARIATIONAL INEQUALITIES*

JIMING LIU†

Abstract. In this paper we consider a generalization of Kojima's strong stability in nonlinear programs to variational inequalities constrained by a system of equations and inequalities. Roughly speaking, strong stability refers to the local existence and uniqueness of a solution of a system under small perturbations. The purpose of the paper is to establish a new and complete characterization for strongly stable generalized Karush–Kuhn–Tucker points and to give a complete characterization for strongly stable stationary solutions under the Mangasarian–Fromovitz constraint qualification.

Key words. strongly stable generalized Karush–Kuhn–Tucker (GKKT) point, strongly stable stationary solution, variational inequality problem, perturbation

AMS subject classifications. 90C33, 49A29, 90C31

1. Introduction.

The notions of strong regularity and strong stability are perhaps the two most important concepts in the stability theory of optimization. The former was introduced by Robinson [22] from an analytic viewpoint as a concept in generalized equations analogous to the classical nonsingularity concept in nonlinear equations, and the latter was proposed by Kojima [8] from a topological point of view attempting to characterize the existence and local uniqueness of the perturbed local solutions of a nonlinear program. These two concepts are equivalent for polyhedrally constrained nonlinear programs [16]. For general nonlinear programs, strong regularity is referred to the nonsingularity of the Karush–Kuhn–Tucker (KKT) points, while the concept of strong stability has more flexibility; a local minimizer with nonunique multipliers, which is not strongly regular, may still have strong stability. These two concepts have been extensively studied in the context of nonlinear programming; see [5], [6], [7], [8], [9], [19], and [24]. Actually, the notion of strong regularity was originally proposed in the framework of generalized equations, or variational inequalities, and a complete characterization for strong regularity was obtained in [22]. In recent years the natural generalizations of strong stability and its related theory to variational inequalities have been addressed in several works [1], [3], [13], [25]. The conceptual aspect of the generalizations is rather straightforward. However, some technical difficulties exist in establishing complete characterizations for strong stability in variational inequalities. Only very recently was a complete characterization for local homeomorphism of normal maps, or equivalently, strong stability of the solutions of the corresponding variational inequalities over polyhedral sets, obtained by Robinson [26] with a very penetrating analysis of the structures of linear normal maps and polyhedral sets; even a short proof of this result is a very challenging task [20], [21], [27]. The major difficulty in this kind of generalization mainly stems from the additional complexity of asymmetric matrices versus symmetric matrices. This difficulty also exists in matrix theory and linear algebra.

The major objectives of this paper are to establish a new and complete characterization of strongly stable generalized Karush–Kuhn–Tucker (GKKT) points and to give the first complete characterization of strongly stable stationary solutions in variational inequalities defined on perturbed sets. To achieve these objectives, some new ideas are needed. We shall first examine the differences between a variational inequality problem and a nonlinear program. It is well known that the Karush–Kuhn–Tucker conditions or stationary conditions of a nonlinear

* Received by the editors November 18, 1992; accepted for publication (in revised form) January 12, 1994. This research was supported by Office of Naval Research contract N00014-89-J-1537.

† Department of Operations Research, George Washington University, Washington, DC 20052 (jiming@seas.gwu.edu).

program can be expressed by a variational inequality. Let us look at the simplest case first. Consider an unconstrained quadratic program,

$$\underset{x}{\text{minimize}}\ \theta(x) = x^{\mathsf{T}}(Ax - b) \quad \text{s.t. } x \in R^n,$$

where A is a symmetric $n \times n$ matrix. The stationary conditions of the above program are equivalent to the following variational inequality:

$$\text{find } x \in R^n \quad \text{such that } (f(x) = (2Ax - b))^{\mathsf{T}}(x' - x) \geq 0 \quad \text{for all } x' \in R^n.$$

If we regard the vector b as the perturbation parameter, then for a particular value, say $b_0 = 0$, the condition which ensures that the above nonlinear program has a unique solution for all b near b_0 is the requirement that A, the Hessian of θ, be positive definite; however, for the above variational inequality, the corresponding condition is that A, the Jacobian matrix of f, is nonsingular. Here the distinction between these two conditions is due to the fact that a solution of a nonlinear program must satisfy some necessary optimality conditions. Now let us examine a more complicated situation:

$$\underset{x}{\text{minimize}}\ \theta(x) = x^{\mathsf{T}}(Ax - b) \quad \text{s.t. } x \in R^n_+.$$

The stationary conditions of this problem can be written down as a linear complementarity problem,

$$\text{find } x \in R^n_+ \quad \text{such that } (2Ax - b)^{\mathsf{T}}(x' - x) \geq 0 \quad \text{for all } x' \in R^n_+.$$

Now, let the vector b be the perturbation parameter and let the unperturbed parameter b_0 be zero. We know that local uniqueness of solutions of the above nonlinear program requires that A be positive definite, and that the corresponding condition for the above linear complementarity problem is that A is a P matrix, i.e., A has positive principal minors. This property is equivalent to the positive definiteness of A in our case since A is symmetric. However, the matrix A is generally asymmetric if the above linear complementarity problem arises in equilibrium problems, and in such cases this property is strictly weaker than the positive definiteness of A. These two examples demonstrate the substantial differences between nonlinear programs and variational inequalities. In the author's opinion, the present stability and sensitivity theory of variational inequalities (see the recent survey paper Kyparisis [13] in this field) follows the stability and sensitivity theory of nonlinear programs too heavily and does not reflect the substantial differences very well.

Now we consider the following more general variational inequality:

$$\text{find } x \in C \quad \text{such that } (Ax - b)^{\mathsf{T}}(x' - x) \geq 0 \quad \text{for all } x' \in C,$$

where C is a convex polyhedral cone, and A is an $n \times n$ matrix. Again, let the vector b be the perturbation parameter and let the unperturbed parameter b_0 be zero. Recently, Robinson [26] showed that the condition that completely characterizes the existence and uniqueness of solutions of the above variational inequality for all $b \in R^n$ is exactly the coherent orientedness condition (Robinson did not restrict C to a cone). To be precise, we shall briefly review the coherent orientedness condition. In what follows, for a $p \times q$ matrix B we shall use B to refer to the matrix B itself or the linear transformation from R^q to R^p represented by the matrix B according to the context. The orientation of a linear transformation B from R^t to R^t is defined to be the sign of the determinant of B, which takes three possible values, -1, 0, and $+1$. Let L be a subspace of R^n and Q be an orthogonal matrix whose columns form a basis for L. The linear transformation represented by $Q^{\mathrm{T}}AQ$ is called the section of A in the subspace L [4]. The coherent orientedness condition is the requirement that the orientations of the sections of A in the subspaces spanned by all faces of C all have a common nonzero sign.

Originally, the coherent orientedness condition was introduced [11], [26] from a piecewise affine map perspective. It is instructive to explain it from a geometric point of view, as we will do here. The advantage of this viewpoint is that it enables us to generalize this important notion to the case where C is not a polyhedral set. In what follows we shall sketch the ideas. Let A be a linear transformation in R^n and let C be a convex cone in R^n. Suppose z is a nonzero vector in C. We say that z is in general positive position with respect to (w.r.t.) a vector set $\{y_1, \dots, y_m\} \subset C$ if there exist $a_i > 0$, $i = 1, \dots, m$, such that

$$z = a_1 y_1 + \cdots + a_m y_m,$$

and the vector set $\{y_1, \dots, y_m\}$ is linearly independent. If z is in general positive position w.r.t. $\{y_1, \dots, y_m\} \subset C$, and z is not in general positive position w.r.t. $\{y_1', \dots, y_{m'}'\} \subset C$ such that $m' > m$, then we say that the vector set $\{y_1, \dots, y_m\}$ is a frame of z on C. Now let $\{y_1, \dots, y_m\}$ be a frame of z on C and let L be the subspace spanned by $\{y_1, \dots, y_m\}$. Suppose $z = a_1 y_1 + \cdots + a_m y_m$. We first assume that $\{y_1, \dots, y_m\}$ is an orthonormal basis of L. The linear transformation A maps z to Az. Consider the rectangular parallelepiped $D = \{t_1 y_1 + \cdots + t_m y_m : 0 \le t_1 \le a_1, \dots, 0 \le t_m \le a_m\}$. It is the image of the rectangle $I = \{(t_1, \dots, t_m)^{\mathrm{T}} : 0 \le t_1 \le a_1, \dots, 0 \le t_m \le a_m\}$ under the matrix $Y = (y_1, \dots, y_m)$. Suppose the linear transformation A maps D onto D'. Then $D' = \{t_1 Ay_1 + \cdots + t_m Ay_m : t = (t_1, \dots, t_m)^{\mathrm{T}} \in I\} = \{AYt : t \in I\}$. Now we project D' on L and obtain the projection $\bar{D}$. Then $\bar{D} = \{(y_1^{\mathrm{T}} AYt, \dots, y_m^{\mathrm{T}} AYt)^{\mathrm{T}} : t \in I\} = \{Y^{\mathrm{T}} AYt : t \in I\}$ in Y space. This means that $\bar{D}$ can be regarded as the image of I under the linear transformation $Y^{\mathrm{T}}AY$. Obviously, it is also a rectangular parallelepiped in L. It is not hard to see that the directed volume of $\bar{D}$ in Y space, denoted by $\mathrm{DV}(\bar{D})$, is equal to $a_1 \times \cdots \times a_m \det(Y^{\mathrm{T}}AY)$, and the directed volume of D in Y space, denoted by $\mathrm{DV}(D)$, is equal to $a_1 \times \cdots \times a_m \det(Y^{\mathrm{T}}Y)$. We define $\mathrm{GAR}(A, Y, z)$, called the general amplification rate of A in the direction z w.r.t. Y, as follows:

$$\mathrm{GAR}(A, Y, z) = \frac{\mathrm{DV}(\bar{D})}{\mathrm{DV}(D)} = \frac{\det(Y^{\mathrm{T}}AY)}{\det(Y^{\mathrm{T}}Y)}. \tag{1.1}$$

The equation above explains the geometric meaning of the determinant of the section of A in the subspace L. Note that although we have assumed that Y is an orthonormal basis of $L = \mathrm{span}(Y)$, the definition is valid even when Y is not orthonormal. To see this, note that there exists an $m \times m$ nonsingular matrix T such that $X = YT$ and X is orthonormal. Substituting $X = YT$ into (1.1) we obtain the same formula. Actually, if two vector sets Y^1 and Y^2 span the same subspace then one has $\mathrm{GAR}(A, Y^1, z) = \mathrm{GAR}(A, Y^2, z)$. This means that if two vector sets Y^1 and Y^2 are frames of z on C, then $\mathrm{GAR}(A, Y^1, z) = \mathrm{GAR}(A, Y^2, z)$ since by the definition of frame, Y^1 and Y^2 span the same subspace. Therefore, for any $z \in C \backslash \{0\}$, we define the general amplification rate of A in the direction z w.r.t. C by

$$\mathrm{GAR}(A, C, z) = \frac{\mathrm{DV}(\bar{D})}{\mathrm{DV}(D)} = \frac{\det(Y^{\mathrm{T}}AY)}{\det(Y^{\mathrm{T}}Y)}, \tag{1.2}$$

where Y is a frame of z on C.

We now introduce the notion of coherent orientedness of a linear transformation on a convex cone in terms of the general amplification rate of the linear transformation w.r.t. this cone. For a polyhedral convex cone C, denote by $\mathrm{rif}(C)$ the set of relative interior points of various faces of C. Then $\mathrm{rif}(C) = C$ if $C \cap (-C) = \{0\}$, i.e., C contains no lines and $\mathrm{rif}(C) = C\backslash\{0\}$ if $C \cap (-C) \neq \{0\}$. So for a convex cone C we define $\mathrm{rif}(C) = (C\backslash\{0\}) \cup (C \cap (-C))$. In stating the definition, we assume that $\mathrm{GAR}(A, C, 0) = 1$ or 0 for any linear transformation A and any convex cone C for the sake of convenience. For simplicity we shall abbreviate the terms "coherently oriented" to "cooriented" and "coherent orientedness" to "coorientedness," etc.

DEFINITION 1.1. *A linear transformation A in R^n is said to be positively (negatively) cooriented on a convex cone C in R^n if for all $z \in \mathrm{rif}(C)$,*

$$\mathrm{GAR}(A, C, z) > (<) 0;$$

A is said to be positively (negatively) semicooriented on C if the above inequality holds for all $z \in \mathrm{rif}(C)$ with the replacement of

$$\mathrm{GAR}(A, C, z) \geq (\leq) 0;$$

A is said to be cooriented (semicooriented) on C if A is either positively cooriented (semicooriented) or negatively cooriented (semicooriented) on C.

In view of (1.2) and the definition of frame, the above definition of coorientedness of a linear transformation on a polyhedral convex cone is equivalent to that of [26]. Evidently, a positive definite (semidefinite) matrix A is positively cooriented (semicooriented) on any convex cone C since we always have

$$\mathrm{GAR}(A, C, z) > (\geq) 0$$

for all $z \in \mathrm{rif}(C)$. The converse is not true; examine the following matrix:

$$A = \begin{bmatrix} 1 & 1 \\ -4 & 1 \end{bmatrix}.$$

It is not difficult to check that A is positively cooriented on R^2_+. However, taking $z = (1, 1)^{\mathrm{T}} \in R^2_+$, we then have $z^{\mathrm{T}} A z = -1 < 0$. Also, we note that A is not positively cooriented on the convex cone $C = \{(x_1, x_2) : x_2 - x_1 \leq 0,\ -x_2 \leq 0\}$, which is a subset of R^2_+. This suggests that the coorientedness of a linear transformation on a convex cone is an incorporating property of the linear transformation and the cone. This property of coorientedness differs substantially from that of positive definiteness. However, it was proved very recently [27] that if the linear transformation can be represented by a symmetric matrix, then the coorientedness of the linear transformation on a polyhedral convex cone containing no lines reduces to positive definiteness.

At this point let us explain our reason for attempting to draw the reader's attention to the notion of coorientedness at the very beginning of the paper. It is perhaps very well known that for nonlinear programs, the combination of the linear independence condition and strong second-order sufficient condition, and the combination of the Mangasarian–Fromovitz constraint qualification and general strong second-order sufficient condition, are necessary and sufficient conditions for strong regularity of KKT points and strong stability of local minimizers, respectively. It is natural to predict that for variational inequality problems a combination of the linear independence condition and a certain coorientedness condition, and a combination of the Mangasarian–Fromovitz constraint qualification and a certain general coorientedness condition, would provide complete characterizations for strongly regular GKKT points and strongly stable stationary solutions, respectively. This is true, as we will see in the following sections.

We organize the rest of the paper in three sections. In §2 we introduce the notions of strong stability in variational inequalities and review the known notation and basic results.

Two results that are useful in understanding strong stability are also given in this section. In §3 we establish a new and complete characterization for strongly stable GKKT points. Other characterizations are discussed. Finally, in §4 we consider strongly stable stationary solutions in the framework of the Mangasarian–Fromovitz constraint qualification and show that the so-called general coorientedness condition completely characterizes the strong stability of stationary solutions.

2. Preliminaries. The variational inequality problems we shall deal with are of the form

$$\text{VI}(f,g) \qquad \text{find } x \in R(g) \quad \text{such that } f(x)^{\mathsf{T}}(x'-x) \geq 0 \quad \text{for all } x' \in R(g),$$

where $g := (g_1, \ldots, g_m)$, $R(g) := \{x \in R^n : g_i(x) \leq 0,\ i \in L_1; g_i(x) = 0,\ i \in L_2\}$, $L_1 := \{1, \ldots, k\}$, $L_2 := \{k+1, \ldots, k+l\}$, $m = k + l$, $f : R^n \to R^n$ is once continuously differentiable, and $g_i : R^n \to R$ $(i = 1, 2, \ldots, m)$ are twice continuously differentiable.

Given a VI(f, g), if x is a local solution to VI(f, g) and an appropriate regularity condition holds at x, then the GKKT conditions or stationary conditions hold at x [13]: there exist multipliers $u \in R^m$ such that

$$\begin{gathered} f(x) + \sum_{i=1}^{m} u_i \nabla g_i(x) = 0, \\ g_i(x) \leq 0, \quad u_i \geq 0, \quad u_i g_i(x) = 0, \quad i \in L_1, \\ g_i(x) = 0, \qquad i \in L_2, \end{gathered} \tag{2.1}$$

where (here and in what follows) the notation ∇ always denotes the derivative with respect to the variable x. It is known that system (2.1) can be expressed in an equivalent form via generalized equations [22]:

$$0 \in \begin{bmatrix} f(x) + \sum_{i=1}^{m} u_i \nabla g_i(x) \\ -g(x) \end{bmatrix} + N_{R^n \times R^k_+ \times R^\ell}(x, u), \tag{2.2}$$

where the notation N denotes the normal cone operator. For a convex set $C \subset R^t$,

$$N_C(x) := \begin{cases} \{y \in R^t \colon y^{\mathsf{T}}(z - x) \leq 0 \text{ for all } z \in C\} & \text{if } x \in C, \\ \emptyset & \text{if } x \notin C. \end{cases}$$

The above generalized equation can be transformed to its corresponding normal equation [26]. To state it, define the normal map

$$F(x,y) := \begin{bmatrix} f(x) + \sum_{i \in L_1} y_i^+ \nabla g_i(x) + \sum_{i \in L_2} y_i \nabla g_i(x) \\ -g_1(x) + y_1^- \\ \vdots \\ -g_k(x) + y_k^- \\ -g_{k+1}(x) \\ \vdots \\ -g_m(x) \end{bmatrix} \tag{2.3}$$

for $(x, y) \in R^n \times R^m$, where for $\mu \in R$,

$$\mu^+ := \max\{\mu, 0\}, \qquad \mu^- := \min\{\mu, 0\}.$$

Then the normal equation of VI(f, g) is the following nonlinear equation:

$$F(x, y) = 0. \tag{2.4}$$

The relation of solutions between (2.2) and (2.4) is rather simple: suppose (x, u) is a solution of (2.2), then (x, y) with $y_i := u + g_i(x)$ $(i = 1, \dots, m)$ solves (2.4); if (x, y) is a solution of (2.4) then (x, u) with

$$u_i := \begin{cases} y_i^+ & \text{if } i \in L_1, \\ y_i & \text{if } i \in L_2, \end{cases}$$

solves (2.2). At this point let us remark that the major advantage of using a normal equation approach is that the normal map is piecewise differentiable provided that the functions involved are differentiable. We shall discuss this in a little more detail later. In this paper we mainly use the normal equation to express the stationary conditions. Therefore, we shall use somewhat different but equivalent terminology. A point $x \in R^n$ satisfying (2.4) with some $y \in R^m$ is said to be a stationary solution of VI(f, g), in symbols, $x \in S(f, g)$, and in such a case the pair (x, y) is said to be a GKKT point of VI(f, g), in symbols, $(x, y) \in$ GKKT(f, g). Let x be a stationary solution of VI(f, g). Denote the set of its associated multipliers by $M(x, f, g) := \{y \in R^m : (x, y) \in \text{GKKT}(f, g)\}$ and the set of extreme points of $M(x, f, g)$ by $E(x, f, g)$. Note that the solution conditions that ensure a stationary solution x to be a local solution of VI(f, g) are generally stronger than those in nonlinear programs. For a convex-constrained variational inequality CCVI(f, g) (i.e., each inequality constraint g_i, $i \in L_1$, is convex in x, and each equation constraint g_i, $i \in L_2$, is affine in x) a stationary solution x is then a local solution of CCVI(f, g).

For convenience, let

$$L_D(x, y) := f(x) + \sum_{i \in L_1} y_i^+ \nabla g_i(x) + \sum_{i \in L_2} y_i \nabla g_i(x).$$

For each $(x, y) \in R^n \times R^m$, some characteristic index sets are defined by

$$\begin{aligned} I_0(x, g) &:= L_2 \cup \{i \in L_1 : g_i(x) = 0\}, \\ I_+(y) &:= \{i \in L_1 : y_i > 0\}, \\ I_N(y) &:= L_2 \cup I_+(y). \end{aligned}$$

Consider the following classes of perturbations of VI(f, g):

$\mathcal{P} := \{(\Delta f, \Delta g) = (\Delta f, \Delta g_1, \dots, \Delta g_m)$: $\Delta f : R^n \to R^n$ is once continuously differentiable,

$\Delta g_i : R^n \to R$, $i \in \{1, \dots, m\}$ are twice continuously differentiable$\}$,

equipped with the family of seminorms

$$\begin{aligned} &\text{norm}(\Delta f, \Delta g, U) \\ &\quad := \sup_{1 \le i \le m} \sup_{x \in U} \max\{\|\Delta f(x)\|, \ \|\nabla \Delta f(x)\|, \ |\Delta g_i(x)|, \ \|\nabla \Delta g_i(x)\|, \ \|\nabla^2 \Delta g_i(x)\|\}, \end{aligned}$$

where U is a subset of R^n, and a subclass $\mathcal{P}^*$ of $\mathcal{P}$ defined by

$$\begin{aligned}\mathcal{P}^* := \{(\Delta f, \Delta g) \in \mathcal{P} : \Delta f(x) = Dx + c,\ \Delta g_i(x) = d_i\ (1 \le i \le m)\\ \text{for some } n \times n \text{ matrix } D,\ \text{some vector } c \in R^n,\ \text{and } d = (d_1, \dots, d_m) \in R^m\}.\end{aligned}$$

For each perturbation $(\Delta f, \Delta g) \in \mathcal{P}$ of the problem VI(f, g), we define the map $\Delta F : R^n \times R^m \to R^{n+m}$ as follows:

$$(2.5) \qquad \Delta F(x,y) := \begin{bmatrix} \Delta f(x) + \sum_{i \in L_1} y_i^+ \nabla \Delta g_i(x) + \sum_{i \in L_2} y_i \nabla \Delta g_i(x) \\ -\Delta g_1(x) \\ \vdots \\ -\Delta g_k(x) \\ -\Delta g_{k+1}(x) \\ \vdots \\ -\Delta g_m(x) \end{bmatrix}.$$

Then the normal equation for the perturbed problem VI$(f + \Delta f, g + \Delta g)$ can be written as

$$F(x, y) + \Delta F(x, y) = 0.$$

We now generalize the notion of Kojima's strong stability to variational inequalities. In what follows, we use $B(z, \delta)$ to denote $\{z' \in R^t : \|z' - z\| \le \delta\}$ for any positive number δ and $z \in R^t$.

DEFINITION 2.1. *Let $\mathcal{P}'$ be a subclass of $\mathcal{P}$. We say that a stationary solution x^* to* VI(f, g) *is strongly stable w.r.t. $\mathcal{P}'$ if for some $r > 0$ and each $\varepsilon \in (0, r]$ there exists a $\delta = \delta(\varepsilon)$ such that whenever $(\Delta f, \Delta g) \in \mathcal{P}'$ and* norm$(\Delta f, \Delta g, B(x^*, r)) \le \delta$, *$B(x^*, \varepsilon)$ contains a stationary solution $x(\Delta f, \Delta g)$ to* VI $(f + \Delta f, g + \Delta g)$, *which is unique in $B(x^*, r)$. Similarly, a GKKT point (x^*, y^*) of* VI(f, g) *is said to be strongly stable w.r.t. $\mathcal{P}'$ if for some $r > 0$ and each $\varepsilon \in (0, r]$ there exists a $\delta = \delta(\varepsilon) > 0$ such that whenever $(\Delta f, \Delta g) \in \mathcal{P}'$ and* norm$(\Delta f, \Delta g, B(x^*, r)) \le \delta$, *$B((x^*, y^*), \varepsilon)$ contains a GKKT point $(x(\Delta f, \Delta g), y(\Delta f, \Delta g))$ to* VI$(f + \Delta f, g + \Delta g)$, *which is unique in $B((x^*, y^*), r)$.*

To explain this important concept, we shall make several remarks in the following sections. In the rest of the paper, when we say that x^* is strongly stable without specifying the perturbation class, we simply mean that the perturbation class is $\mathcal{P}$.

Remark 2.2. From Definition 2.1, if x^* is a strongly stable stationary solution (w.r.t. $\mathcal{P}'$), then there are some $r, s > 0$, and some mapping $x(\cdot)$ defined on $W(0, s) := \{(\Delta f, \Delta g) \in \mathcal{P}' : \text{norm}(\Delta f, \Delta g, B(x^*, r)) \le s\}$ such that

(1) $x(\cdot)$ is continuous at $(\Delta f, \Delta g) = 0$ (w.r.t. the above norm) with $x(0) = x^*$;

(2) $S(f+\Delta f, g+\Delta g) \cap B(x^*, r) = \{x(\Delta f, \Delta g)\}$ for all $(\Delta f, \Delta g)$ belonging to $W(0, s)$.

Properties (1) and (2) may be explained informally as follows: if a stationary solution is strongly stable, then the slightly perturbed stationary solution is locally unique and continuous as a function of the perturbation.

Remark 2.3. The above definition about strong stability of GKKT points is different from Kojima's definition when restricted to nonlinear programs. According to Kojima [8], a KKT point (x^*, y^*) of a nonlinear program NLP(f, g) with objective function f and feasible set $R(g)$ is said to be strongly stable (w.r.t. $\mathcal{P}'$) if and only if x^* is a strongly stable stationary solution to NLP(f, g) (w.r.t. $\mathcal{P}'$).

Remark 2.4. (1) If $\mathcal{P}'$ and $\mathcal{P}''$ are subclasses of $\mathcal{P}$ such that $\mathcal{P}' \subset \mathcal{P}''$, then strong stability w.r.t. $\mathcal{P}''$ implies strong stability w.r.t. $\mathcal{P}'$. (2) Strong stability of GKKT points implies local uniqueness of the perturbed stationary solution as well as local uniqueness of the perturbed multipliers. However, a strongly stable stationary solution may have nonunique associated multipliers.

We now turn to discuss an important property, called piecewise differentiability, of the normal map $F(\cdot\,,\cdot)$, which plays a key role in the study of normal maps. We first define the following matrices: Given a GKKT point $(x^*\,, y^*)$, define

$$B = (\nabla g_i(x^*)), \quad i \in I_N(y^*); \qquad C(J) = (\nabla g_i(x^*)), \quad i \in J \quad \text{for } J \subset I_0(x^*, g)\backslash I_N(y^*);$$

and

$$M(J) = \begin{bmatrix} \nabla L_D & B & C(J) \\ -B^{\mathsf{T}} & 0 & 0 \\ -C(J)^{\mathsf{T}} & 0 & 0 \end{bmatrix}$$

for $J \subset I_0(x^*, g)\backslash I_N(y^*)$, where $\nabla L_D := \nabla L_D(x^*, y^*)$. For the given GKKT point (x^*, y^*), we can induce a subdivision K of $R^n \times R^m$ as follows: for any $J \subset I_0(x^*, g)\backslash I_N(y^*)$ define

$$\tau(J) = R^n \times \{y \in R^m : y_i \geq 0,\ i \in J \cup I_N(y^*);\ y_i \leq 0,\ i \in \{1, \ldots, m\}\backslash(J \cup I_N(y^*))\}.$$

Let

$$K = \{\tau(J) : J \subset I_0(x^*, g)\backslash I_N(y^*)\}.$$

Then K is a subdivision of R^{n+m} [8], and $F(\cdot\,,\cdot)$ is continuously differentiable in each piece $\tau(J) \cap B((x^*, y^*), \delta)$, where δ is a positive number, and the corresponding Jacobian matrix $DF(x, y; \tau(J))$ at a point $(x, y) \in \tau(J) \cap B((x^*, y^*), \delta)$ is

$$\bar{M}(J) = \begin{bmatrix} \nabla L_D & B & C(J) & 0 \\ -B^{\mathsf{T}} & 0 & 0 & 0 \\ -C(J)^{\mathsf{T}} & 0 & 0 & 0 \\ -C(\bar{J}) & 0 & 0 & I \end{bmatrix},$$

where $B = (\nabla g_i(x)),\ i \in I_N(y^*)$; $C(J) = (\nabla g_i(x)),\ i \in J$, $C(\bar{J}) = (\nabla g_i(x)), i \in \bar{J}$, ∇L_D is evaluated at (x, y), and $\bar{J} = (I_0(x^*, g)\backslash I_N(y^*))\backslash J$. It is easy to see that $\det \bar{M}(J) = \det M(J)$. The signs of the determinants of $F(x^*, y^*; \tau(J))$ are closely related to the strong stability at (x^*, y^*) and will be demonstrated later. For convenience we shall assume that the sign of the determinant of the matrix (0) is plus one.

According to the subdivision K, we can naturally induce a piecewise affine map LF about the point (x^*, y^*) as follows:

$$LF(x, y) := F(x^*, y^*) + DF(x^*, y^*; \sigma)((x, y) - (x^*, y^*))$$

for every $(x, y) \in \sigma$ and $\sigma \in K$, which is the linearization of the normal map $F(\cdot\,,\cdot)$ about the point (x^*, y^*).

Several commonly used regularity conditions which may hold at a stationary solution x^* to VI(f, g) are as follows:

(a) The linear independence condition (LI) holds at x^* if the set

$$\{\nabla g_i(x^*) : i \in I_0(x^*, g)\}$$

is linearly independent.

(b) The Mangasarian–Fromovitz constraint qualification (MFCQ) holds at x^* if

(i) the set $\{\nabla g_i(x^*) : i \in L_2\}$ is linearly independent,

(ii) there exists z such that $\nabla g_i(x^*)^\mathsf{T} z = 0,\ i \in L_2;\ \nabla g_i(x^*)^\mathsf{T} z < 0,\ i \in L_1 \cap I_0(x^*, g)$.

(c) The strong second-order condition (SSOC) holds at (x^*, y^*) if

$$z^\mathsf{T} \nabla_x L_D(x^*, y^*) z > 0 \quad \text{for all } z \neq 0,\ z \in Z(x^*, y^*),$$

where

$$Z(x^*, y^*) := \{z \in R^n : \nabla g_i(x^*)^\mathsf{T} z = 0,\ i \in I_N(y^*)\}.$$

(d) The general strong second-order condition (GSSOC) holds at x^* if SSOC holds at (x^*, y^*) for every $y^* \in M(x^*, f, g)$.

It is worth pointing out that strong stability is an "open" property. To be precise we have the following result, which is similar to the one given by Klatte and Tammer [7] for strong stability in nonlinear programs.

PROPOSITION 2.5. *Suppose that x^* is a strongly stable stationary solution to* VI(f, g) (*w.r.t. $\mathcal{P}'$*) *and that MFCQ holds at x^*. Let $x(\cdot)$ be the mapping appearing in Remark* 2.2. *Then for any* $(\Delta f, \Delta g) \in W(0, s)$

(i) *$x(\cdot)$ is continuous at $(\Delta f, \Delta g)$ w.r.t. the norm used in Definition* 2.1;

(ii) *$x(\Delta f, \Delta g)$ is a strongly stable stationary solution to* VI$(f + \Delta f, g + \Delta g)$ (*w.r.t. $\mathcal{P}'$*).

As in nonlinear programs (see Klatte and Tammer [7]) we have a relation between strongly stable GKKT points and strongly stable stationary solutions.

PROPOSITION 2.6. *Consider a GKKT point (x^*, y^*) to* VI(f, g). *Then the following two conditions are equivalent*:

(1) *(x^*, y^*) is a strongly stable GKKT point to* VI(f, g) (*w.r.t. $\mathcal{P}' \supset \mathcal{P}^*$*).

(2) *x^* is a strongly stable stationary solution with associated multipliers y^* to* VI(f, g) (*w.r.t. $\mathcal{P}' \supset \mathcal{P}^*$*) *and the LI condition holds at x^* for* VI(f, g).

A major tool that we use in this study is degree theory. We shall briefly review the general conditions under which a degree of a continuous function can be defined and some important properties of degrees of continuous functions. For more information the interested reader may consult Lloyd [16] and Ortega and Rheinboldt [17]. For each subset D of R^t, we use int D, bd D and $\bar{D}$ to denote the interior of D, the boundary of D, and the closure of D, respectively. Let $\phi : R^t \to R^t$ be a continuous function, D be a bounded open subset of R^t, and p be a point in R^t. If $p \notin \phi(\text{bd } D)$, then the Brouwer degree of the map ϕ at p with respect to D, denoted by $\deg(\phi, D, p)$, is well defined. Some properties of degree are as follows:

(i) If $\deg(\phi, D, p) \neq 0$ then there is a solution of $\phi(x) = p$ in D.

(ii) Suppose $\deg(\phi, D, p)$ is defined. Let ψ be a continuous function on $\bar{D}$. If $\sup\{\|\phi(x) - \psi(x)\| : x \in D\} < d(p, \phi(\text{bd } D))$, then $\deg(\psi, D, p)$ is defined and is equal to $\deg(\phi, D, p)$.

(iii) Suppose that $H : [0, 1] \times \bar{D} \to R^t$ is continuous and $p \notin H(t, \text{bd } D)$ for all $t \in [0, 1]$. Then $\deg(H(t, \cdot), D, p)$ is independent of t.

(iv) Suppose that $\deg(\phi, D, p)$ is defined. Let S be a compact subset of D such that there are no solutions of the equation $\phi(x) = p$ in S. Then $\deg(\phi, D \backslash S, p)$ is defined and is equal to $\deg(\phi, D, 0)$.

Note that from (i) we know that if $\deg(\phi, D, 0) \neq 0$, then $\phi(x) = 0$ is solvable in D. Moreover, it follows from (ii) that the solvability is preserved if the function is slightly perturbed. This fact is particularly useful for stability analysis.

3. Characterizations of strongly stable GKKT points. In the context of nonlinear programming, a strongly stable KKT point is a strongly regular KKT point and vice versa [7]. It is now well known that both concepts are related to the local homeomorphism of the linearization of the normal map $F(x,y)$ at a solution of the normal equation in question. It is not difficult to show that these relations are also retained in the context of variational inequalities. So in this section we sometimes use the term "strongly regular" and at other times we use "strongly stable" according to the context, assuming the reader is well aware of this fact. In Robinson's original paper on generalized equations [22], he obtained a complete characterization for strongly regular GKKT points, which we will state first. Then, we shall establish a new and complete characterization for strongly stable GKKT points with the aid of the idea of coorientedness. Some results in this section are direct extensions of known results from nonlinear programs. We present them here for the sake of completeness and easy reference.

We first recall Robinson's definition of a strongly regular GKKT point. Adapting Robinson's general definition of a strongly regular generalized equation [22] to our case, we say that a GKKT point (x^*,y^*) of $\mathrm{VI}(f,g)$ is strongly regular if the normal map $F_L(\cdot,\cdot)$ corresponding to the following affine variational inequality:

$$\mathrm{LVI}(f,g) \qquad \text{find } x \in LR(g) \quad \text{such that } Lf(x)^{\mathrm{T}}(x'-x) \geq 0 \quad \text{for all } x' \in LR(g),$$

where $Lf(x) := f(x^*) + \nabla f(x^*)(x - x^*)$ and $LR(g) := \{x \in R^n : Lg_i(x) := g_i(x^*) + \nabla g_i(x^*)^{\mathrm{T}}(x-x^*) \leq 0,\ i \in L_1;\ Lg_i(x) := g_i(x^*) + \nabla g_i(x^*)^{\mathrm{T}}(x-x^*) = 0,\ i \in L_2\}$, is Lipschitzian invertible at (x^*,y^*).

Recall that if a matrix H is partitioned as

$$H = \begin{bmatrix} A & B \\ C & D \end{bmatrix}$$

with A nonsingular, then the Schur complement of A in H, written (H/A), is defined as $D - CA^{-1}B$. An excellent and extensive study of this concept may be found in Oullette [18].

Below we summarize several basic results about strongly stable GKKT points and strongly regular GKKT points. A similar theorem was given by Klatte and Tammer [7] in the context of nonlinear programs.

THEOREM 3.1. *Suppose that (x^*,y^*) is a GKKT point of* $\mathrm{VI}(f,g)$. *Then the following are equivalent*:

(1) (x^*,y^*) *is a strongly regular GKKT point*;

(2) (*Robinson* [22])

(i) $M(\emptyset)$ *is nonsingular*,

(ii) $I_0(x^*,g)\backslash I_N(y^*) = \emptyset$ *or* $M(I_0(x^*,g)\backslash I_N(y^*))/M(\emptyset)$ *has positive principal minors*;

(3) (x^*,y^*) *is a strongly stable GKKT point* (*w.r.t.* $\mathcal{P}$);

(4) (x^*,y^*) *is a strongly stable GKKT point* (*w.r.t.* $\mathcal{P}^*$);

(5) $LF(\cdot,\cdot)$ *is a homeomorphism from* R^{n+m} *to* R^{n+m}.

We postpone giving the proof until we establish some basic facts for variational inequalities.

Robinson [22] applied the equivalence of (1) and (2) in Theorem 3.1 to nonlinear programming problems and proved that the LI condition and strong second-order sufficient condition are sufficient for (2). He also gave an example to show that these conditions are not necessary in general. However, combining the above theorem with Corollary 6.6 of Kojima [8], we see that if a stationary solution of a nonlinear program is known to be a local minimizer a priori, then the LI condition and strong second-order sufficient condition are also necessary for strong regularity.

The main drawback of the above characterization (Theorem 3.1 (2)) for strongly regular GKKT points may be that it uses higher-dimensional matrices instead of the Jacobian of the Lagrange function $L_D(x^*, y^*)$. Kyparisis [13] extended Robinson's results to variational inequalities showing that the LI condition and strong second-order condition are sufficient for strongly regular GKKT points. As was explained in the introduction, the strong second-order condition, although quite suitable for nonlinear programs, may not be very effective in stability analysis of variational inequalities. However, the concept of coorientedness, as we will see in the following theorem, can completely characterize the strong regularity of variational inequalities. In stating this result, for a given GKKT point (x^*, y^*), we use $C(x^*, y^*)$ to denote the critical cone at this point:

$$\begin{aligned} &C(x^*, y^*) \\ &\quad = \{z \in R^n : \nabla g_i(x^*)^{\mathsf{T}} z = 0,\ i \in I_N(y^*);\ \nabla g_i(x^*)^{\mathsf{T}} z \leq 0,\ i \in I_0(x^*, g) \backslash I_N(y^*)\}. \end{aligned}$$

THEOREM 3.2. *Let (x^*, y^*) be a GKKT point of* VI(f, g). *Then (x^*, y^*) is a strongly regular GKKT point if and only if*: (1) *The LI condition holds at x^**; *and* (2) *the coorientedness condition* (*CC*) *holds at (x^*, y^*), i.e., $\nabla L_D(x^*, y^*)$ is cooriented on $C(x^*, y^*)$.*

The coorientedness condition may not be easy to verify in general. However, in some particular cases the difficulty can be reduced. Let us see two extreme cases first. If $I_0(x^*, g) = I_N(y^*)$, i.e., the strict complementary slackness condition holds, then CC is equivalent to the condition that the restriction of $\nabla L_D(x^*, y^*)$ on the subspace $C(x^*, y^*)$ is nonsingular (in this case $C(x^*, y^*)$ is a subspace); if $I_0(x^*, g) = \emptyset$, then CC is just the nonsingular condition in the standard calculus. For the general case, we can perform the reduction procedure [27] as follows: Let E be the affine hull of $C(x^*, y^*)$, i.e., $E = \{z \in R^n : \nabla g_i(x^*)^{\mathsf{T}} z = 0,\ i \in I_N(y^*)\}$; let L be the lineality space of $C(x^*, y^*)$, i.e., $L = \{z \in R^n : \nabla g_i(x^*)^{\mathsf{T}} z = 0,\ i \in I_0(x^*, g)\}$. Now choose an orthonormal matrix partitioned as $(L\ M\ T)$ such that the columns of L, $E = (L\ M)$, and $(L\ M\ T)$ form the bases of the subspaces L, E, and R^n, respectively. Let d_L and d_E be the dimensions of L and E, respectively. If $d_L > 0$, then the section of $\nabla L_D(x^*, y^*)$ in the subspace E is

$$E^{\mathsf{T}} \nabla L_D(x^*, y^*) E = \begin{bmatrix} L^{\mathsf{T}} \nabla L_D(x^*, y^*) L & L^{\mathsf{T}} \nabla L_D(x^*, y^*) M \\ M^{\mathsf{T}} \nabla L_D(x^*, y^*) L & M^{\mathsf{T}} \nabla L_D(x^*, y^*) M \end{bmatrix}.$$

Denote by aff(M) the subspace spanned by M. If the dimension of aff(M) that is equal to $(d_E - d_L)$ is much smaller than that of $C(x^*, y^*)$ then the difficulty of checking CC can be reduced, as we will see in the following result.

THEOREM 3.2$'$. *Let (x^*, y^*) be a GKKT point of* VI(f, g). *Then (x^*, y^*) is a strongly regular GKKT point if and only if*:

(1) *the LI condition holds at x^**;

(2i) *when $d_L = 0$, $E^{\mathsf{T}} \nabla L_D(x^*, y^*) E$ is positively cooriented on* aff$(M) \cap C(x^*, y^*)$;

(2ii) *when $d_L > 0$, $L^{\mathsf{T}} \nabla L_D(x^*, y^*) L$ is nonsingular, and $E^{\mathsf{T}} \nabla L_D(x^*, y^*) E / L^{\mathsf{T}} \nabla L_D(x^*, y^*) L$ is positively cooriented on* aff$(M) \cap C(x^*, y^*)$.

Before proceeding to prove the theorems, we give an example to demonstrate a situation where the GKKT point being considered is strongly regular, but SSOC does not hold. Note that SSOC implies CC and not vice versa.

Example 3.3. Consider the following VI(f, g) with $f(x) = (1,\ x_2 + x_3,\ -4x_2 + x_3)^{\mathsf{T}}$, $g_1(x) = -x_1,\ g_2(x) = -x_2,\ g_3(x) = -x_3,\ L_1 = \{1, 2, 3\}$, and $L_2 = \emptyset$. Let $x^* = (0, 0, 0)^{\mathsf{T}}$.

It is easy to check that the point (x^*, y^*) with $y^* = (1,0,0)^{\mathrm{T}}$ is a GKKT point, and that

$$\nabla L_D(x^*, y^*) = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 1 & 1 \\ 0 & -4 & 1 \end{bmatrix},$$

and $I_0(x^*, g) = \{1,2,3\}$, $I_N(y^*) = \{1\}$. It is not difficult to check that $C(x^*, y^*) = \{z \in R^n : z_1 = 0,\ z_2 \geq 0,\ z_3 \geq 0\}$, $L = \{0\}$, aff$(M) = \{z \in R^n : z_1 = 0,\ z_2 \in R,\ z_3 \in R\}$, and $Z(x^*, y^*) =$ aff(M). Obviously, LI holds at x_0. Also, we can easily verify that condition (2i) in Theorem 3.2′ holds. Therefore, it follows that (x^*, y^*) is a strongly regular GKKT point.

We now show that SSOC does not hold at x_0. Take $z = (0,1,1)^{\mathrm{T}} \in Z(x^*, y^*)$. An algebraic calculation yields that $z^{\mathrm{T}} \nabla L_D(x^*, y^*) z = -1 < 0$. Thus SSOC does not hold at x_0.

To prove Theorem 3.2 we need several technical lemmas. The first one below presents an alternative form of characterization (2) in Theorem 3.1.

LEMMA 3.4. *Suppose (x^*, y^*) is a GKKT point of* VI(f, g). *It is a strongly regular GKKT point if and only if* signdet$M(J)$ *is a nonzero constant for all $J \subset I_0(x^*, g) \backslash I_N(y^*)$.*

Proof. It is obvious that under the hypotheses of Lemma 3.4, $M(\emptyset)$ must be nonsingular. Note that any principal submatrix of $M(I_0(x^*, g) \backslash I_N(y^*))/M(\emptyset)$ takes the form $C(J)^{\mathrm{T}} M(\emptyset)^{-1} C(J)$ for some $J \subset I_0(x^*, g) \backslash I_N(y^*)$. Conversely, for any $J \subset I_0(x^*, g) \backslash I_N(y^*)$, there is a principal submatrix of $M(I_0(x^*, g) \backslash I_N(y^*))/M(\emptyset)$ which takes the form $C(J)^{\mathrm{T}} M(\emptyset)^{-1} C(J)$. In addition, by the determinant formula of the Schur complement, we have

$$\det M(J) = \det(C(J)^{\mathrm{T}} M(\emptyset)^{-1} C(J)) \det M(\emptyset).$$

Therefore, signdet$M(J)$ has the common sign $=$ signdet$M(\emptyset)$ for all $J \subset I_0(x^*, g) \backslash I_N(y^*)$ if and only if $M(I_0(x^*, g) \backslash I_N(y^*))/M(\emptyset)$ has positive principal minors. Hence the desired conclusion follows from (2) in Theorem 3.1. □

The next lemma says that if the orientation of the section of a linear transformation in a subspace is zero, then small perturbations to the linear transformation can make the orientations of the sections of the perturbed linear transformations take any value. This lemma is an extension of Lemma 3.4 in [8].

LEMMA 3.5. *Let N be an $n \times n$ square matrix and C an $n \times p$ matrix with* rank$(C) = p$ *and $p \leq n$. Assume that* $\det(C^{\mathrm{T}} N C) = 0$. *Then there exist $n \times n$ matrices Q^+ and Q^- such that for any $\gamma > 0$,*

$$\det(C^{\mathrm{T}}(N + \gamma Q^+)C) > 0 \tag{3.1}$$

and

$$\det(C^{\mathrm{T}}(N + \gamma Q^-)C) < 0. \tag{3.2}$$

Proof. Since $C^{\mathrm{T}} N C$ is a $p \times p$ matrix with $\det(C^{\mathrm{T}} N C) = 0$, it is similar to a unique Jordan matrix D, i.e., there exists a $p \times p$ nonsingular matrix P such that

$$C^{\mathrm{T}} N C = P D P^{-1},$$

where D is a block-diagonal matrix of the form

$$D = \begin{bmatrix} D_1 & 0 & \cdots & 0 \\ 0 & D_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & D_s \end{bmatrix}.$$

Each D_i is a $p_i \times p_i$ Jordan matrix of a common (complex) eigenvalue λ_j $(j = p_1 + \cdots + p_{i-1} + 1, \ldots, p_1 + \cdots + p_{i-1} + p_i)$; λ_i and λ_j belonging to different D_i and D_j are distinct; $p_1 + p_2 + \cdots + p_s = p$; and there is at least one j such that $\lambda_j = 0$. Without loss of generality we assume that $\lambda_i = 0$ $(1 \le i \le k)$ and $0 \ne \lambda_j$ $(k+1 \le j \le p)$. Choose $\mu_i \in \{-1, 0, 1\}$ $(1 \le i \le p)$ such that

$$\mu_1 \cdots \mu_k \lambda_{k+1} \cdots \lambda_p > 0 \quad \text{and} \quad \mu_j = 0 \qquad (k+1 \le j \le p). \tag{3.3}$$

Now define a $p \times p$ matrix

$$V^+ = P \begin{bmatrix} \mu_1 & & 0 \\ & \ddots & \\ 0 & & \mu_p \end{bmatrix} P^{-1}.$$

From this construction, it is not difficult to see that

$$\det(C^{\mathrm{T}} N C + \gamma V^+) = \gamma^k \det(P) \det(P^{-1}) \mu_1 \ldots \mu_k \lambda_{k+1} \ldots \lambda_p > 0.$$

On the other hand, since C has full column rank, without loss of generality we can assume that the set of the first p rows of C is linearly independent, and then partition the matrix C into

$$C = \begin{bmatrix} C_1 \\ C_2 \end{bmatrix}$$

with $p \times p$ matrix C_1 nonsingular. Define the $n \times n$ matrix Q^+ as follows:

$$Q^+ = \begin{bmatrix} (C_1^{-1})^{\mathrm{T}} V^+ C_1^{-1} & 0 \\ 0 & 0 \end{bmatrix}.$$

Then a simple calculation shows that Q^+ satisfies (3.1) for every $\gamma > 0$. The existence of an $n \times n$ matrix Q^- satisfying (3.2) for any $\gamma > 0$ can be proved similarly by replacing (3.3) with

$$\mu_1 \ldots \mu_k \lambda_{k+1} \ldots \lambda_p < 0 \quad \text{and} \quad \mu_j = 0 \quad (k+1 \le j \le p).$$

This completes the proof. □

The following lemma presents a connection between the orientation of a linear transformation with a particular form and the orientation of the section of the linear transformation in a particular subspace. This lemma extends Theorem 3.5 of Kojima [8], where he considered the symmetric case. The proof is analogous to that of the aforementioned result in [8] except that it uses Lemma 3.5 instead of the results in [8].

LEMMA 3.6. *Consider the following square matrix M:*

$$M = \begin{bmatrix} N & B \\ -B^{\mathrm{T}} & 0 \end{bmatrix}$$

with an $n \times n$ square matrix N and an $n \times m$ matrix B, and $n \ge m$. If $\operatorname{rank}(B) = m$ *then* $\operatorname{signdet}(M) = \operatorname{signdet}(N(\ker(B^{\mathrm{T}})))$, *where* $N(\ker(B^{\mathrm{T}}))$ *is a matrix that represents the section of the linear transformation N in the subspace* $\ker(B^{\mathrm{T}})$.

Using Lemma 3.6, we can easily give the proof of Theorem 3.2.

Proof of Theorem 3.2. We prove this theorem by showing the equivalence of the hypotheses of the theorem and those of Lemma 3.4. We shall show first that the assumptions of Theorem

3.2 imply those of Lemma 3.4. Take any $J \subset I_0(x^*, g)\backslash I_N(y^*)$. Let $N = \nabla L_D(x^*, y^*)$ and let $B = (\nabla g_i(x^*)), \ (i \in I_N(y^*))$ and $C(J) = (\nabla g_i(x^*)), \ (i \in J)$. Then

$$\begin{aligned}\ker((B\ C(J))^{\mathsf{T}}) = \{z \in R^n : \nabla g_i(x^*)^{\mathsf{T}} z = 0,\ i \in I_N(y^*);\\ \nabla g_i(x^*)^{\mathsf{T}} z = 0,\ i \in J\}.\end{aligned}$$

Denote $\ker((B\ C(J))^{\mathsf{T}})$ by L. It is easy to see that $F = L \cap C(x^*, y^*)$ is a face of $C(x^*, y^*)$. Choose an orthogonal basis $V = (v_i)$ of the subspace L. By definition $V^{\mathsf{T}} N V$ is a section of N in the subspace L. Since the LI condition holds, Lemma 3.6 is applicable here, and consequently we obtain that $\mathrm{signdet}(M(J)) = \mathrm{signdet}(V^{\mathsf{T}} N V)$. Therefore, CC implies that $\mathrm{signdet} M(J)$ is a nonzero constant for all $J \subset I_0(x^*, g)\backslash I_N(y^*)$.

Similarly it is not hard to show that the hypotheses of Lemma 3.6 imply those of Theorem 3.2; we omit the details. □

Proof of the equivalence of Theorems 3.2 *and* 3.2′. The equivalence of Theorems 3.2 and 3.2′ readily follows from some results in Robinson [27]. Note that the homeomorphism condition of a normal map induced by a linear transformation A on a polyhedral convex cone C is equivalent to the coorientedness of A on C. Therefore, Propositions 2.2 and 2.3 in [27] imply that $\nabla L_D(x^*, y^*)$ is cooriented on $C(x^*, y^*)$ if and only if $E^{\mathsf{T}} \nabla L_D(x^*, y^*) E$ is cooriented on $\mathrm{aff}(M) \cap C(x^*, y^*)$ if $d_L = 0$, $L^{\mathsf{T}} \nabla L_D(x^*, y^*) L$ is nonsingular, and $E^{\mathsf{T}} \nabla L_D(x^*, y^*) E / L^{\mathsf{T}} \nabla L_D(x^*, y^*) L$ is cooriented on $\mathrm{aff}(M) \cap C(x^*, y^*)$ if $d_L > 0$. On the other hand, since $\mathrm{aff}(M) \cap C(x^*, y^*)$ is a polyhedral convex cone containing no lines, any linear transformation B is cooriented on $\mathrm{aff}(M) \cap C(x^*, y^*)$ if and only if it is positively cooriented. This completes the proof. □

Below we cite two results from [8] that are applicable here. The first one says that the linearization $LF(\cdot\,,\cdot)$ of the normal map $F(\cdot\,,\cdot)$ about the point (x^*, y^*) is a homeomorphism if and only if the determinants of $DF(x^*, y^*; \sigma)$ for all pieces σ in K have a common sign.

PROPOSITION 3.7 [8, Thm. 3.3]. *Let* (x^*, y^*) *be a GKKT point to* VI(f, g). *The linearization* $LF(\cdot\,,\cdot)$ *is a homeomorphism if and only if* $\mathrm{signdet} DF(x^*, y^*; \sigma)$ *is a nonzero constant for all* $\sigma \in K$.

The second proposition below says that if the normal map itself is a local homeomorphism then the local degree of the normal map is ± 1 and the determinants of $DF(x^*, y^*; \sigma)$ obey some inequality. Note that the homeomorphism property of $LF(\cdot\,,\cdot)$ implies the local homeomorphism property of $F(\cdot, \cdot)$ but the converse is not true in general.

PROPOSITION 3.8 [8, Lem. 2.3]. *Suppose that* $F(\cdot\,,\cdot)$ *is a local homeomorphism at* (x^*, y^*). *Then, for some* $\delta^* > 0$ *one has*

$$\deg(F, \mathrm{int}\, B(x^*, y^*, \delta), 0) = +1 \quad \text{for all } \delta \in (0, \delta^*] \tag{3.4}$$

or

$$\deg(F, \mathrm{int}\, B(x^*, y^*, \delta), 0) = -1 \quad \text{for all } \delta \in (0, \delta^*]. \tag{3.5}$$

Moreover, if (3.4) (*or* (3.5)) *holds then*

$$\mathrm{det} DF(x, y; \sigma) \geq 0 \quad (\text{or } \leq 0)$$

for all $(x, y) \in \sigma \cap \mathrm{int}\, B(x^*, y^*, \delta^*)$ *and* $\sigma \in K$.

Now we are ready to prove Theorem 3.1.

Proof of Theorem 3.1. The equivalence of (1) ⇔ (2) was shown in [22] and the equivalence of (2) ⇔ (5) follows from Lemma 3.4 and Proposition 3.7. To complete the proof, we shall show the following implications: (1) ⇒ (3), (3) ⇒ (4), and (4) ⇒ (5).

(1) $\Rightarrow$ (3): This implication readily follows from Theorem 2.1 of [22].

(3) $\Rightarrow$ (4): Since $\mathcal{P}^* \subset \mathcal{P}$, the implication is obvious.

(4) $\Rightarrow$ (5): From Proposition 2.6 we see that the LI condition holds at x^*. The rest of the proof is similar to that of Theorem 4.2 in [8] except that it uses Lemma 3.5, Lemma 3.6, Proposition 3.7, and Proposition 3.8 instead of Lemma 3.4, Theorem 3.5, Theorem 3.3, and Lemma 2.3 in [8], respectively. □

From Theorem 3.1 and Proposition 3.8 we know that the local degree of the normal map at a strongly stable GKKT point is either $+1$ or -1. We now claim that the sign of the local degree of the normal map at a strongly stable GKKT point can be determined if we have additional information about the coorientedness condition.

COROLLARY 3.9. *Let (x^*, y^*) be a strongly stable GKKT point of* VI(f, g). *Then for some $\delta^* > 0$ one has*

$$\deg(F, \operatorname{int}B(x^*, y^*, \delta), 0) = +(-)1 \quad \textit{for all } \delta \in (0, \delta^*]$$

if and only if $\nabla L_D(x^*, y^*)$ is positively (negatively) cooriented on $C(x^*, y^*)$.

Proof. Since (x^*, y^*) is a strongly stable GKKT point, the normal map $F(\cdot\,, \cdot)$ is then a local homeomorphism. Therefore, by Proposition 3.8 there exists $\delta^* > 0$ such that

$$\deg(F, \operatorname{int}B(x^*, y^*, \delta), 0) = +1$$

or

$$\deg(F, \operatorname{int}B(x^*, y^*, \delta), 0) = -1$$

for all $\delta \in (0, \delta^*]$. By Theorem 3.2 we know that $\nabla L_D(x^*, y^*)$ is either positively or negatively cooriented on $C(x^*, y^*)$. On the other hand, choose any $\sigma \in K$. If $\nabla L_D(x^*, y^*)$ is positively (negatively) cooriented on $C(x^*, y^*)$, from the proof of Theorem 3.2 we know that $\operatorname{sign}\det DF(x^*, y^*; \sigma) = +(-)1$. Then the desired result follows from the second conclusion in Proposition 3.8. □

Finally, we note a connection of strongly stable GKKT points with Clark's notion of nonsingularity of nondifferentiable maps: a GKKT point (x^*, y^*) is strongly stable if and only if $\partial F(x^*, y^*)$, the generalized Jacobian of the normal map $F(\cdot\,, \cdot)$ at (x^*, y^*), is nonsingular. The interested reader may consult Jongen, Klatte, and Tammer [6] for details.

4. Strongly stable stationary solutions.

As we have seen in the previous section, various complete characterizations for strongly stable GKKT points from different perspectives have been established, and strong stability behavior of GKKT points has been more or less well understood. However, when we are concerned with a parametric variational inequality, the assumption of the LI condition seems to be too strong because at some value of the parameter more than n constraints can be active, and in such a case the LI condition certainly fails. A reasonable replacement of the LI condition will be MFCQ. It is now known that the latter constraint qualification is necessary and sufficient for the topological stability of the feasible set [2]. In this section we shall establish the first complete characterization of strongly stable stationary solutions in the framework of MFCQ.

For convenience of reference, we list below some important properties of MFCQ.

PROPOSITION 4.1. *Suppose x^* is a feasible point of* VI(f, g) *and MFCQ holds at x^* for* VI(f, g). *Then there exist some r, $\delta > 0$ such that*

(i) *$M(x^*, f, g)$ is bounded*;

(ii) *MFCQ holds at any feasible point $x \in B(x^*, r)$ of* VI$(f + \Delta f, g + \Delta g)$ *provided that* $\operatorname{norm}(\Delta f, \Delta g, B(x^*, r)) \le \delta$;

(iii) *$M(x, f + \Delta f, g + \Delta g)$ as a function of $(x, \Delta f, \Delta g)$ is upper semicontinuous.*

Our characterization for strongly stable stationary solutions under MFCQ should reduce to the one for strongly stable stationary solutions under the LI condition when MFCQ reduces to the LI condition. Note that under MFCQ the set of associated multipliers with a stationary solution may not be a singleton. Therefore, it is reasonable to imagine that when MFCQ holds, but the LI condition does not, we shall impose a condition similar to the LI case for all GKKT points associated with the stationary solution in question. Such a condition, called the general coorientedness condition, together with MFCQ provides a complete characterization for strongly stable stationary solutions.

To motivate the general coorientedness condition, we shall briefly examine what additional difficulties would occur if the LI condition were weakened to MFCQ. We have seen in the previous sections that the coorientedness condition requires that the sections of the Jacobian matrix $\nabla L_D(x,y)$ in all subspaces spanned by the faces of the polyhedral cone $C(x,y)$ have the same nonzero orientation, so it is indeed closely related to the combinatorial structures of $C(x,y)$. Note that although both the LI condition and MFCQ are sufficient for the topological stability of the feasible set, an essential difference between the LI condition and MFCQ is that the former ensures the stability of the combinatorial structure of the polyhedral set $C(x,y)$ but the latter does not. Therefore, in the case in which MFCQ holds but the LI condition does not, we have to take care of all possible combinatorial structures that a slightly perturbed polyhedral cone $C(x,y)$ can have.

We now introduce the general coorientedness condition.

DEFINITION 4.2. *Let $x \in R^n$ be a stationary solution to* VI(f,g). *We say that the general positive (or negative) coorientedness condition (GPCC or GNCC) holds at x if for each $y \in M(x,f,g)$,*

$$\nabla L_D(x,y) \text{ is positively (or negatively) cooriented on } C(x,y;J) \quad \text{for all } J \subset I_0(x,g)\backslash I_N(y),$$

where $C(x,y;J)$ is defined as

$$C(x,y;J) = \{z \in R^n : \nabla g_i(x)^{\mathsf{T}} z = 0,\ i \in I_N(y);\ \nabla g_i(x)^{\mathsf{T}} z \le 0,\ i \in J\}.$$

We say that the general coorientedness condition (GCC) holds at x if either GPCC or GNCC holds at x.

Here is the main result of the paper.

THEOREM 4.3. *Suppose that x^* is a stationary solution of* VI(f,g) *and that MFCQ holds there. Then x^* is a strongly stable stationary solution if and only if GCC holds at x^*.*

The verification of GCC involves checking CC for each multiplier in $M(x^*,f,g)$, and this may not be an easy task. We will give more verifiable sufficient conditions later. Note that the reduction procedure described in §3 is applicable for verifying CC at a particular GKKT point (x^*,y^*) on $C(x^*,y^*;J)$.

The proof of our main result is rather long. We shall first establish some technical lemmas and then go to the main part of the proof. Although an elementary and self-contained proof is possible, we have decided to present a proof which relies on degree theory. The reasons for this choice are not only that such a proof is much shorter, but that we think the generic nature of the LI condition is the essence of the problem here. This is more in keeping with the spirit of degree theory.

An important property of GCC is that it is locally preservable. To be precise, we have the following result.

LEMMA 4.4. *Assume that x^* is a stationary solution to* VI(f,g) *and that MFCQ holds at x^*. Assume further that GPCC (or GNCC) holds at x^* for* VI(f,g). *Then there exist some $r,\ \delta > 0$ such that for any stationary solution $x \in B(x^*,r)$ to* VI$(f+\Delta f, g+\Delta g)$

satisfying $(\Delta f, \Delta g) \in \mathcal{P}$ *and* $\operatorname{norm}(\Delta f, \Delta g, B(x^*, r)) \leq \delta$, *GPCC* (*or GNCC*) *holds at* x *for* $\mathrm{VI}(f + \Delta f, g + \Delta g)$.

Proof. We only prove the case where GPCC holds at x^*. The other case can be shown similarly. Suppose the contrary. Then there exist sequences of perturbations $\{(\Delta^j f, \Delta^j g)\} \subset \mathcal{P}$, stationary solutions $\{x_j\} \subset S(f + \Delta^j f, g + \Delta^j g)$, multipliers $\{y_j\} \in M(x_j, f + \Delta^j f, g + \Delta^j g)$, index sets $\{J_j\} \subset I_0(x_j, g + \Delta^j g) \backslash I_N(y_j)$, and some orthonormal matrices $\{W_j\}$ such that $\operatorname{norm}(\Delta^j f, \Delta^j g, B(x^*, r)) \to 0$, $x_j \to x^*$, the columns of the matrices $\{W_j\}$ form bases of the subspaces spanned by some faces F_j of $C(x, y; J_j)$, and $\det((W_j)^{\mathrm{T}} \nabla L_D(x_j, y_j)(W_j)) \leq 0$. Without loss of generality we may assume that

(1) $y_j \to y^*$ (Proposition 4.1);

(2) $J_j \to J \subset I_0(x^*, g)$;

(3) $\{W_j\} \to$ some matrix W.

Then $J \subset I_0(x^*, g) \backslash I_N(y^*)$, the columns of the matrix W form a basis for a face of $C(x^*, y^*; J)$, and $\det(W^{\mathrm{T}} \nabla L_D(x^*, y^*) W) \leq 0$. This contradicts GPCC. □

In addition to the hypotheses of Lemma 4.4, if the multiplier set $M(x^*, f, g)$ is a singleton, say $\{y^*\}$, then (x^*, y^*) is a strongly stable GKKT point; if $M(x^*, f, g)$ is not a singleton, we can take an extreme point of $M(x^*, f, g)$, say y^*. We then perturb the original problem by defining

$$\begin{aligned} &\Delta^\varepsilon f(x) = 0, \\ &\Delta^\varepsilon g_i(x) = \begin{cases} 0 & \text{if } i \in I_N(y^*), \\ -\varepsilon & \text{otherwise} \end{cases} \end{aligned} \tag{4.1}$$

for $i = 1, \ldots, m$, $\varepsilon \geq 0$, and $x \in R^n$. Let

$$y_i^\varepsilon = \begin{cases} y_i^* & \text{if } i \in I_N(y^*), \\ y_i^* - \varepsilon & \text{otherwise} \end{cases}$$

for $i = 1, \ldots, m$. Then

$$F(x^*, y^\varepsilon) + \Delta^\varepsilon F(x^*, y^\varepsilon) = F(x^*, y^*) = 0 \tag{4.2}$$

for all $\varepsilon \geq 0$. This shows that x^* is also a stationary solution to $\mathrm{VI}(f + \Delta^\varepsilon f, g + \Delta^\varepsilon g)$ for all $\varepsilon \geq 0$. Actually, we can obtain more than this conclusion.

LEMMA 4.5. *Assume the hypotheses of Lemma* 4.4. *Then in the above setting there exist* δ, $\varepsilon^* > 0$ *such that* x^* *is the unique stationary solution to* $\mathrm{VI}(f + \Delta^\varepsilon f, g + \Delta^\varepsilon g)$ *in* $B(x^*, \delta)$ *for all* $\varepsilon \in [0, \varepsilon^*]$.

Proof. To avoid piecewise differentiability arguments, we shall use the generalized equation (2.2) to express the stationary conditions in the proof. Also, we shall assume without loss of generality that all constraints are active at x^*, i.e., $I_0(x^*, g) = \{1, \ldots, m\}$. Suppose the assertion is false. Then there exist sequences of perturbations $\{(\Delta^{\varepsilon_j} f, \Delta^{\varepsilon_j} g)\} \in \mathcal{P}$ and GKKT points $\{(x^j, u^j)\}$ to $\mathrm{VI}(f + \Delta^{\varepsilon_j} f, g + \Delta^{\varepsilon_j} g)$ such that $\varepsilon_j \to 0$ and $x^j \neq x^*$. Note that for all $x \in R^n$, $(f + \Delta^{\varepsilon_j} f)(x) = f(x)$ and $\nabla(g + \Delta^{\varepsilon_j} g)(x) = \nabla g(x)$. Hence for each j the point (x^j, u^j) satisfies

$$0 \in \begin{bmatrix} L_D(x^j, u^j) \\ -(g + \Delta^{\varepsilon_j} g)(x^j) \end{bmatrix} + N_{R^n \times R^k_+ \times R^l}(x^j, u^j),$$

where $L_D(x^j, u^j) := f(x^j) + \sum_{i=1}^m u_i^j \nabla g_i(x^j)$. Without loss of generality we can assume that

(i) $u^j \to \bar{u} \in M(x^*, f, g)$ and $I_+(u^j) = I_+$ for all j;

(ii) $z_j = (x^j - x^*)/\|(x^j - x^*)\| \to z$.

Then we claim that

(iii) $I_+ \supset I_+(\bar{u}),\ I_N \supset I_N(\bar{u})$ and $z \neq 0$, where $I_N = I_+ \cup L_2$;

(iv) $\nabla g_i(x^*)^{\mathsf{T}} z = 0$ for $i \in I_N(\bar{u})$ and $\nabla g_i(x^*)^{\mathsf{T}} z \geq 0$ for $i \in I_N \backslash I_N(\bar{u})$;

(v) $(\nabla_x L_D(x^*, \bar{u})z)^{\mathsf{T}} y \geq 0$ if y satisfies $\nabla g_i(x^*)^{\mathsf{T}} y = 0$ for $i \in I_N(\bar{u})$ and $\nabla g_i(x^*)^{\mathsf{T}} y \leq 0$ for $i \in I_N \backslash I_N(\bar{u})$.

Conclusion (iii) above is obvious. To prove (iv), we note first that for any $i \in I_+(y^*)$,

$$g_i(x^j) \leq 0,$$

and for any $i \in L_2$,

$$g_i(x^j) = 0.$$

Therefore, for $i \in I_+(y^*)$,

$$g_i(x^j) = g_i(x^*) + \nabla g_i(x^*)^{\mathsf{T}}(x^j - x^*) + o(\|x^j - x^*\|) \Rightarrow \nabla g_i(x^*)^{\mathsf{T}} z \leq 0, \tag{4.3}$$

and for $i \in L_2$,

$$g_i(x^j) = g_i(x^*) + \nabla g_i(x^*)^{\mathsf{T}}(x^j - x^*) + o(\|x^j - x^*\|) \Rightarrow \nabla g_i(x^*)^{\mathsf{T}} z = 0. \tag{4.4}$$

Since

$$f(x^*) + \sum_{i \in I_+(y^*)} y_i^* \nabla g_i(x^*) + \sum_{i \in L_2} y_i^* \nabla g_i(x^*) = 0,$$

we obtain from (4.3) and (4.4) that

$$f(x^*)^{\mathsf{T}} z = - \sum_{i \in I_+(y^*)} y_i^* \nabla g_i(x^*)^{\mathsf{T}} z - \sum_{i \in L_2} y_i^* \nabla g_i(x^*)^{\mathsf{T}} z \geq 0. \tag{4.5}$$

On the other hand, we have that for any $i \in I_+ = I_+(u^j)$,

$$g_i(x^j) = \varepsilon_j \quad \text{or} \quad 0 \geq 0.$$

Using the above argument we can deduce that for any $i \in I_+$,

$$\nabla g_i(x^*)^{\mathsf{T}} z \geq 0. \tag{4.6}$$

From

$$f(x^*) + \sum_{i \in I_+(\bar{u})} \bar{u}_i \nabla g_i(x^*) + \sum_{i \in L_2} \bar{u} \nabla g_i(x^*) = 0, \tag{4.7}$$

we have that using (4.4), (4.6), and the fact that $I_+(\bar{u}) \subset I_+$,

$$f(x^*)^{\mathsf{T}} z = - \sum_{i \in I_+(\bar{u})} \bar{u}_i \nabla g_i(x^*)^{\mathsf{T}} z - \sum_{i \in L_2} \bar{u}_i \nabla g_i(x^*)^{\mathsf{T}} z \leq 0. \tag{4.8}$$

Inequalities (4.5) and (4.8) imply that

$$f(x^*)^{\mathsf{T}} z = 0. \tag{4.9}$$

Then from (4.4), (4.6), (4.8), and (4.9) we can deduce that for any $i \in I_N(\bar{u})$

$$\nabla g_i(x^*)^{\mathsf{T}} z = 0.$$

Hence conclusion (iv) readily follows from the above assertion and (4.6).

We now proceed to prove conclusion (v). Since

$$\begin{aligned}
0 &= L_D(x^j, u^j) - L_D(x^*, \bar{u}) \\
&= f(x^j) + \sum_{i \in I_N(\bar{u})} u_i^j \nabla g_i(x^j) + \sum_{i \in I_N \backslash I_N(\bar{u})} u_i^j \nabla g_i(x^j) - \left[f(x^*) + \sum_{i \in I_N(\bar{u})} \bar{u}_i \nabla g_i(x^*) \right] \\
&= \nabla L_D(x^*, \bar{u})(x^j - x^*) + \sum_{i \in I_N(\bar{u})} (u_i^j - \bar{u}_i) \nabla g_i(x^*) \\
&\quad + \sum_{i \in I_N \backslash I_N(\bar{u})} u_i^j \nabla g_i(x^*) + o(\|x^j - x^*\|),
\end{aligned}$$

it follows that for any y such that $\nabla g_i(x^*)^{\mathsf{T}} y = 0$ for $i \in I_N(\bar{u})$ and $\nabla g_i(x^*)^{\mathsf{T}} y \leq 0$ for $i \in I_N \backslash I_N(\bar{u})$,

$$\begin{aligned}
&[\nabla L_D(x^*, \bar{u})(x^j - x^*) + o(\|x^j - x^*\|)]^{\mathsf{T}} y \\
&\quad = - \sum_{i \in I_N(\bar{u})} (u_i^j - \bar{u}_i) \nabla g_i(x^*)^{\mathsf{T}} y - \sum_{i \in I_N \backslash I_N(\bar{u})} u_i^j \nabla g_i(x^*)^{\mathsf{T}} y \geq 0.
\end{aligned}$$

Dividing by $\|x^j - x^*\|$ and passing to the limit, we obtain

$$(\nabla L_D(x^*, \bar{u}) z)^{\mathsf{T}} y \geq 0.$$

Hence we have completed the proof of conclusion (v).

Now consider the following affine variational inequality:

$$\text{find } x \in K \quad \text{such that } (\nabla L_D(x^*, \bar{u}) x + b)^{\mathsf{T}} (x' - x) \geq 0 \quad \text{for all } x' \in K,$$

where $K = \{x \in R^n : \nabla g_i(x^*)^{\mathsf{T}} x = 0 \text{ for } i \in I_N(\bar{u}) \text{ and } \nabla g_i(x^*)^{\mathsf{T}} x \leq 0 \text{ for } i \in I_N \backslash I_N(\bar{u})\}$ and $b \in R^n$. By assumption we know that $\nabla L_D(x^*, \bar{u})$ is cooriented on K. Therefore, it follows from Robinson's homeomorphism theorem [26, Thm. 4.3] that the above variational inequality has a unique solution for each $b \in R^n$. However, it is easy to verify that both $x = 0$ and $x = -z$ are solutions of the above variational inequality when $b = \nabla L_D(x^*, \bar{u}) z$. Thus we have a contradiction. □

The reason for our interest in the perturbed problems is that in addition to the fact that for small $\varepsilon \geq 0$, x^* is the unique stationary solution to $\mathrm{VI}(f + \Delta^\varepsilon f, g + \Delta^\varepsilon g)$ in a neighborhood of x^*, the LI condition holds at x^* for $\mathrm{VI}(f + \Delta^\varepsilon f, g + \nabla^\varepsilon g)$ provided that $\varepsilon > 0$. Since GCC is an open property, under the hypotheses of Lemma 4.4, GCC and the LI condition hold at x^* for $\mathrm{VI}(f + \Delta^\varepsilon f, g + \Delta^\varepsilon g)$ with small $\varepsilon > 0$, and, consequently, in such a case (x^*, y^ε) is a strongly stable GKKT point to $\mathrm{VI}(f + \Delta^\varepsilon f, g + \Delta^\varepsilon g)$. Therefore, we can find a sequence of strongly stable GKKT points (x^*, y^ε) to $\mathrm{VI}(f + \Delta^\varepsilon f, g + \Delta^\varepsilon g)$ which converges to (x^*, y^*). This fact readily leads to the following degree conclusion, which plays a key role in deriving the existence result for our main theorem.

LEMMA 4.6. *Assume that x^* is a stationary solution to* $\mathrm{VI}(f, g)$ *and that MFCQ holds at x^*. If GPCC (or GNCC) holds at x^* for* $\mathrm{VI}(f, g)$ *then there exist some compact set $C \subset R^m$ and $\delta^* > 0$ such that $M(x^*, f, g) \subset \mathrm{int} C$ and*

$$\deg(F, \mathrm{int}(B(x^*, \delta) \times C), 0) = +(\text{or}-)1 \quad \textit{for all } \delta \in (0, \delta^*].$$

Proof. We only prove the result when GPCC holds at x^* for VI(f, g); the other case can be proved similarly. From Proposition 4.1 and Lemmas 4.4 and 4.5 we conclude that there exist positive numbers α, δ^*, ε^*, and a compact set $C \subset R^m$ such that for any $x' \in B(x^*, \delta^*)$ and $(\Delta f, \Delta g) \in \mathcal{P}$ with $\text{norm}(\Delta f, \Delta g, B(x^*, \delta^*)) \leq \alpha$ one has

(i) $M(x', f + \Delta f, g + \Delta g) \subset \text{int}C$;

(ii) MFCQ and GPCC hold at x' for VI$(f + \Delta f, g + \Delta g)$ provided that x' is a stationary solution to VI$(f + \Delta f, g + \Delta g)$;

(iii) x^* is the unique stationary solution to VI$(f + \Delta^\varepsilon f, g + \Delta^\varepsilon g)$ in $B(x^*, \delta^*)$ for all $\varepsilon \in [0, \varepsilon^*]$, where $(\Delta^\varepsilon f, \Delta^\varepsilon g)$ was defined by (4.1).

From (i) and (iii) we have that

$$\text{GKKT}(f, g) \cap (B(x^*, \delta) \times C) = \{x^*\} \times M(x^*, f, g) \subset \text{int}(B(x^*, \delta) \times C)$$

$$\text{for all } \delta \in (0, \delta^*],$$

so $\deg(F, \text{int}(B(x^*, \delta) \times C), 0)$ is well defined for $\delta \in (0, \delta^*]$.

Without loss of generality we may assume that $\text{norm}(\Delta^\varepsilon f, \Delta^\varepsilon g, B(x^*, \delta^*)) \leq \alpha$ for all $\varepsilon \in [0, \varepsilon^*]$. Now let $\delta \in (0, \delta^*]$. From (i) and (iii) we see that for each $\varepsilon \in [0, \varepsilon^*]$ the map $F + \Delta^\varepsilon F$ cannot take the zero value on $\text{bd}(B(x^*, \delta) \times C)$. Hence from degree property (iii) listed in §2 we deduce that

$$\deg(F, \text{int}(B(x^*, \delta) \times C), 0) = \deg(F + \Delta^\varepsilon F, \text{int}(B(x^*, \delta) \times C), 0) = \text{constant} \tag{4.10}$$

for any $\varepsilon \in [0, \varepsilon^*]$. On the other hand, since LI and GPCC hold at x^* for VI$(f + \Delta^\varepsilon f, g + \Delta^\varepsilon g)$, if ε is positive and sufficiently small, it follows from Corollary 3.9 and the fact that in such a case the map $F + \Delta^\varepsilon F$ takes the zero value only at (x^*, y^ε) in $B(x^*, \delta) \times C$ that

$$\deg(F + \Delta^\varepsilon F, \text{int}(B(x^*, \delta) \times C), 0) = +1$$

for sufficiently small $\varepsilon \in (0, \varepsilon^*]$. Thus from (4.10) we obtain

$$\deg(F, \text{int}(B(x^*, \delta) \times C), 0) = +1.$$

This completes the proof. □

The following lemma gives a result concerning the subspaces spanned by the faces of a polyhedral set.

LEMMA 4.7. *Let I, J be finite index sets and let $a_i \in R^t$, $i \in I$, and $b_j \in R^t$, $j \in J$. Assume that the set $\{a_i,\ i \in I\}$ is linearly independent. Consider a polyhedral cone P defined by $P = \{z \in R^t\colon a_i^\mathsf{T} z = 0,\ i \in I;\ b_j^\mathsf{T} z \leq 0,\ j \in J\}$. Suppose that F is a face of P and L is the subspace in R^t spanned by F. Then there is a subset J_L of J such that $L = \{z \in R^t : a_i^\mathsf{T} z = 0,\ i \in I;\ b_j^\mathsf{T} z = 0,\ j \in J_L\}$ and the set $\{a_i,\ i \in I;\ b_j,\ j \in J_L\}$ is linearly independent.*

Proof. Note that L can be expressed by $\{z \in R^t : a_i^\mathsf{T} z = 0,\ i \in I;\ b_j^\mathsf{T} z = 0,\ j \in J_0\}$, where J_0 (maybe empty) is a subset of J. Therefore, we can choose a subset J_L of J_0 such that the set $\{a_i,\ i \in I;\ b_j,\ j \in J_L\}$ is a maximal linearly independent subset of $\{a_i,\ i \in I;\ b_j,\ j \in J_0\}$. Obviously we have $L = \{z \in R^t : a_i^\mathsf{T} z = 0,\ i \in I;\ b_j^\mathsf{T} z = 0,\ j \in J_L\}$. □

The next lemma provides necessary conditions for strongly stable stationary solutions under MFCQ.

LEMMA 4.8. *Assume that x^* is a strongly stable stationary solution to* VI(f, g) *and that MFCQ holds at x^*. Then there exist some compact set $C \subset R^m$ and $\delta^* > 0$ such that $M(x^*, f, g) \subset \text{int}C$ and*

$$\deg(F, \text{int}(B(x^*, \delta) \times C), 0) = +1 \quad \textit{for all } \delta \in (0, \delta^*] \tag{4.11}$$

or

$$\deg(F, \mathrm{int}(B(x^*,\delta) \times C), 0) = -1 \quad \textit{for all } \delta \in (0, \delta^*]. \tag{4.12}$$

Furthermore, if (4.11) (*or* (4.12)) *holds then for any* $y^* \in M(x^*, f, g)$ *and any* $J \subset I_0(x^*, g) \backslash I_N(y^*)$, $\nabla L_D(x^*, y^*)$ *is positively* (*or negatively*) *cooriented on* $C(x^*, y^*; J)$.

Proof. Without loss of generality we assume that $I_0(x^*, g) = \{1, \ldots, m\}$. Since x^* is a strongly stable stationary solution to VI(f, g) and MFCQ holds at x^* for VI(f, g), it follows from Propositions 2.5 and 4.1 that there exist positive numbers α, δ^*, a compact set $C \subset R^m$, and a mapping $x(\cdot)$ defined on $U(0, \alpha) = \{(\Delta f, \Delta g) \in \mathcal{P} : \mathrm{norm}(\Delta f, \Delta g, B(x^*, \delta^*)) \leq \alpha\}$ such that for any $x' \in B(x^*, \delta^*)$ and $(\Delta f, \Delta g) \in U(0, \alpha)$ one has

(i) $M(x', f + \Delta f, g + \Delta g) \subset \mathrm{int} C$;

(ii) $x(\Delta f, \Delta g)$ is the unique stationary solution and a strongly stable stationary solution to VI($f + \Delta f, g + \Delta g$) in $B(x^*, \delta^*)$.

Let $\delta \in (0, \delta^*]$. Choose an extreme point of $M(x^*, f, g)$, say $y^* \in E(x^*, f, g)$. Consider the perturbed problem VI($f + \Delta^\varepsilon f, g + \Delta^\varepsilon g$) defined by (4.1) for $\varepsilon > 0$. Then x^* is a stationary solution with associated multipliers y^ε to VI($f + \Delta^\varepsilon f, g + \Delta^\varepsilon g$) and the LI condition holds at x^* for VI($f + \Delta^\varepsilon f, g + \Delta^\varepsilon g$). On the other hand, for every sufficiently small ε satisfying $(\Delta^\varepsilon f, \Delta^\varepsilon g) \in U(0, \alpha)$, statement (ii) above implies $\{x^*\} = x(\Delta^\varepsilon f, \Delta^\varepsilon g)$. Hence, for such an ε, it follows from Proposition 2.6 that (x^*, y^ε) is a strongly stable GKKT point to VI($f + \Delta^\varepsilon f, g + \Delta^\varepsilon g$).

By (i) and (ii) we know that $\deg(F, \mathrm{int}(B(x^*, \delta) \times C), 0)$ and $\deg(F + \Delta^\varepsilon F, \mathrm{int}(B(x^*, \delta) \times C), 0)$ are well defined for every sufficiently small ε. Using degree property (iii) in §2 and the above properties (i) and (ii) again, we can deduce from Corollary 3.9 that

$$\deg(F, \mathrm{int}(B(x^*, \delta) \times C), 0) = \deg(F + \Delta^\varepsilon F, \mathrm{int}(B(x^*, \delta) \times C), 0) = \pm 1$$

for sufficiently small ε. This proves the first conclusion.

For the second conclusion we only prove the case when (4.11) holds. The other case can be proved in the same way. Suppose (4.11) holds. For the desired conclusion we shall divide our proof into three steps. In the first step we consider any extreme point of $M(x^*, f, g)$. In the second step we show that for any $y^* \in M(x^*, f, g)$ and any $J \subset I_0(x^*, g) \backslash I_N(y^*)$, $\nabla L_D(x^*, y^*)$ is cooriented on $C(x^*, y^*; J)$. In the third and final step we prove that $\nabla L_D(x^*, y^*)$ is positively cooriented on $C(x^*, y^*; J)$.

Take any $y^* \in E(x^*, f, g)$ and any $J \subset I_0(x^*, g) \backslash I_N(y^*)$. To show that $\nabla L_D(x^*, y^*)$ is positively cooriented on $C(x^*, y^*; J)$, it is sufficient to prove that for any face F of $C(x^*, y^*; J)$, letting L be the subspace spanned by F, and Q be a matrix whose columns form a basis for L, we have that $\det(Q^\mathsf{T} \nabla L_D(x^*, y^*) Q) > 0$. Now by Lemma 4.7 we find that there exists a subset J_L of J such that $L = \{z \in R^n : \nabla g_i(x^*)^\mathsf{T} z = 0,\ i \in I_N(y^*);\ \nabla g_j(x^*)^\mathsf{T} z = 0,\ j \in J_L\}$ and the set $\{\nabla g_i(x^*),\ i \in I_N(y^*);\ \nabla g_j(x^*),\ j \in J_L\}$ is linearly independent. Define the following perturbations:

$$\Delta_L^\varepsilon f(x) = 0,$$

and

$$\Delta_L^\varepsilon g_i(x) = \begin{cases} 0 & \text{if } i \in I_N(y^*) \cup J_L, \\ -\varepsilon & \text{otherwise} \end{cases}$$

for $i = 1, \ldots, m$, $\varepsilon \geq 0$, and $x \in R^n$. Let

$$y_{L_i}^\varepsilon = \begin{cases} y_i^* & \text{if } i \in I_N(y^*) \cup J_L, \\ y_i^* - \varepsilon & \text{otherwise} \end{cases}$$

for $i = 1, \ldots, m$. Then we have

$$F(x^*, y_L^\varepsilon) + \Delta_L^\varepsilon F(x^*, y_L^\varepsilon) = F(x^*, y^*) = 0$$

for all $\varepsilon \geq 0$. This means that x^* is a stationary solution to VI$(f + \Delta_L^\varepsilon f, g + \Delta_L^\varepsilon g)$ for all $\varepsilon \geq 0$. Also, the LI condition holds at x^* for VI$(f + \Delta_L^\varepsilon f, g + \Delta_L^\varepsilon g)$ when $\varepsilon > 0$ since $I_0(x^*, g + \Delta_L^\varepsilon g) = I_N(y^*) \cup J_L$. From assertion (ii) above and Proposition 2.6 we conclude that (x^*, y_L^ε) is a strongly stable GKKT point to VI $(f + \Delta_L^\varepsilon f, g + \Delta_L^\varepsilon g)$ for every sufficiently small $\varepsilon > 0$. Using the same reasoning as we did in the first part of the proof yields that for $\delta \in (0, \delta^*]$ we have

$$\deg(F + \Delta_L^\varepsilon F, \operatorname{int}(B(x^*, \delta) \times C), 0) = +1 \tag{4.13}$$

for sufficiently small $\varepsilon > 0$. Note that the function $L_D(x, y)$ takes the same value at (x^*, y^*) for the problem VI(f, g) and at (x^*, y_L^ε) for the problem VI$(f + \Delta_L^\varepsilon f, g + \Delta_L^\varepsilon g)$, and that L itself is a face of $C(x, y)$ at (x^*, y_L^ε) for VI$(f + \Delta_L^\varepsilon f, g + \Delta_L^\varepsilon g)$. Therefore, using (4.13) we find from Corollary 3.9 that $\det(Q^{\mathrm{T}} \nabla L_D(x^*, y^*) Q) > 0$. Hence we have proved that for any $y^* \in E(x^*, f, g)$, $\nabla L_D(x^*, y^*)$ is positively cooriented on $C(x^*, y^*; J)$, where $J \subset I_0(x^*, g) \backslash I_N(y^*)$.

We now show that for any $y^* \in M(x^*, f, g)$, any $J \subset I_0(x^*, g) \backslash I_N(y^*)$, $\nabla L_D(x^*, y^*)$ is cooriented on $C(x^*, y^*; J)$. To do so, note that for polyhedrally constrained variational inequalities, strong stability is equivalent to the coorientedness condition [16]. We shall first relax the original problem to discard those inequality constraints whose indexes are not in $I_N(y^*) \cup J$, and then consider the linearization of the relaxed problem about (x^*, y^*) to obtain the desired conclusion. Define

$$\begin{aligned}
&f^\varepsilon(x) := f(x^*) + \nabla L_D(x^*, y^*)(x - x^*), \\
&g_i^\varepsilon(x) := g_i(x^*) + \nabla g_i(x^*)^{\mathrm{T}}(x - x^*) \quad \text{if } i \in I_N(y^*) \cup J, \\
&g_i^\varepsilon(x) := g_i(x) - \varepsilon, \quad \text{otherwise}
\end{aligned}$$

for all $\varepsilon \geq 0$ and $x \in R^n$. Consider the following VI problem:

$$\text{VI}(\varepsilon) \qquad \text{find } x \in K^\varepsilon \quad \text{such that } f^\varepsilon(x)^{\mathrm{T}}(x' - x) \geq 0 \quad \text{for all } x' \in K^\varepsilon,$$

where $K^\varepsilon = \{x \in R^n : g_i^\varepsilon(x) \leq 0,\ i \in L_1; g_i^\varepsilon(x) = 0,\ i \in L_2\}$. The normal map corresponding to VI(ε) takes the following form:

$$F^\varepsilon(x, y) = \begin{bmatrix} f(x^*) + \nabla L_D(x^*, y^*)(x - x^*) + \sum_{i \in L_1} y_i^+ \nabla g_i^\varepsilon(x) + \sum_{i \in L_2} y_i \nabla g_i^\varepsilon(x) \\ -g_i^\varepsilon(x) + y_i^- \\ -g_j^\varepsilon(x) \end{bmatrix}.$$

For any $\gamma > 0$ such that $B(x^*, 2\gamma) \subset B(x^*, \delta^*)$, define

$$\begin{aligned}
&f^{\gamma,\varepsilon}(x) := \lambda(x, \gamma) f^\varepsilon(x) + (1 - \lambda(x, \gamma)) f(x), \\
&g_i^{\gamma,\varepsilon}(x) := \lambda(x, \gamma) g_i^\varepsilon(x) + (1 - \lambda(x, \gamma)) g_i(x), \\
&F^{\gamma,\varepsilon}(x, y) := \lambda(x, \gamma) F^\varepsilon(x, y) + (1 - \lambda(x, \gamma)) F(x, y),
\end{aligned}$$

where the function $\lambda(x, \gamma)$ is defined on R^n such that $\lambda(x, \gamma) \in C^2$, $0 \leq \lambda(x, \gamma) \leq 1$, and

$$\begin{aligned}
&\text{for any } x \in B(x^*, \gamma) \Rightarrow \lambda(x, \gamma) = 1, \\
&\text{for any } x \notin B(x^*, 2\gamma) \Rightarrow \lambda(x, \gamma) = 0.
\end{aligned}$$

The existence of such functions $\lambda(x,\gamma)$ is obvious. Define the following VI problem:

$$\text{VI}(\gamma,\varepsilon) \quad \text{find } x \in K^{\gamma,\varepsilon} \quad \text{such that } f^{\gamma,\varepsilon}(x)^{\mathsf{T}}(x'-x) \geq 0 \quad \text{for all } x' \in K^{\gamma,\varepsilon},$$

where $K^{\gamma,\varepsilon} = \{x \in R^n : g_i^{\gamma,\varepsilon}(x) \leq 0,\ i \in L_1;\ g_i^{\gamma,\varepsilon}(x) = 0,\ i \in L_2\}$. Note first that x^*, together with

$$y_i^\varepsilon = \begin{cases} y_i^* & \text{if } i \in I_N(y^*) \cup J, \\ y_i^* - \varepsilon & \text{otherwise,} \end{cases}$$

is a solution to $F^\varepsilon(x,y) = 0$ and $F^{\gamma,\varepsilon}(x,y) = 0$. Hence x^* is a stationary solution to VI(ε) and VI(γ,ε). Furthermore, we can choose sufficiently small γ' and ε' such that $(f^{\gamma',\varepsilon'} - f, g^{\gamma',\varepsilon'} - g) \in U(0,\alpha)$. Then x^* is a strongly stable stationary solution to VI(γ',ε') from statement (ii). However, from the definitions of VI(ε) and VI(γ,ε) we know that there exists a neighborhood N of x^* such that for all $x \in N$, $f^\varepsilon(x) = f^{\gamma,\varepsilon}(x)$ and $g_i^\varepsilon(x) = g_i^{\gamma,\varepsilon}(x)$, and hence VI($\varepsilon'$) is equivalent to VI($\gamma',\varepsilon'$) in this neighborhood. So x^* is also a strongly stable stationary solution of VI (ε'). Now note that for VI(ε') the possible nonlinear constraints are those whose indexes do not belong to the set $I_N(y^*) \cup J$, and these constraints can be eliminated locally at x^* without loss of generality. Therefore, VI(ε') can be treated as a polyhedrally constrained variational inequality in a small neighborhood (depending on ε) of x^*. Hence, from the fact that x^* is a strongly stable stationary solution to VI(ε'), we can deduce that the cooriented condition holds at x^* for VI(ε') [15], [26]. Therefore, $\nabla L_D(x^*,y^*)$ is cooriented on $C(x^*,y^*;J)$.

Finally, we prove that for any $y^* \in M(x^*,f,g)$ and any $J \subset I_0(x^*,g)\backslash I_N(y^*)$, $\nabla L_D(x^*,y^*)$ is positively cooriented on $C(x^*,y^*;J)$. Suppose the contrary. Then $\nabla L_D(x^*,y^*)$ is negatively cooriented on $C(x^*,y^*;J)$ since we have shown that $\nabla L_D(x^*,y^*)$ is cooriented on $C(x^*,y^*;J)$. Therefore, there exist a subspace L spanned by a face of $C(x^*,y^*;J)$ and a matrix Q whose columns form a basis of L such that $\det(Q^{\mathsf{T}}\nabla L_D(x^*,y^*)Q) < 0$. Choose a face F of $M(x^*,f,g)$ such that y^* is in the relative interior of F, and choose an extreme point y' of $M(x^*,f,g)$ such that $y' \in F$. Then we have $I_+(y') \subset I_+(y^*)$ and thus $I_N(y') \subset I_N(y^*)$. Therefore, there exists a face of $C(x^*,y';J)$ that spans L. Recall that in the second step we showed that $\nabla L_D(x^*,y')$ is positively cooriented on $C(x^*,y';J)$. This implies that $\det(Q^{\mathsf{T}}\nabla L_D(x^*,y')Q) > 0$. Define

$$\begin{aligned} y(\lambda) &:= \lambda y^* + (1-\lambda)y', \\ d(\lambda) &:= \det(Q^{\mathsf{T}}\nabla L_D(x^*,y(\lambda))Q) \end{aligned}$$

for $\lambda \in [0,1]$. Then for any $\lambda \in (0,1)$, we have that $y(\lambda) \in M(x^*,f,g)$ and $d(\lambda)$ is continuous on $[0,1]$. Therefore, since $d(0) \times d(1) < 0$ we can find a $\lambda' \in (0,1)$ such that $d(\lambda') = 0$. However, it is easy to see that there exists a face of $C(x^*,y(\lambda');J)$ that spans L. Hence $\nabla L_D(x^*,y(\lambda'))$ is not cooriented on $C(x^*,y(\lambda');J)$. This is a contradiction. □

Now we are ready to give the proof of our main result, Theorem 4.3.

Proof of Theorem 4.3. The "if" part of the proof is as follows. Under the hypotheses we know that either GPCC or GNCC holds at x^*. We only show the case where GPCC holds. The other case can be proved in the same way. From the proof of Lemma 4.6 we find that there exist positive numbers α^*, δ^*, and a compact set $C \subset R^m$ such that for any $x' \in B(x^*,\delta^*)$ and $(\Delta f,\Delta g) \in \mathcal{P}$ with $\text{norm}(\Delta f,\Delta g,B(x^*,\delta^*)) \leq \alpha^*$, one has

(i) $M(x',f+\Delta f,g+\Delta g) \subset \text{int}C$;

(ii) MFCQ and GPCC hold at x' for VI($f+\Delta f,g+\Delta g$) provided that x' is a stationary solution of VI($f+\Delta f,g+\Delta g$);

(iii) x^* is the unique stationary solution of VI(f,g) in $B(x^*,\delta^*)$;

(iv) $\deg(F, \text{int}(B(x^*, \delta^*) \times C)) = +1$.

Then we have

$$\text{GKKT}(f, g) \cap (B(x^*, \delta^*) \times C) = \{x^*\} \times M(x^*, f, g) \subset \text{int}(B(x^*, \delta^*) \times C).$$

Hence, from degree property (iii) there exists an $\alpha \in (0, \alpha^*]$ such that if $\text{norm}(\Delta f, \Delta g, B(x^*, \delta^*)) \leq \alpha$, one has

$$F(x, y) + \Delta F(x, y) \neq 0 \quad \text{for any } (x, y) \in (B(x^*, \delta^*) \times C)\backslash\text{int}(B(x^*, \delta^*/2) \times C)$$

and

$$\deg(F + \Delta F, \text{int}(B(x^*, \delta^*/2) \times C)) = +1. \tag{4.14}$$

Then by degree property (i) we conclude that there exists a solution $(x', y') \in \text{int}(B(x^*, \delta^*/2) \times C)$ to $F(x, y) + \Delta F(x, y) = 0$. So x' is a stationary solution to $\text{VI}(f + \Delta f, g + \Delta g)$.

Now we shall show the uniqueness of the stationary solution to $\text{VI}(f + \Delta f, g + \Delta g)$. Since MFCQ and GPCC hold at any stationary solution x' to $\text{VI}(f + \Delta f, g + \Delta g)$, by Lemma 4.5 we find that x' is isolated. So $B(x^*, \delta^*/2)$ contains at most a finite number of stationary solutions of $\text{VI}(f + \Delta f, g + \Delta g)$, say $x^1, \ldots, x^t$.

For each $p = 1, \ldots, t$ by statement (ii) we know that MFCQ and GPCC hold at each x^p, $p = 1, \ldots, t$ for $\text{VI}(f + \Delta f, g + \Delta g)$. Hence there exist positive numbers $\delta^1, \ldots, \delta^t$ such that

$$B(x^*, \delta^p) \times C \subset B(x^*, \delta^*/2) \times C \quad \text{and} \quad (B(x^*, \delta^i) \times C) \cap (B(x^*, \delta^j) \times C) = \emptyset \quad (i \neq j),$$

and

$$\deg(F + \Delta F, \text{int}(B(x^*, \delta^i) \times C), 0) = +1. \tag{4.15}$$

Using the degree we deduce that

$$\sum_{p=1}^{t} \deg(F + \Delta F, \text{int}(B(x^*, \delta^i) \times C), 0) = \deg(F + \Delta F, \text{int}(B(x^*, \delta^*/2) \times C), 0).$$

From (4.14) and (4.15) we find that $t = 1$. This completes the "if" part of the proof.

The "only if" part readily follows from Lemma 4.8. □

Below we present a verifiable sufficient condition for strongly stable stationary solutions. It is a direct consequence of Theorem 4.3 and the fact that GSSOC implies GPCC.

THEOREM 4.9. *Suppose x^* is a stationary solution to* $\text{VI}(f, g)$. *If MFCQ and GSSOC hold at x^*, then x^* is a strongly stable stationary solution to* $\text{VI}(f, g)$.

We conclude the paper by mentioning the article [10] in which the notion of a strongly stable Nash equilibrium point of an n-person noncooperative game was introduced and characterized.

Acknowledgments. The author thanks Professors S. M. Robinson, J. S. Pang, A. V. Fiacco, and M. Kojima for their interests and valuable suggestions, and the two anonymous referees and the associate editor who handled this paper for the many constructive comments and corrections that greatly improved the exposition in the paper. In particular, several discussions with Professor Pang led the author to employ degree theory in this study.

REFERENCES

[1] M. S. Gowda and J. S. Pang, *Stability analysis of variational inequities and nonlinear complementarity problems, via the mixed linear complementarity problems and degree theory*, Math. Oper. Res., to appear.

[2] J. Guddat, H. T. Jongen, and J. Rueckmann, *On stability and stationary points in nonlinear optimization*, J. Austral. Math. Soc., Ser. B, 28 (1986), pp. 36–56.

[3] C. D. Ha, *Stability of the linear complementarity problem at a solution point*, Math. Programming, 31 (1985), pp. 327–338.

[4] A. S. Householder, *The Theory of Matrices in Numerical Analysis*, Dover, New York, 1975.

[5] H. T. Jongen, T. Mobert, J. Rückmann, and K. Tammer, *On inertia and Schur complement in optimization*, Linear Algebra Appl., 95 (1987), pp. 97–109.

[6] H. T. Jongen, D. Klatte, and K. Tammer, *Implicit functions and sensitivity of stationary points*, Math. Programming, 49 (1990), pp. 123–138.

[7] D. Klatte and K. Tammer, *Strong stability of stationary solutions and Karush–Kuhn–Tucker points in nonlinear optimization*, Ann. Oper. Res., 27 (1990), pp. 285–310.

[8] M. Kojima, *Strongly stable stationary solutions in nonlinear programs*, in Analysis and Computation of Fixed Points, S. M. Robinson, ed., Academic Press, New York, 1979, pp. 93–138.

[9] M. Kojima and R. Hirabayashi, *Continuous deformation of nonlinear programs*, Math. Programming Study, 21 (1984), pp. 150–198.

[10] M. Kojima, A. Okada, and S. Shindoh, *Strongly stable equilibrium points of N-person noncooperative games*, Math. Oper. Res., 10 (1985), pp. 650–663.

[11] D. Kuhn and R. Lowen, *Piecewise affine bijections of R^n, and the equation $Sx^+ - Tx^- = y$*, Linear Algebra Appl., 96 (1987), pp. 109–129.

[12] J. Kyparisis, *On uniqueness of Kuhn–Tucker multipliers in nonlinear programming*, Math. Programming, 32 (1985), pp. 242–246.

[13] ———, *Sensitivity analysis framework for variational inequalities*, Math. Programming, 38 (1987), pp. 203–213.

[14] ———, *Sensitivity analysis for variational inequalities and nonlinear complementarity problems*, Ann. Oper. Res., 27 (1990), pp. 143–174.

[15] J. Liu, *Nonsingular solutions of finite-dimensional variational inequalities: Theory and methods*. Tech. report T-562, Department of Operations Research, George Washington University, Washington, DC, 1992.

[16] N. G. Lloyd, *Degree Theory*, Cambridge University Press, Cambridge, 1978.

[17] J. M. Ortega and W. C. Rheinboldt, *Iterative Solution of Nonlinear Equations in Several Variables*, Academic Press, New York, 1970.

[18] D. V. Oullette, *Schur complements and statistics*, Linear Algebra Appl., 36 (1981), pp. 187–295.

[19] D. Ralph, *Global convergence of damped Newton's method for nonsmooth equations, via the path search*, Math. Oper. Res., 39 (1994), pp. 352–389.

[20] ———, *On branching numbers of normal manifolds*, Tech. report TR 92-1283, Department of Computer Science, Cornell University, Ithaca, NY, 1992.

[21] ———, *A new proof of Robinson's homeomorphism theorem for PL-normal maps*, Linear Algebra Appl., 178 (1993), pp. 249–260.

[22] S. M. Robinson, *Strongly regular generalized equations*, Math. Oper. Res., 5 (1980), pp. 43–62.

[23] ———, *Generalized equations and their solutions, Part II: Applications to nonlinear programming*, Math. Programming Study, 19 (1982), pp. 200–221.

[24] ———, *An implicit-function theorem for a class of nonsmooth functions*, Math. Oper. Res., 16 (1991), pp. 292–309.

[25] ———, *Homeomorphism conditions for normal maps of polyhedra*, Optimization and Nonlinear Analysis, A. Ioffe, M. Marcus, and S. Reich, eds., Longman Scientific and Technical, Harlow, U.K., 1992, to appear.

[26] ———, *Normal maps induced by linear transformations*, Math. Oper. Res., 17 (1992), pp. 691–714.

[27] ———, *Nonsingularity and symmetry for linear normal maps*, Math. Programming Study, to appear.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 33, No. 3, pp. 750–764, May 1995

© 1995 Society for Industrial and Applied Mathematics

006

THE DISTURBANCE DECOUPLING PROBLEM FOR SYSTEMS OVER A RING*

G. CONTE† AND A. M. PERDON‡

Abstract. Up to now the use of geometric methods in the study of disturbance decoupling problems (DDPs) for systems over a ring has provided only necessary conditions for the existence of solutions. In this paper we study such problems, considering separately the case in which only static feedback solutions are allowed, and the one in which dynamic feedback solutions are admitted. In the first case, we give a complete geometric characterization of the solvability conditions of such problems for injective systems with coefficients in a commutative ring. Practical procedures for testing the solvability conditions and for constructing solutions, if any exist, are given in the case of systems with coefficients in a principal ideal domain (PID). In the second case, we give a complete geometric characterization of the solvability conditions for systems with coefficients in a PID.

Key words. systems over rings, dynamical feedback disturbance decoupling, invariant subspaces

AMS subject classifications. 93B27, 93B52

1. Introduction.

In the last decade the geometric approach to the theory of linear dynamical systems has provided deep insights and elegant solutions to many synthesis and control problems, such as the disturbance decoupling problem (DDP), the model matching problem, and the tracking and regulation problem (see [10] and the references therein).

Systems with coefficients in a ring turn out to be useful in many interesting situations for modeling, e.g., delay-differential systems, families of systems depending on one parameter, and time-varying systems (see [7], [8], [9]). This motivates attempts to extend the geometric approach to the theory of systems with coefficients in a ring. The first results in this direction were obtained by M. Hautus (see [5] and [6]) and by the authors in 1982. The main difference encountered in dealing with coefficients in a ring instead of coefficients in a field concerns the fundamental notion of (A, B)-invariant subspace or controlled invariant subspace [1], [10]. From a dynamical point of view, the controlled invariance property for a subspace V of the state space of a system Σ means that the trajectories starting in V can be kept inside V by a suitable choice of the control. What makes the (A, B)-invariant subspaces so useful in the theory of systems with coefficients in a field is the fact that a control with the above property can be generated using a state feedback. This means, in other words, that by choosing a suitable state feedback, any (A, B)-invariant subspace can be made invariant with respect to the resulting closed loop dynamics. Unfortunately, such a crucial feature is generally lost in the framework of systems with coefficients in a ring, where only the so-called (A, B)-invariant submodules of feedback type [5] can be made invariant by means of a state feedback.

For that reason, the classical results obtained by means of geometric techniques for systems with coefficients in a field can be generalized only partially to systems with coefficients in a ring. An illustration of this is provided in [5] in the section devoted to the DDP and in [6]. It turns out that the well-known geometric conditions which characterize the solvability of such a problem in the case of systems with coefficients in a field [1], [10] by means of a static state feedback are necessary but not sufficient, except in restricted classes of systems, for assuring the existence of solutions in the case of coefficients in a ring.

For the case of systems with coefficients in a principal ideal domain (PID), the solvability of the DDP by means of a static state feedback was characterized in [4] in terms of a necessary and sufficient condition based on the outcome of a recursive procedure. Unlike the field case, when the coefficients are taken in a ring, one can reasonably expect to achieve more using dynamic feedback rather than just static feedback. However, no general results on the DDP

* Received by the editors August 6, 1992; accepted for publication (in revised form) January 21, 1994.

† Dipartimento di Elettronica e Automatica, Università di Ancona, via Brecce Bianche, Ancona, 60131 Italy.

‡ Dipartimento di Matematica, Università di Ancona, via Brecce Bianche, Ancona, 60131 Italy.

seem to be known if dynamic feedback is allowed. So, in general, the problem of finding a complete characterization in geometric terms of the solvability conditions of that problem has remained open up to now (see the discussion in §2).

In this paper we consider the DDP for systems with coefficients in a ring from a geometric point of view, taking into account a restrictive formulation in which only static feedback solutions are allowed, and a more general one in which dynamic feedback solutions are accepted.

In the first case, the approach we develop differs slightly from the classical one. More precisely, instead of looking for solutions based on the properties of some maximal controlled invariant submodule (cf. [10]), we consider, as was partially done in [4], solutions constructed by means of minimal elements in particular lattices of controlled invariant submodules. This approach turns out to be feasible for systems that verify a mild geometric condition, which from a practical point of view is essentially equivalent to injectivity. For systems of this kind, in fact, it is possible to define the smallest controlled invariant submodule containing a given module. The properties of this geometric object allow us to obtain in §3 the first main result of the paper (Theorem 3.4), which provides a complete geometric characterization of the solvability conditions for the DDP by means of a static state feedback for systems with coefficients in a commutative ring under the sole assumption of injectivity.

Sections 4 and 5 are concerned with the problems of checking the solvability conditions mentioned above and constructing a static feedback solution, if any exists, to a given DDP. More precisely, in §4 we describe the computation of the relevant controlled invariant submodules, while in §5 we focus on the case of systems with coefficients in a PID. For these systems, we show that all the steps required for checking the solvability conditions of Theorem 3.4 and finding a solution can be performed by suitable algorithms, except for the construction of the maximal controlled invariant submodules contained in a given module. Actually, the algorithm used for this purpose in the case of coefficients in a field needs to be substituted by a procedure, described in this paper, that involves the computation of the limit of a possibly infinite sequence of submodules.

In §6 we look for solutions consisting of a dynamic state feedback and restrict our attention to the case of systems with coefficients in a PID. In this way we can take for granted that the submodules we need to consider have a basis and all the relevant constructions can be practically performed. The main result we obtain in this section is a complete geometric characterization of the solvability conditions for the DDP by means of a dynamic state feedback for systems with coefficients in a PID. If the disturbance is measurable, the solvability conditions are the same as those in the case of coefficients in a field. If the disturbance is not measurable, the solvability conditions are only technically more restrictive than the corresponding ones in the case of coefficients in a field. In conclusion, we can say that, allowing dynamic feedback solutions, the disturbance decoupling problem for systems with coefficients in a PID behaves essentially as it does in the case of coefficients in a field.

The last section describes some applications of the results of the paper to delay-differential systems and to families of parameter dependent systems.

2. Preliminaries and statement of the problems.

Let R denote a commutative ring. By a *system with coefficients in* R, or a *system over* R, we mean a discrete-time linear dynamical system Σ whose evolution is described by a set of difference equations of the form

$$\left\{\begin{array}{l} x(t+1) = Ax(t) + Bu(t), \\ y(t) = Cx(t), \end{array}\right. \tag{2.1}$$

where x belongs to the free state module $X = R^n$, u belongs to the free input module

$U = R^m$, y belongs to the free output module $Y = R^p$, and A, B, C are matrices of suitable dimensions with entries in R.

The control problems we want to consider in the framework of the systems over R are stated in the following section.

2.1. Statement of the DDPs. Given a system Σ over R, described by the equations

$$\left\{ \begin{array}{l} x(t+1) = Ax(t) + Bu(t) + Dq(t), \\ y(t) = Cx(t), \end{array} \right. \tag{2.2}$$

where $q \in Q = R^q$ is a disturbance, let us consider the feedback law

$$\left\{ \begin{array}{l} z(t+1) = A_{21}x(t) + A_{22}z(t) + G_1q(t), \\ u(t) = F_1x(t) + F_2z(t) + G_2q(t), \end{array} \right. \tag{2.3}$$

where z belongs to the free module $Z = R^{n'}$, and A_{21}, A_{22}, F_1, F_2, G_1, G_2 are matrices of suitable dimensions with entries in R.

The DDPs we consider consist, from a general point of view, of finding a feedback law of the form (2.3) such that the output of the closed loop system obtained by compensating Σ does not depend on q.

More precisely, we will speak of

(i) the disturbance decoupling problem by static feedback (static feedback DDP) if the class of solutions is restricted by requiring that n' equal zero and G_2 be the null matrix;

(ii) the disturbance decoupling problem by static feedback with measurable disturbance (static feedback DDPMD) if the class of solution is restricted by requiring that n' equal zero;

(iii) disturbance decoupling problems by dynamic feedback (dynamic feedback DDP) if the class of solutions is restricted by requiring that G_1 and G_2 be null matrices;

(iv) disturbance decoupling problems by dynamic feedback with measurable disturbance (dynamic feedback DDPMD) in the general case.

In the case of linear systems with coefficients in a field, a DDP that is solvable by means of a dynamic feedback law, i.e., a feedback law of the form (2.3) with n' different from zero, is also solvable by means of a static feedback law, i.e., a feedback law of the form (2.3) with n' equal to zero. As shown in Example 7.3, this is not the case for systems over a ring. In the field case, elegant solutions to the above control problems are provided by the so-called geometric approach (see [1], [10]). Early attempts to extend the geometric approach to the framework of systems over a ring and to apply geometric techniques to the solutions of the static feedback DDP and the static feedback DDPMD are reported in [5]. In order to describe the results obtained and to discuss the differences and the difficulties encountered in dealing with systems over a ring instead of a field, let us recall the following definition.

DEFINITION 2.1 [5]. *Given a system Σ over R described by* (2.1), *a submodule V of the state module X is called*

(i) *(A, B)-invariant, or controlled invariant, if and only if* $AV \subset V +$ Im B;

(ii) *(A, B)-invariant of feedback type if and only if there exists an R-linear map $F : X \to U$ such that $(A + BF)V \subset V$.*

A feedback F, if any exists, with the property described in (ii) *above is called a friend of V.*

The proof of the following proposition follows immediately from [1], [10].

PROPOSITION 2.2. *Given a system Σ over R described by* (2.2), *the static feedback DDP* (*respectively, the static feedback DDPMD*) *is solvable if and only if there exists an (A, B)-invariant submodule V of feedback type such that* Im $D \subset V \subset$ Ker C (*respectively, such that* Im $D \subset V +$ Im B *and* $V \subset$ Ker C).

For systems with coefficients in a field, the notions of (A, B)-invariance and (A, B)-invariance of feedback type described in Definition 2.1 are known to coincide. Moreover, the family of the (A, B)-invariant subspaces contained in ker C has a maximum element, usually denoted by V^*, which can be found as the limit of the sequence of subspaces V_k defined recursively by

$$\begin{cases} V_0 = \operatorname{Ker} C, \\ V_{k+1} = V_k \cap A^{-1}(V_k + \operatorname{Im} B). \end{cases} \tag{2.4}$$

Then, the necessary and sufficient condition of Proposition 2.2 can be stated as Im $D \subset V^*$ (respectively, Im $D \subset V^* +$ Im B) and, since the sequence (2.4) converges in a finite number of steps, it can be easily checked.

For systems with coefficients in a ring, the two notions described in Definition 2.1 do not coincide. In fact, any (A, B)-invariant submodule of feedback type is, in particular, (A, B)-invariant, but the converse is not true. The family of (A, B)-invariant submodules contained in Ker C also has in this case a maximum element V^* (see [5, Prop. 2.5]), but the condition $\operatorname{Im} D \subset V^*$ (respectively, $\operatorname{Im} D \subset V^* + \operatorname{Im} B$), which is, of course, necessary for the existence of solutions to the static feedback DDP (respectively, the static feedback DDPMD), is sufficient only if V^* happens to be of feedback type. Unfortunately, this is generally untrue for systems with coefficients in a ring (see [5], in particular, Ex. 5.6). Hence, the geometric conditions which characterize the existence of static solutions to the DDP in the case of coefficients in a field are too weak to provide a complete characterization of the solvability in the case of coefficients in a ring. Moreover, since, in general, a maximum (A, B)-invariant submodule of feedback type contained in Ker C does not exist (see [5, §5]), the above conditions cannot be modified in a straightforward way.

3. Solvability conditions for the static feedback DDPs. In order to overcome the difficulties we pointed out in the previous section, the strategy employed in [5] and [6] consists in characterizing those systems for which V^* turns out to be of feedback type. Systems of this kind, however, form only a small subset of the class at issue and, therefore, such an attempt does not provide a completely satisfactory result. Here we develop a different approach, based on a particular (A, B)-invariant submodule, that allows us to state a necessary and sufficient condition for the solvability of the static feedback DDP and the static feedback DDPMD for injective systems.

Let us start by stating some preliminary results about the lattice of (A, B)-invariant submodules of the state module X of a system Σ described by (2.1).

PROPOSITION 3.1. *Given Σ, let $V \subset X$ be an (A, B)-invariant submodule such that $V \cap$ Im $B = \{0\}$. Then, we have the following*:

(i) *if $V' \subset V$ and $V'' \subset V$ are two (A, B)-invariant submodules, then $V' \cap V''$ is an (A, B)-invariant submodule*;

(ii) *if V is of feedback type, then any (A, B)-invariant submodule V' such that $V' \subset V$ is of feedback type*; *moreover, if $F : X \to U$ is a friend of V, then F is also a friend of V'*;

(iii) *if $B : U \to X$ is injective, $F : X \to U$ is a friend of V, and $F' : X \to U$ is a friend of $V' \subset V$, then $F_{|V'} = F'_{|V'}$.*

Remark 3.2. The above results are a slight generalization of the results about self-bounded invariant subspaces given in [2] for systems with coefficients in a field. The condition $V \cap$ Im $B = \{0\}$ implies, in fact, that the invariant subspaces contained in V are self-bounded with respect to V, i.e., they are such that a trajectory starting in one of them, say V', and remaining in V, cannot escape from V'. The proofs given in [2] for the field case also work in our situation, and are reported here just for completeness and easy reference.

Proof of Proposition 3.1. (i) Given an element $v \in V' \cap V''$, by the (A, B)-invariance of V' and V'', we have $Av = v' + b' = v'' + b''$ for some $v' \in V'$, $v'' \in V''$, and b', $b'' \in$ Im B. Since v', $v'' \in V$, and $V\cap$ Im $B = \{0\}$, it follows that $v' = v'' \in V' \cap V''$ and $A(V' \cap V'') \subset (V' \cap V'')+$ Im B.

(ii) Given an element $v \in V' \subset V$, let $F : X \to U$ be a friend of V. By the (A, B)-invariance of V' and the property of F, we have $(A + BF)v = v' + b' + BFv = v''$ for some $v' \in V'$, $v'' \in V$, and $b' \in$ Im B. By $V\cap$ Im $B = \{0\}$, it follows that $v' = v'' \in V'$ and $b' = -BFv$, hence $(A + BF)V' \subset V'$.

(iii) By applying the same arguments as above, it is easy to get $BFv = BF'v$, and hence $Fv = F'v$ for any $v \in V'$.

Remark 3.3. We recall that a system Σ described by (2.1) is said to be injective if for any state x_0 in X and any pair of input sequences $\mathbf{u} = \{u(t), 0 \leq t\}$, $\mathbf{v} = \{v(t), 0 \leq t\}$ the equality $\mathbf{y}(x_0, \mathbf{u}) = \mathbf{y}(x_0, \mathbf{v})$ of the corresponding output sequences implies $\mathbf{u} = \mathbf{v}$. If Σ is injective and V^* denotes the maximum (A, B)-invariant submodule contained in Ker C, then $V^*\cap$ Im $B = \{0\}$ and B is injective. In fact, the injectivity of B is obviously implied from that of Σ and, if $B\bar{u} = x_0 \in V^*$, and $\mathbf{u} = \{u(t) = u_t, 0 \leq t\}$ is an input sequence that keeps the state trajectory starting at the initial state x_0 inside V^*, by $\mathbf{y}(0, \mathbf{u}') = \mathbf{y}(0, \mathbf{0_u})$, where $\mathbf{u}' = \{u(0) = \bar{u},\ u(t + 1) = u_t,\ 0 \leq t\}$ and $\mathbf{0_u}$ is the null input sequence, we have $\bar{u} = 0$ and $B\bar{u} = x_0 = 0$. The results of the above proposition, therefore, apply to the (A, B)-invariant submodules contained in Ker C. In particular, by Proposition 3.1 (i), we have that any nonempty lattice of (A, B)-invariant submodules contained in Ker C has a minimal element. From a practical point of view, injectivity is a very weak assumption since input sequences that produce the same output can be identified.

Notation. In the following, given a system Σ for which the condition $V^*\cap$ Im $B = \{0\}$ holds, and given a lattice $\mathcal{L}$ of (A, B)-invariant submodules of Ker C, we will denote by $V_*(\mathcal{L})$ the minimal element in the lattice.

We can now state the main result of this section.

THEOREM 3.4. *Let Σ be the system described by* (2.2) *and assume that* $V^*\cap$ Im $B = \{0\}$, *where V^* is the maximum (A, B)-invariant submodule of* Ker C. *Then, the static feedback DDP for Σ is solvable if and only if* Im $D \subset V^*$ *and* $V_*(\mathcal{L})$, *where $\mathcal{L}$ is the lattice of all the (A, B)-invariant submodules of* Ker C *containing* Im D, *is of feedback type. Analogously, the static feedback DDPMD is solvable if and only if* Im $D \subset V^*+$ Im B *and* $V_*(\mathcal{L})$, *where $\mathcal{L}$ is the lattice of all the (A, B)-invariant submodules V of* Ker C *such that* Im $D \subset V+$ Im B, *is of feedback type.*

Proof. The proof follows from Propositions 2.2 and 3.1(ii).

The above theorem gives a complete geometric characterization of the solvability conditions of the static feedback DDP and static feedback DDPMD. By comparing it to the results of [5], it appears clear that Theorem 3.4 represents a substantial improvement over what was previously known. In order for us to use it to solve the static feedback DDP and the static feedback DDPMD, we should be able to construct the submodules V^* and $V_*(\mathcal{L})$ for the lattice $\mathcal{L}$ we are interested in, and check whether the latter is of feedback type. These problems are considered in the next sections.

4. Properties and computation of $V_*(\mathcal{L})$. The results of Theorem 3.4 assume a practical relevance if a procedure for computing $V_*(\mathcal{L})$ is given. In order to relate $V_*(\mathcal{L})$ to other known geometric objects, let us recall the following definition.

DEFINITION 4.1. *Given a system Σ over R described by* (2.1), *a submodule S of the state module X is called (A, C)-invariant if and only if $A(S \cap$ Ker $C) \subset S$.*

As in the case of systems with coefficients in a field, the existence of the minimum (A, C)-invariant submodule containing a given submodule $K \subset X$, denoted by $S^*(K)$, is easily proved. For the construction of $S^*(K)$, let us consider the sequence of submodules S_k of X defined recursively by

$$(4.1)\qquad \begin{cases} S_0 = K, \\ S_{k+1} = S_k + A\,(S_k \cap \operatorname{Ker} C). \end{cases}$$

As in the case of coefficients in a field [1], $\{S_k\}$ is a nondecreasing sequence and becomes stationary for $k \geq \bar{k}$ if $S_{\bar{k}+1} = S_{\bar{k}}$. If R is a Noetherian ring we can therefore state that $\{S_k\}$ converges in a finite number of steps, and, moreover, we can prove in the same way as in [1] that the limit of $\{S_k\}$ is $S^*(K)$. The computation of $S^*(K)$ as the limit of $\{S_k\}$ is usually referred to, in the case of coefficients in a field, as the conditionally invariant subspace algorithm.

It will be useful in the following sections to note the following facts about the submodule $V^* \cap S^*(\operatorname{Im} B)$.

LEMMA 4.2. (i) *The submodule $V^* \cap S^*(\operatorname{Im} B)$ is the smallest (A, B)-invariant submodule containing $V^* \cap \operatorname{Im} B$.*

(ii) *The condition $V^* \cap \operatorname{Im} B = \{0\}$ is equivalent to the condition $V^* \cap S^*(\operatorname{Im} B) = \{0\}$.*

Proof. (i) Let us first show that any element $x \in S_k$, the kth element in the sequence defined by (4.1) with $K = \operatorname{Im} B$, can be written as $x = b + As$ with $b \in \operatorname{Im} B$ and $s \in S_{k-1} \cap \operatorname{Ker} C$. The statement is true for $k = 1$. Then, assuming it holds for $k - 1$, and writing $x \in S_k$ as $x = s + As'$ with $s \in S_{k-1}$ and $s' \in S_{k-1} \cap \operatorname{Ker} C$, one has, by $s = b + As''$ with $b \in \operatorname{Im} B$ and $s'' \in S_{k-2} \cap \operatorname{Ker} C \subset S_{k-1} \cap \operatorname{Ker} C$, $x = b + A(s'' + s')$ with $(s'' + s') \in S_{k-1} \cap \operatorname{Ker} C$.

By the above result it follows that any element $x \in V^* \cap S^*(\operatorname{Im} B)$ can be written, for some integer k, as $x = b_k + As_{k-1} = b_k + A(b_{k-1} + As_{k-2}) = b_k + Ab_{k-1} + A^2 b_{k-2} + \cdots + A^{k-1} b_1 + A^k s_0$, where $b_i \in \operatorname{Im} B$ for $1 \leq i \leq k$, $s_i = b_i + As_{i-1} \in S_i \cap \operatorname{Ker} C$ for $1 \leq i \leq k - 1$, and $s_0 \in \operatorname{Im} B \cap \operatorname{Ker} C$. In particular, the above relations imply that a trajectory starting at s_i for $0 \leq i \leq k - 1$ can be kept by a suitable input inside Ker C. Then, since V^* is the largest submodule of Ker C whose points have that property [5, §2], all the points s_i for $0 \leq i \leq k - 1$ turn out to belong to V^* and, in particular, s_0 belongs to $V^* \cap$ Im B.

Now, let $V \subset V^*$ be an (A, B)-invariant submodule containing $V^* \cap \operatorname{Im} B$ and let x be a point of $V^* \cap S^*(\operatorname{Im} B)$. With the above notation, letting $x = b_k + Ab_{k-1} + A^2 b_{k-2} + \cdots + A^{k-1} b_1 + A^k s_0$, we have that s_0 belongs to V. Choosing u_1 such that $(Bu_1 + As_0)$ belongs to V, we have $(b_1 + As_0) - (Bu_1 + As_0) = (b_1 - Bu_1) \in V^* \cap \operatorname{Im} B \subset V$. It also follows that $s_1 = b_1 + As_0$ belongs to V and, iterating this argument, that x belongs to V. Hence $V^* \cap S^*(\operatorname{Im} B)$ is contained in all the (A, B)-invariant submodules containing $V^* \cap \operatorname{Im} B$.

(ii) This part is proved by (i) and the fact that $\{0\}$ is obviously an (A, B)-invariant submodule.

Then, if we restrict our attention to systems for which the condition $V^* \cap \operatorname{Im} B = \{0\}$ holds, recalling Theorem 3.4 we have the following results (see [2] for the case of systems with coefficients in a field).

PROPOSITION 4.3. *Given a system Σ over R described by* (2.2), *assume that $V^* \cap \operatorname{Im} B = \{0\}$, where V^* is the maximum (A, B)-invariant submodule of* Ker C, *and* Im$D \subset V^*$. *If $\mathcal{L}$ denotes the lattice of all the (A, B)-invariant submodules of* Ker C *containing* Im D, *then we have $V_*(\mathcal{L}) = V^* \cap S^*(\operatorname{Im} D + \operatorname{Im} B)$.*

Proof. The proof given in the field case in [2, Thm. 2.3] works in our situation.

PROPOSITION 4.4. *Given a system Σ over R described by* (2.2), *assume that $V^* \cap$ Im $B = \{0\}$, where V^* is the maximum (A, B)-invariant submodule of* Ker C, *and* Im $D \subset V^* +$Im B. *If $\mathcal{L}$ denotes the lattice of all the (A, B)-invariant submodules V of* Ker C *such that* Im $D \subset$ $V +$ Im B, *then we have $V_*(\mathcal{L}) = V^* \cap S^*(\operatorname{Im} D + \operatorname{Im} B)$.*

Proof. Let us temporarily denote $V^* \cap S^*(\operatorname{Im} D + \operatorname{Im} B)$ by V. Then, by

$$\begin{aligned} AV &= A(V^* \cap S^*(\operatorname{Im} D + \operatorname{Im} B)) \\ &\subset (V^* + \operatorname{Im} B) \cap S^*(\operatorname{Im} D + \operatorname{Im} B) \\ &\subset (V^* \cap S^*(\operatorname{Im} D + \operatorname{Im} B)) + \operatorname{Im} B \\ &= V + \operatorname{Im} B, \end{aligned}$$

V is shown to be (A, B)-invariant. By $\operatorname{Im} D \subset V^* + \operatorname{Im} B$ and $\operatorname{Im} D \subset S^*(\operatorname{Im} D + \operatorname{Im} B)$ we have

$$\begin{aligned} \operatorname{Im} D &\subset (V^* + \operatorname{Im} B) \cap S^*(\operatorname{Im} D + \operatorname{Im} B) \\ &= (V^* \cap S^*(\operatorname{Im} D + \operatorname{Im} B)) + \operatorname{Im} B \\ &= V + \operatorname{Im} B, \end{aligned}$$

hence V belongs to $\mathcal{L}$.

Now, if V' belongs to $\mathcal{L}$, the submodule $V_1 = V' \cap V$ is also (A, B)-invariant and such that $\operatorname{Im} D \subset V_1 + \operatorname{Im} B$. By

$$\begin{aligned} A((V_1 + S^*(\operatorname{Im} B)) \cap \operatorname{Ker} C) &= A(V_1 + (S^*(\operatorname{Im} B) \cap \operatorname{Ker} C)) \\ &\subset A(V_1) + S^*(\operatorname{Im} B) \\ &\subset V_1 + \operatorname{Im} B + S^*(\operatorname{Im} B) \\ &\subset V_1 + S^*(\operatorname{Im} B), \end{aligned}$$

the submodule $(V_1 + S^*(\operatorname{Im} B))$ is shown to be (A, C)-invariant. Moreover, by

$$\operatorname{Im} D + \operatorname{Im} B \subset V_1 + \operatorname{Im} B \subset (V_1 + S^*(\operatorname{Im} B))$$

and the minimality of $S^*(\operatorname{Im} D + \operatorname{Im} B)$, recalling that $V^* \cap \operatorname{Im} B = \{0\}$ implies $V^* \cap S^*(\operatorname{Im} B) = \{0\}$, we have

$$(V^* \cap S^*(\operatorname{Im} D + \operatorname{Im} B)) + \operatorname{Im} B \subset (V^* \cap (V_1 + S^*(\operatorname{Im} B))) + \operatorname{Im} B = V_1 + \operatorname{Im} B;$$

then, by $\operatorname{Im} B \cap V^* = \{0\}$, it follows that $V = V_1 \subset V'$. This implies that V is minimal in $\mathcal{L}$, that is $V_*(\mathcal{L}) = V^* \cap S^*(\operatorname{Im} D + \operatorname{Im} B)$.

As a result of the above propositions, the computation of $V_*(\mathcal{L})$ for the lattice $\mathcal{L}$ we are interested in is reduced to the computation of V^* and $S^*(\operatorname{Im} D + \operatorname{Im} B)$.

For the computation of $S^*(\operatorname{Im} D + \operatorname{Im} B)$, we have that (4.1) initialized with $S_0 = \operatorname{Im} D + \operatorname{Im} B$ yields, if R is a Noetherian ring, $S^*(\operatorname{Im} D + \operatorname{Im} B) = S_{\bar{k}}$, where $\bar{k}$ is the smallest integer for which $S_{\bar{k}+1} = S_{\bar{k}}$.

The situation is quite different for the computation of V^*. In the case of coefficients in a field, as already mentioned, V^* can be computed as the limit of the sequence of submodules V_k defined recursively by (2.4), which converges in a finite number of steps. This is no longer true for systems over a ring R, since the convergence of $\{V_k\}$ in a finite number of steps is not guaranteed. In the next section we will give a procedure for finding V^* when R is a PID, but in general its construction remains a difficult problem.

Finally, for solving a given static feedback DDP or static feedback DDPMD by means of Theorem 3.4, one needs to check whether the submodule $V_*(\mathcal{L})$ is of feedback type or not. Since no alternative general characterization of Definition 2.1(ii) is known, this may be a quite difficult task. In the next section we will show how to accomplish this when the coefficient ring R is a PID.

5. Static feedback DDPs for systems over a PID. In this section we will consider the static feedback DDP and the static feedback DDPMD in the case in which R is a PID. This assumption gives us the possibility of characterizing the (A, B)-invariant submodules of feedback type in a quite simple way and, moreover, allows us to find a procedure for the construction of V^*. To this end, let us first recall the following notions and results of the authors from [3] and [5].

DEFINITION 5.1 [3, Def. 1.9]. *Let R be a PID and let $V \subset X = R^n$ be a submodule. By the closure of V in X we mean the submodule*

$$\overline{V} = \{x \in X, \textit{ such that } ax \in V \textit{ for some } a \neq 0,\ a \in R\}.$$

If V coincides with its closure $\overline{V}$, we say that V is closed in X.

Remark 5.2. It is useful to note the following [3, Prop. 1.10]:

(i) the closure $\overline{V}$ of V in X is the smallest closed submodule of X containing V;

(ii) any submodule V of the finitely generated free module X over a PID R is closed if and only if it is a direct summand of X;

(iii) for any submodule V of X, one has $\dim_R V = \dim_R \overline{V}$.

In addition to Remark 5.2, we point out that the computation of the closure of a submodule V can be performed by computing the Smith factorization of a suitable matrix. In fact, if D is a square matrix whose first k columns form a basis of V and whose remaining columns are zero, and if $D = PSQ$ with $S = \operatorname{diag}(a_1, \dots, a_k, 0, \dots, 0)$ is a Smith factorization, then the first k columns of $D' = PS'Q$ with $S' = \operatorname{diag}(1, \dots, 1, 0 \dots, 0)$ form a basis of $\overline{V}$. Algorithms for constructing a Smith factorization of D are known.

PROPOSITION 5.3 [5, Prop. 5.3]. *Let R be a PID and let V be an (A, B)-invariant submodule of X. Then V is of feedback type only if its closure $\overline{V}$ is (A, B)-invariant.*

PROPOSITION 5.4 [5, Prop. 5.2]. *Let R be a PID and let V be a closed (A, B)-invariant submodule of X; then V is of feedback type.*

After constructing $V_*(\mathcal{L})$, the notion of closure and Proposition 5.3 gives us the possibility, when R is a PID, of checking the solvability condition for the DDP and the DDPMD stated in Theorem 3.4. More precisely, Theorem 3.4 becomes the following theorem.

THEOREM 5.5. *Let Σ be the system described by* (2.2) *over the PID R, and assume that $V^* \cap \operatorname{Im} B = \{0\}$, where V^* is the maximum (A, B)-invariant submodule of* $\operatorname{Ker} C$. *Then, the DDP for Σ is solvable if and only if* $\operatorname{Im} D \subset V^*$ *and the closure $\overline{V_*(\mathcal{L})}$ of $V_*(\mathcal{L})$ in X, where $\mathcal{L}$ is the lattice of all the (A, B)-invariant submodules of* $\operatorname{Ker} C$ *containing* $\operatorname{Im} D$, *is (A, B)-invariant. Analogously, the DDPMD is solvable if and only if* $\operatorname{Im} D \subset V^* + \operatorname{Im} B$ *and $\overline{V_*(\mathcal{L})}$, where $\mathcal{L}$ is the lattice of all the (A, B)-invariant submodules V of* $\operatorname{Ker} C$ *such that* $\operatorname{Im} D \subset V + \operatorname{Im} B$, *is (A, B)-invariant.*

As a result, we are now left with the problem of constructing V^* in the case in which R is a PID. To this end, let us consider the following result.

PROPOSITION 5.6. *Given a system Σ described by* (2.2) *over a PID R, consider the sequence of submodules $\{V_k\}$ of X defined recursively by* (2.4) *and denote its limit by $V = \bigcap_{k=0}^{\infty} V_k$.*

(i) *If $V_k = V_{k+1}$ for some k, then $V_k = V = V^*$, that is, the sequence $\{V_k\}$ converges in a finite number of steps to the maximum (A, B)-invariant submodule of* $\operatorname{Ker} C$.

(ii) *If $V_k \neq V_{k+1}$ for all k, then $\overline{V} \overset{\subset}{\neq}$* $\operatorname{Ker} C$.

Proof. (i) The proof is the same as that in [1].

(ii) Let $\dim_R V = r$ and S be an $n \times r$ matrix whose columns span V. Choose an $r \times r$ minor S_M of S whose determinant is nonzero and let $\det S_M = d_1 d_2 \dots d_q$ be a prime decomposition of $\det S_M$. Now assume that $\overline{V} = \operatorname{Ker} C$; since $V \subset V_k \subset V_0 = \operatorname{Ker} C$ for all k and $\dim V = \dim \overline{V}$, this implies that $r = \dim V = \dim \overline{V} = \dim \operatorname{Ker} C = \dim V_k$ for all k.

Denoting by S_k an $n \times r$ matrix whose columns span V_k, since $V \subset V_k \subset V_{k-1} \subset \cdots \subset V_0 = \operatorname{Ker} C$ for all k, we have $S = S_k Q_k = S_{k-1} Q_{k-1} Q_k = S_0 Q_1 Q_2 \ldots Q_k$, where Q_i denotes a nonsingular $r \times r$ matrix for $i = 0, \ldots, k$. In particular $S_M = S_{0M} Q_1 Q_2 \ldots Q_k$, where S_{0M} is a suitable minor of S_0, and det $S_M =$ (det S_{0M})(det Q_1)(det Q_2)$\ldots$(det Q_k). Now, since prime decompositions in R are unique, if $k > q$ det Q_i is a unit in R for some i and $V_i = V_{i+1}$, we have a contradiction.

A procedure for constructing V^* in the case in which R is a PID can now be described as follows.

Step 1. By (2.4) compute the sequence $\{V_k\}$ and its limit $V^1 = \cap_{k=0}^{\infty} V_k$. If V^1 is (A, B)-invariant, in particular, if $V^1 = V_k$ for some k, $V^1 = V^*$ and the procedure stops.

Step q. In (2.4) set $V_0 = \overline{V^{q-1}}$ and compute the resulting sequence $\{V_k\}$ and its limit $V^q = \cap_{k=0}^{\infty} V_k$. If V^q is (A, B)-invariant, in particular, if $V^q = V_k$ for some k, then $V^q = V^*$ and the procedure stops.

The key fact to point out in the above procedure is that at each step either the procedure stops or the dimension of the submodule that is used for initializing (2.4) is reduced by at least one. Hence, if $r = \dim \operatorname{Ker} C$, r steps of the above procedure allow us to find V^*. No algorithm for computing the limit of $\{V_k\}$ can obviously be given, and, in general, one has to find an explicit description of V_k that facilitates the computation of the limit.

So, in conclusion, the solvability condition for the static feedback DDP and the static feedback DDPMD for a system over a PID R can be checked, and a solution can possibly be constructed by using (4.1) and the above procedure, Propositions 4.3 or 4.4, Proposition 5.3, and finally Theorem 5.5.

Remark 5.7. Let us assume for a given system Σ described by (2.2) over the PID R with $V^* \cap \operatorname{Im} B = \{0\}$, that the static feedback DDP is solvable with a static feedback F. Choosing a basis of X whose first elements are a basis of $\overline{V_*(\mathcal{L})}$, $\mathcal{L}$ being the lattice of all the (A, B)-invariant submodules of Ker C containing Im D, the compensated system Σ_F is described by equations of the form

$$\text{(5.1)} \qquad \begin{cases} x_1(t+1) = \tilde{A}_{11} x_1(t) + \tilde{A}_{12} x_2(t) + \tilde{D}_1 q(t), \\ x_2(t+1) = \tilde{A}_{22} x_2(t), \\ y(t) = \tilde{C} x_2(t). \end{cases}$$

The subsystem $x_1(t+1) = \tilde{A}_{11} x_1(t)$ evolves on $\overline{V_*(\mathcal{L})}$ with a dynamic described, up to a change of basis, by the restriction of $(A + BF)$ to such a submodule. The minimality of $V_*(\mathcal{L})$ implies that this is a fixed dynamic for the considered static feedback DDP in the sense that it is present in the compensated system for any solution F. This has important consequences if we modify the static feedback DDP or the static feedback DDPMD by requiring the additional condition that $\det(zI - A - BF)$ belongs to some specific subset $\mathcal{D}$ of $R[z]$. In particular, $\mathcal{D}$ can be chosen as in [6] to characterize a notion of internal stability. Then, the static feedback DDP or the static feedback DDPMD with the additional requirement of internal stability can be solved only if $\det(zI - \tilde{A}_{11})$ belongs to $\mathcal{D}$.

6. Solvability conditions for the dynamic feedback DDPs. In this section we consider the dynamic feedback DDPs described in §2.1 (iii) and (iv) for a system Σ over a PID R. Let us first remark that, quite obviously, we can view such problems as the search for a suitable dynamic extension Σ_E of Σ of the form

$$\text{(6.1)} \qquad \begin{cases} x(t+1) = Ax(t) + Bu(t) + Dq(t), \\ z(t+1) = A_{21} x(t) + A_{22} z(t) + G_1 q(t), \\ y(t) = Cx(t), \end{cases}$$

where possibly, if the disturbance is not measurable, G_1 is the null matrix, for which the resulting DDP admits a static feedback solution. Adopting this point of view, we are able to make use of the results of the previous sections. The first main result we obtain in this section is the following theorem.

THEOREM 6.1. *Let Σ be a system described by* (2.2) *over the PID R. Then, the dynamic feedback DDP for Σ is solvable if and only if* $\overline{\operatorname{Im} D} \subset V^*$, *where V^* is the maximum (A, B)-invariant submodule of* Ker C.

Proof. We first prove sufficiency. The proof is constructive and consists of finding a suitable dynamic extension Σ_E of Σ for which the resulting DDP admits a static feedback solution. To this end, let us assume without loss of generality that the columns of

$$\begin{bmatrix} I_h \\ 0 \end{bmatrix},$$

where I_h denotes the $h \times h$ identity matrix, form a basis of $\overline{\operatorname{Im} D}$, and the columns of

$$\begin{bmatrix} I_h & 0 \\ 0 & V \end{bmatrix},$$

where V is a suitable $(n-h) \times k$ matrix, form a basis of V^*. By (A, B)-invariance, there exist a suitable $(h+k) \times (h+k)$ matrix G and a suitable $m \times (h+k)$ matrix H such that the following matrix equality holds:

$$A \begin{bmatrix} I_h & 0 \\ 0 & V \end{bmatrix} = \begin{bmatrix} I_h & 0 \\ 0 & V \end{bmatrix} G + BH. \tag{6.2}$$

Then, partitioning $[0_{k\times h}\, I_k]\, G$, where $0_{k\times h}$ is the null matrix with k rows and h columns, as $[0_{k\times h}\, I_k]\, G = [A_0\, A_{22}]$, where A_0 and A_{22} have, respectively, h and k columns, it is easy to see that the following matrix equality holds:

$$\begin{bmatrix} A & 0 \\ [A_0\, 0_{k\times(n-h)}] & A_{22} \end{bmatrix} \begin{bmatrix} I_h & 0 \\ 0 & V \\ 0 & I_k \end{bmatrix} = \begin{bmatrix} I_h & 0 \\ 0 & V \\ 0 & I_k \end{bmatrix} G + \begin{bmatrix} B \\ 0_{k\times m} \end{bmatrix} H. \tag{6.3}$$

Now, letting $A_{21} = [A_0\, 0_{k\times(n-h)}]$, we get the dynamic extension

$$\Sigma_E = \begin{cases} x(t+1) = Ax(t) + Bu(t) + Dq(t), \\ z(t+1) = A_{21}x(t) + A_{22}z(t), \\ y(t) = Cx(t), \end{cases}$$

which exhibits the desired property. In fact, the submodule V_E, spanned in $X_E = X \oplus R^k$ by the columns of the matrix

$$\begin{bmatrix} I_h & 0 \\ 0 & V \\ 0 & I_k \end{bmatrix},$$

is easily seen by (6.3) to be

$$\left(\begin{bmatrix} A & 0 \\ A_{21} & A_{22} \end{bmatrix}, \begin{bmatrix} B \\ 0_{k\times m} \end{bmatrix} \right)\text{-invariant;}$$

it is contained in Ker $C_E = \text{Ker}\,[C\,0_{p\times k}]$ and contains

$$\text{Im}\,D_E = \text{Im}\begin{bmatrix} D \\ 0_{k\times q}\end{bmatrix}.$$

Moreover, since its basis can be completed to a basis of X_E, V_E is of feedback type by Remark 5.2(ii) and Proposition 5.4. A solution to the dynamic feedback DDDP is then given by Σ_E and any friend $[F_1\,F_2] : X \oplus R^k \to U$ of V_E.

We now prove necessity. Assume that a solution of the dynamic feedback DDP consists, in particular, of a dynamic extension Σ_E of the form (6.1), with G_1 equal to the null matrix and state module $X_E = X \oplus R^k$. Also, recalling Proposition 2.2, let V_E be an

$$\left(\begin{bmatrix} A & 0 \\ A_{21} & A_{22}\end{bmatrix}, \begin{bmatrix} B \\ 0_{k\times m}\end{bmatrix}\right)\text{-invariant submodule}$$

of feedback type, contained in Ker $C_E = \text{Ker}\,[C\,0_{p\times k}]$ and containing

$$\text{Im}\,D_E = \text{Im}\begin{bmatrix} D \\ 0_{k\times q}\end{bmatrix}.$$

Since Ker C_E is closed, by Propositions 5.3 and 5.4 we can assume without loss of generality that V_E contains $\overline{\text{Im}\,D_E}$. Denoting the canonical projection by $\pi : X \oplus R^k \to X$, we have that $\overline{\text{Im}\,D} = \pi(\overline{\text{Im}\,D_E}) \subset \pi(V_E) \subset \pi(\text{Ker}\,C_E) \subset \text{Ker}\,C$. Moreover, letting

$$B_E = \begin{bmatrix} B \\ 0_{k\times m}\end{bmatrix},$$

for any element $v \in V_E$ we have

$$\pi\left(\begin{bmatrix} A & 0 \\ A_{21} & A_{22}\end{bmatrix} v\right) = A\pi(v) = \pi(v' + B_E u) = \pi(v') + Bu$$

for some $v' \in V_E$ and $u \in U$, hence $\pi(V_E)$ is (A, B)-invariant and $\overline{\text{Im}\,D} \subset V^*$.

PROPOSITION 6.2. *Let Σ be a system described by* (2.2) *over the PID R, and let $V^* \cap \overline{\text{Im}\,B} = \{0\}$, where V^* is the maximum (A, B)-invariant submodule of* Ker C. *Assuming $\overline{\text{Im}\,D}$ is contained in V^*, let $k = \dim V_*(\mathcal{L}) - \dim(\overline{\text{Im}\,D})$, where $\mathcal{L}$ is the lattice of all the (A, B)-invariant submodules of* Ker C *containing* $\overline{\text{Im}\,D}$. *Then, the dynamic feedback DDP for Σ can be solved by a dynamic feedback law of the form* (2.3) *with* $\dim Z = k$.

Proof. The proof is obvious when we substitute $V_*(\mathcal{L})$ for V^* in the proof of Theorem 6.1.

Remark 6.3. The key point in constructing a solution to the dynamic feedback DDDP in Theorem 6.1 is the possibility of finding an integer h and a matrix V such that, in a suitable basis, Im D is contained in the submodule spanned by the columns of

$$\begin{bmatrix} I_h \\ 0 \end{bmatrix},$$

and the columns of

$$\begin{bmatrix} I_h & 0 \\ 0 & V \end{bmatrix}$$

form a basis of an (A, B)-invariant submodule of Ker C. If such a condition holds, the construction of a solution can be carried on in the same way as in the proof of the theorem on any ring R.

The above theorem shows that, allowing dynamic feedback, we get a solvability condition for the DDP akin to the one we have in the case of coefficients in a field. The situation is even better if the disturbance is measurable, since in this case the condition is the same as in the field case, the only difference being that, in the latter, the existence of dynamic feedback solutions is equivalent to the existence of static feedback solutions.

THEOREM 6.4. *Let Σ be a system described by* (2.2) *over the PID R. Then, the dynamic feedback DDPMD for Σ is solvable if and only if* $\operatorname{Im} D \subset V^* + \operatorname{Im} B$, *where V^* is the maximum (A, B)-invariant submodule of* $\operatorname{Ker} C$.

Proof. We first prove sufficiency. The proof is constructive and consists of finding a suitable dynamic extension Σ_E of Σ for which the resulting DDP admits a static feedback solution. To this end, let V be a suitable $n \times k$ matrix whose columns form a basis of V^*. By (A, B)-invariance, there exist a suitable $k \times k$ matrix A_{22} and a suitable $m \times k$ matrix H such that the following matrix equality holds:

$$AV = VA_{22} + BH; \tag{6.4}$$

hence, by

$$\begin{bmatrix} A & 0 \\ 0 & A_{22} \end{bmatrix} \begin{bmatrix} V \\ I_k \end{bmatrix} = \begin{bmatrix} V \\ I_k \end{bmatrix} A_{22} + \begin{bmatrix} B \\ 0_{k\times m} \end{bmatrix} H, \tag{6.5}$$

we have that the submodule V_E spanned by the columns of

$$\begin{bmatrix} V \\ I_k \end{bmatrix}$$

in $X_E = X \oplus R^k$ is

$$\left(\begin{bmatrix} A & 0 \\ 0 & A_{22} \end{bmatrix}, \begin{bmatrix} B \\ 0_{k\times m} \end{bmatrix} \right)\text{-invariant.}$$

Since its basis can be completed to a basis of X_E, V_E is of feedback type and is clearly contained in $\operatorname{Ker} C_E = [C \; 0_{p\times k}]$. Now, let G_1 and K be matrices of dimensions $k \times q$ and $m \times q$, respectively, such that $D = VG_1 + BK$. Then, in the extended system Σ_E given by

$$\begin{cases} x(t+1) = Ax(t) + Bu(t) + Dq(t), \\ z(t+1) = A_{22}z(t) + G_1 q(t), \\ y(t) = Cx(t), \end{cases} \tag{6.6}$$

the image of the disturbance

$$\operatorname{Im} D_E = \begin{bmatrix} D \\ G_1 \end{bmatrix}$$

is contained in V_E as shown by the equality

$$\begin{bmatrix} D \\ G_1 \end{bmatrix} = \begin{bmatrix} V \\ I_k \end{bmatrix} G_1 + \begin{bmatrix} B \\ 0_{k\times m} \end{bmatrix} K,$$

and the DDP admits a static feedback solution.

We now prove necessity. Assume that a solution of the dynamic feedback DDPMD consists, in particular, of a dynamic extension Σ_E of the form (6.1) with state space $X_E = X \oplus R^k$. Also, recalling Proposition 2.2, let V_E be an

$$\left(\begin{bmatrix} A & 0 \\ A_{21} & A_{22} \end{bmatrix}, \begin{bmatrix} B \\ 0_{k\times m} \end{bmatrix} \right)\text{-invariant submodule}$$

of feedback type, contained in Ker $C_E = \text{Ker}\,[C\ 0_{p\times k}]$ such that

$$\text{Im}\,D_E = \text{Im}\begin{bmatrix} D \\ G_1 \end{bmatrix}$$

is contained in $V_E + \text{Im}\,B_E$, where

$$B_E = \begin{bmatrix} B \\ 0_{k\times m} \end{bmatrix}.$$

Denoting by $\pi : X \oplus R^k \to X$ the canonical projection, we have that $\text{Im}\,D = \pi(\text{Im}\,D_E) \subset \pi(V_E + \text{Im}\,B_E) = \pi(V_E) + \pi(\text{Im}\,B_E) = \pi(V_E) + \text{Im}\,B$. As shown in the proof of Theorem 6.1, $\pi(V_E)$ is (A, B)-invariant and hence $\text{Im}\,D \subset V^* + \text{Im}\,B$.

7. Examples.

Example 7.1. Let Σ be the delay-differential system defined by

$$\Sigma = \begin{cases} \dot{x}_1(t) = q(t-2\partial) + q(t-\partial), \\ \dot{x}_2(t) = q(t-2\partial) + 2q(t-\partial) + q(t), \\ \dot{x}_3(t) = x_1(t) + u_1(t-\partial), \\ \dot{x}_4(t) = x_2(t) + u_2(t), \\ y_1(t) = x_3(t), \\ y_2(t) = x_4(t), \end{cases}$$

where ∂ represents a fixed delay and q is a disturbance. By introducing the delay operator Δ, defined for any time function f by $\Delta(f)(t) = f(t-\partial)$, the static feedback DDP for Σ consists of finding a delay feedback law $u(t) = F(\Delta)(x(t))$, where $F(\Delta)$ is a matrix of polynomials in Δ such that the output of the compensated system $\Sigma_{F(\Delta)}$ does not depend on q. Formally, we can associate with Σ a system of the form (2.2) defined over the ring of polynomials $R[\Delta]$ by the matrices

$$A = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{bmatrix}, \quad B = \begin{bmatrix} 0 & 0 \\ 0 & 0 \\ \Delta & 0 \\ 0 & 1 \end{bmatrix}, \quad D = \begin{bmatrix} \Delta(\Delta+1) \\ (\Delta+1)^2 \\ 0 \\ 0 \end{bmatrix},$$

$$C = \begin{bmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}.$$

The static feedback DDP for Σ can then be dealt with using the methods we developed in the previous sections.

First, an easy computation shows that Ker $C = \{(a, b, 0, 0)^T; a, b \in R[\Delta]\}$ and that $V^* = \{(\Delta a, b, 0, 0)^T; a, b \in R[\Delta]\}$. Clearly, Im D is contained in V^*, whose closure $\overline{V^*}$ coincides with Ker C. Since Ker C is not (A, B)-invariant, V^* is not of feedback type and, therefore, the condition Im $D \subset V^*$ is not sufficient for assuring the solvability of the static feedback DDP.

Note, moreover, that for this example there does not exist a maximum (A, B)-invariant submodule of feedback type contained in Ker C. In fact, both submodules $V_1 = \{(\Delta a, (\Delta + 1)a, 0, 0)^T; a \in R[\Delta]\}$ and $V_2 = \{(\Delta^2 a, (\Delta + 1)a, 0, 0)^T; a \in R[\Delta]\}$ are (A, B)-invariant and closed. Hence, the maximum (A, B)-invariant submodule of feedback type contained in Ker C, if it exists, must contain both V_1 and V_2. Then, since it may be assumed to be closed, it must also contain $\overline{V_1 + V_2}$, but as $\overline{V_1 + V_2}$ coincides with Ker C, no maximum exists.

A simple computation shows that Im D is (A, B)-invariant, hence, denoting by $\mathcal{L}$ the lattice of all the (A, B)-invariant submodules of Ker C containing Im D, we have $V_*(\mathcal{L}) = \operatorname{Im} D$. Since $D_1 = \overline{V_*(\mathcal{L})} = \{(\Delta a, (\Delta + 1)a, 0, 0)^T; a \in R[\Delta]\}$ is (A, B)-invariant, $V_*(\mathcal{L})$ is of feedback type, and there exists a solution to the considered static feedback DDP.

In order to compute a solution, let us extend the basis $\{(\Delta, \Delta + 1, 0, 0)^T\}$ of $\overline{V_*(\mathcal{L})}$ to a basis $\{(\Delta, \Delta + 1, 0, 0)^T, (1, 1, 0, 0)^T, (0, 0, 1, 0)^T, (0, 0, 0, 1)^T\}$ of $(R[\Delta])^4$, and, noting that $A((\Delta, \Delta + 1, 0, 0)^T) = B((1, \Delta + 1)^T)$, let us define a map $f : (R[\Delta])^4 \to (R[\Delta])^2$ by $f((\Delta, \Delta + 1, 0, 0)^T) = (1, \Delta + 1)^T$, and, e.g., $f((1, 1, 0, 0)^T) = f((0, 0, 1, 0)^T) = f((0, 0, 0, 1)^T) = (0, 0)^T$. The matrix F associated with f with respect to the canonical basis is

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

The compensated system Σ_F now takes the following form:

$$\Sigma_F = \begin{cases} \dot{x}_1(t) = q(t - 2\partial) + q(t - \partial), \\ \dot{x}_2(t) = q(t - 2\partial) + 2q(t - \partial) + q(t), \\ \dot{x}_3(t) = x_1(t) + x_1(t - \partial) - x_2(t - \partial), \\ \dot{x}_4(t) = x_1(t) + x_1(t - \partial) - x_2(t - \partial), \\ y_1(t) = x_3(t), \\ y_2(t) = x_4(t), \end{cases}$$

from which one can check that the second derivative $\ddot{y}(t)$ of the output $y(t)$ is equal to zero, thus proving that $y(t)$ is actually independent of q.

Example 7.2. Let $\Sigma(\mathbf{p})$ be the family of parameter dependent systems defined by

$$\Sigma(\mathbf{p}) = \begin{cases} \dot{x}_1(t) = \mathbf{p}(\mathbf{p} + 1)q(t), \\ \dot{x}_2(t) = (\mathbf{p} + 1)^2 q(t), \\ \dot{x}_3(t) = x_1(t) + \mathbf{p}u_1(t), \\ \dot{x}_4(t) = x_2(t) + u_2(t), \\ y_1(t) = x_3(t), \\ y_2(t) = x_4(t), \end{cases}$$

where $\mathbf{p}$ represents a parameter which may take values in a fixed set $\mathbf{P}$, and q is a disturbance. The static feedback DDP for the Σ we are interested in solving consists of finding a feedback law $u(t) = F(\mathbf{p})(x(t))$, where $F(\mathbf{p})$ is a matrix that depends polynomially on $\mathbf{p}$ such that the output of the compensated system $\Sigma_{F(\mathbf{p})}$ does not depend on q. Formally, we can associate with Σ a system of the form (2.2) defined over the ring of polynomials $R[\mathbf{p}]$ by the same matrices A, B, C, D as in Example 7.1 with Δ replaced by $\mathbf{p}$. The conclusion follows as in Example 7.1 and the following solution is found:

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

Example 7.3. Let us consider the PID $R[X]$ consisting of the polynomials in one variable with real coefficients, and the system Σ over $R[X]$ given by

$$\Sigma = \begin{cases} \begin{bmatrix} x_1(t+1) \\ x_2(t+1) \end{bmatrix} = \begin{bmatrix} 0 & 0 \\ 1 & 0 \end{bmatrix} \begin{bmatrix} x_1(t) \\ x_2(t) \end{bmatrix} + \begin{bmatrix} 0 \\ X \end{bmatrix} u(t) + \begin{bmatrix} X \\ 0 \end{bmatrix} q(t), \\ y(t) = [0 \quad 1] \begin{bmatrix} x_1(t) \\ x_2(t) \end{bmatrix}, \end{cases}$$

where q is a disturbance. One can interpret Σ, e.g., as the system formally associated with a delay-differential system (cf. Example 7.1) or with a family of systems depending on a parameter (cf. Example 7.2). The analysis of the DDP for Σ is easily done. The maximum (A, B)-invariant submodule contained in Ker C is the submodule V^* spanned by $\left[\begin{smallmatrix} X \\ 0 \end{smallmatrix}\right]$ and verifies the condition $V^* \cap \operatorname{Im} B = \{0\}$. We have that $\operatorname{Im} D = V^*$ and $V^* = V_*(\mathcal{L})$, if $\mathcal{L}$ is either the lattice of all the (A, B)-invariant submodules of Ker C containing $\operatorname{Im} D$, or the lattice of all the (A, B)-invariant submodules V of Ker C such that $\operatorname{Im} D \subset V + \operatorname{Im} B$. Moreover, V^* is not closed and its closure $\overline{V^*}$ is not (A, B)-invariant. As a consequence, the static feedback DDP, the static feedback DDPMD, and the dynamic feedback DDP are not solvable, respectively, by Theorem 5.5(i), (ii) and Theorem 6.1. However, the dynamic feedback DDPMD is solvable, and a solution, obtained by means of a suitable system extension as in the proof of Theorem 6.4, is given by the dynamic feedback

$$\begin{cases} z(t+1) = q(t), \\ u(t) = -z(t). \end{cases}$$

REFERENCES

[1] G. Basile and G. Marro, *Controlled and conditioned invariant subspace in linear system theory*, J. Optim. Theory Appl., 3 (1969), pp. 2305–2315.

[2] ———, *Self-bounded controlled invariant subspaces: A straightforward approach to constrained controllability*, J. Optim. Theory Appl., 38 (1982), pp. 71–81.

[3] G. Conte and A. M. Perdon, *Systems over principal ideal domains: A polinomial model approach*, SIAM J. Control Optim., 20 (1982), pp. 112–124.

[4] ———, *The disturbance decoupling problem for systems over a principal ideal domain*, in New Trends in Systems Theory, G. Conte, A. M. Perdon, and B. Wyman, eds., Birkhäuser, Basel, Switzerland, 1991.

[5] M. L. J. Hautus, *Controlled invariance in systems over rings*, Lecture Notes in Control and Inform. Sci. 39, Springer, New York, 1982.

[6] ———, *Disturbance rejection for systems over rings*, Lecture Notes in Control and Inform. Sci. 58, Springer, New York, 1984.

[7] E. Kamen, *Lectures on Algebraic System Theory. Linear Systems over Rings*, National Aeronautics and Space Administration Contractor Report 3016, 1976.

[8] E. Sontag, *Linear systems over rings: A survey*, Ricerche di Automatica, 7 (1976), pp. 1–34.

[9] ———, *Linear systems over rings: A (partial) update survey*, in Proc. International Federation on Automatic Control 81, Kyoto, Japan, 1981.

[10] M. Wohnam, *Linear Multivariable Control: A Geometric Approach*, 3rd ed., Springer-Verlag, Berlin, New York, 1985.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 33, No. 3, pp. 765–803, May 1995

© 1995 Society for Industrial and Applied Mathematics
007

CONTROL OF TRUNK LINE SYSTEMS IN HEAVY TRAFFIC*

HAROLD J. KUSHNER†

Abstract. The paper deals with the heavy traffic modelling and optimal control of trunk line networks. The controls concern either acceptance or rerouting. Both discounted and ergodic cost functions are used. It is shown that the sequence of optimal value functions of the physical process converges to the optimal value function for the limit model. Also, good controls for the limit system are good for the physical system. Single trunk systems with both priority and other inputs as well as networks are addressed. For the network case, the limit model is nonstandard, since the control occurs only on the boundary, and is a control over the direction of reflection.

Key words. trunk line networks, loss networks, queueing networks, optimal routing, heavy traffic limits, ergodic control, singular control, rerouting strategies, weak convergence

AMS subject classifications. 90B22, 60K25, 60K30, 60F17, 93E20, 93E25

1. Introduction.

This paper deals with a variety of questions concerning the modelling and optimal control of large trunk line systems under heavy traffic. The basic physical structure is that of a telephone or data communications network where there are several nodes or sources of calls or data, each of which is connected to many others via a trunk line, and each trunk line contains many individual circuits. Loosely speaking, each trunk has a primary purpose, which is to handle the traffic of a priority source that wishes to communicate between its end nodes. However, the trunks are also asked to carry other calls. For example, there might be calls or data of lower priority. In addition, if the circuits in some trunk are all occupied when a request for service between its end nodes arrives, then one might attempt to reroute that request via a pair of lines through some intermediate node and that connect the end nodes of interest. One wishes to operate the network with minimal cost or loss. The control problem consists in the decisions concerning admission of the lower priority requests or of the rerouting.

There is a very large literature on the subject, e.g., [10], [20], [21], [25], [26], [23], [7], [6]. A wealth of material is in [11]. The actual physical problem is quite difficult to treat due to the sizes of the networks and the large number of individual lines. Frequently, some sort of "fluid" model or aggregation approach is taken, and the analysis presumes stationarity. Nevertheless, a good analysis from a control theory perspective seems to be lacking, and many basic questions concerning approximations and the use of the approximations to deduce good policies remain to be answered. In particular, there does not seem to be a rigorous justification of the heavy traffic limits (i.e., when the number of lines goes to infinity) for the control and network problems. This is particularly true when the system is to be used for a long time, and the ergodic control problem is of interest.

The papers [25], [26], [7], [6] do treat heavy traffic limits. They deal with a single trunk line with a main source and (in some cases) with an extra source that is modelled essentially as a "fluid" source. [25] and [26] show via numerical comparisons that the limit model can be used to obtain nearly optimal threshold policies. For the network, the optimal policies are not usually of the threshold type and need to

* Received by the editors April 3, 1992; accepted for publication (in revised form) November 12, 1993.

† Division of Applied Mathematics, Brown University, Providence, Rhode Island, 02912. The work was partially supported by Air Force Office of Scientific Research grant AFOSR 89-0015 and National Science Foundation grant ECS-8913351.

be deduced. [7] and [6] contain a heavy traffic analysis of a single trunk line when all the interarrival and service times are exponentially distributed but do not deal explicitly with controls. The methods do not seem to be powerful enough to deal with the general control problems. The heavy traffic limits are particularly useful since numerical methods are available for the solution of the optimal control problems for the type of reflected diffusions that arise. These processes might have their control on the boundary and so are nonstandard in the control literature. They raise new questions that are of theoretical importance. Further details on the numerical methods will appear in another paper [18].

This paper starts with the analysis of a single trunk line, since this is of practical interest and since the basic ideas for the more complicated networks can be more easily developed for this simpler case. Section 2 introduces the terminology for the uncontrolled single trunk line case. The weak convergence arguments that are basic to getting the limit models are in §3. It is simpler to get the basic weak convergence ideas without a control and then to extend to the controlled case, so these early sections do not involve controls. Section 4 introduces the control problem for the single trunk line case and gives the basic stability and weak convergence result. In this case, the control concerns only the admission of the low-priority class. It is shown that the sequence of controlled physical systems converges to a well-defined controlled reflected diffusion as the number of individual circuits goes to infinity. The ergodic control problem is defined in §5. This is of particular interest if the system is to be operated over a long time period. Basically, via an occupation measure argument, it is shown that the sequence of average costs per unit time for the physical system converges to the ergodic cost for a limit process. Quite similar arguments will be used for the more complex network case. Section 6 deals with the convergence of the sequence of optimal costs for the trunk lines to the optimal cost for the control problem for the limit model, under a discounted cost criterion.

The controlled network is introduced in §7. We work with a three-node network, which thus forms a triangle. This is the simplest network of interest, but all of the results and methods work without change (except for the more complex notation) for arbitrary networks. Here the control is the decision whether to reroute or not, although admission of low-priority inputs can be added with little extra complication. The basic weak convergence result is proved, and it is seen that the limit model has certain "singular" features, since the reflection term is what is being controlled. The control occurs on the boundary of the state space, since it is only when one trunk is fully occupied that the control question arises. Optimality results for the network under a discounted cost criterion are given in §8. It is shown that the optimal costs for the network are well approximated by optimal costs for the heavy traffic limit. We emphasize that very similar methods can be used to show the convergence of numerical approximations via the Markov chain approximation method discussed in [13], [15], [16]. Indeed, we have conducted extensive numerical studies of the three-dimensional system in order to get guides concerning the structure of good policies for systems of arbitrary sizes. Some comments appear below the statement of Theorem 8.3. Controls for the three-dimensional system were adapted for use on general networks. The resulting performance on large systems (with hundreds of trunks) was very good (and relatively easy to implement), when compared in simulations to several current alternatives. The details are in [18].

The ergodic cost problem for the network is dealt with in §9 and in the Appendix (§10). We do not require that the systems be stationary. The ergodic problem is particularly difficult, since little is known about ergodic properties of boundary-controlled reflected diffusions. An outline of the results and basic ideas is given in §9, and one half of the desired convergence result is proved there. The other half is proved in the Appendix, where we use a natural "barrier" method to approximate the system by one that is better understood and to which standard "Girsanov measure transformation" methods can be used. The proof is completed by exploiting the ideas for the ergodic control problem for recurrent strong Feller processes that were developed in [12]. The major point of the Appendix is the demonstration that, for each $\epsilon > 0$, there is a smooth ϵ-optimal control for the heavy traffic limit. The reader who is willing to accept this can skip the Appendix.

We note that appropriate jump terms can be added to account for sudden changes, without changing the machinery significantly, as can state dependent arrival and service rates. Also, discrete approximations to the heavy traffic limit can be used to explore rerouting strategies via simulation, and these might be simpler than simulating the actual physical system directly. Indeed, they provide a natural aggregation of the state space.

2. An uncontrolled trunk line system: Introduction and terminology. This section defines the form of the uncontrolled one-dimensional system whose weak convergence will be proved in the next section. Controls will be added in §4. Let the trunk line contain N individual lines, with the service time on each being exponentially distributed with rate $\mu > 0$ (the mean service time is $1/\mu$). Let $\{\alpha_k^N, k \geq 1\}$ denote the sequence of interarrival intervals of the requests for service, assumed mutually independent and identically distributed (i.i.d.) for each N and independent of the service times. The sequence of service times are also i.i.d. for each N. The system starts at time zero with some given number of lines already occupied. Define $\overline{\alpha}^N = E\alpha_k^N$. On arrival of a request, if any line is available, then the arrival is assigned to some available line and service begins. Otherwise the arrival is rejected and disappears from the system.

In the heavy traffic regime in which we will be working, the mean service rate (over all lines) needs to be "nearly" equal to the mean arrival rate. If all individual lines are occupied, then the mean rate of completion of service is $N\mu$. Thus, we suppose that the mean arrival rate $(\overline{\alpha}^N)^{-1}$ satisfies

$$(\overline{\alpha}^N)^{-1} = \mu N - b_1\sqrt{N}, \tag{2.1}$$

for some given real number b_1. The $O(\sqrt{N})$ difference between the service and arrival rates is essential if the heavy traffic limit is to be nontrivial. In particular, it can be shown that if the difference were of a larger order, then the system would be fully occupied as $N \to \infty$ if $b_1 < 0$ and a negligible percentage would be occupied if $b_1 > 0$. We also will use the condition that the set of random variables

$$\left\{ \left(\frac{\alpha_k^N}{\overline{\alpha}^N} \right)^2 ; k < \infty, N < \infty \right\} \tag{2.2}$$

is uniformly integrable and that there is $\sigma^2 < \infty$ such that

$$E[1 - \alpha_k^N/\overline{\alpha}^N]^2 = \frac{\text{var } \alpha_k^N}{(\bar{\alpha}^N)^2} \equiv \sigma_N^2 \to \sigma^2. \tag{2.3}$$

Define

$A^N(t)$ = [number of arrivals by time t] $/\sqrt{N}$,
$D^N(t)$=[number of service completions by time $t]/\sqrt{N}$,
$Q^N(t)$ = [number of circuits occupied at $t]/\sqrt{N}$,
$X^N(t) = \sqrt{N} - Q^N(t)$ = [free circuits at $t]/\sqrt{N}$,
$Y^N(t)$ = [number of arrivals rejected by time $t]/\sqrt{N}$.

Now, we can write the dynamical equation

$$Q^N(t) = Q^N(0) + A^N(t) - D^N(t) - Y^N(t),$$

or equivalently for the normalized number of available lines,

$$X^N(t) = X^N(0) - A^N(t) + D^N(t) + Y^N(t). \tag{2.4}$$

The above arrival model is used because it describes common situations. In fact, $A^N(\cdot)$ can be any sequence such that the process $\overline{A}^N(\cdot)$ defined by

$$\bar{A}^N(t) = A^N(t) - \mu\sqrt{N}t$$

converges weakly to a Wiener process with constant drift and the increments

$$\{\bar{A}^N(n+1) - \bar{A}^N(n)\,; n, N\}$$

are uniformly integrable.

In order to simplify the notation, we assume that only one arrival or departure event can occur at a time. The general case only requires that we define an order for the events that occur simultaneously. This is not hard to do, but it does complicate the notation.

3. Weak convergence of $\{X^N(\cdot)\}$. The basic ideas of the weak convergence methods for the more complex network and controlled problems starting in §4 are essentially contained in the development in this section. In all of the weak convergence analysis to follow, for appropriate values of k we use the Skorohod topology on $D^k[0,\infty)$, the space of $I\!R^k$-valued functions that are right continuous and have left hand limits [4], [1]. If $k = 1$, we write just $D[0,\infty)$. For each $t > 0$, both $A^N(t)$ and $D^N(t)$ go to infinity as $N \to \infty$. In order to prove the weak convergence of $\{X^N(\cdot)\}$ and get the desired representation for the limit process, we first need to go through the "usual" procedure of representing $A^N(t)$ and $D^N(t)$ in more useful forms; i.e., as sums of the dominant part of their values (namely, $\mu\sqrt{N}t$, which will cancel each other) plus terms that will converge to either a (bounded) drift or a martingale in the limit.

Tightness in $D[0,\infty)$ will be proved via the following special case of the "Aldous–Kurtz" criterion [4, Chap. 3, Thm. 8.6c]: Let $Z_n(\cdot), n = 1, 2, \ldots$, be a sequence of processes with paths in $D[0,\infty)$ w.p.1. For each $T < \infty$, let the set of random variables $\{Z_n(s), s \leq T, n \geq 1\}$ be tight and suppose that

$$\lim_{\Delta\to 0} \sup_n \sup_{\tau\leq T} \sup_{s\leq\Delta} E\min\{1, |Z_n(\tau+s) - Z_n(\tau)|\} = 0, \tag{3.1}$$

where τ is an arbitrary stopping time bounded by T. Then $\{Z_n(\cdot)\}$ is tight.

Representation of the arrival process $A^N(\cdot)$. If α is a positive real number, Σ_1^α always denotes the sum up to the *integer part* of α. Let $\mathcal{B}_t^N$ denote the σ-algebra induced by $\{A^N(s), D^N(s), s \leq t; X^N(0)\}$. Define the auxiliary process

$$\tilde{A}^N(t) = \frac{1}{\sqrt{N}} \sum_{k=1}^{Nt} \left(1 - \frac{\alpha_k^N}{\overline{\alpha}^N}\right).$$

Let $\tilde{B}_A^N(t)$ denote the σ-algebra that measures $\{\tilde{A}^N(s), s \leq t\}$. Then the sequence of sampled values $\{\tilde{A}^N(i/N), i = 0, 1, \ldots\}$ is a $\tilde{B}_A^N(i/N)$-martingale. Define

$$\begin{aligned} S^N(t) &= \frac{1}{N} \times \max\left\{n : \sum_{k=1}^{n} \alpha_k^N \leq t\right\} \\ &= \frac{1}{N} \times [\text{number of arrivals by time } t]. \end{aligned}$$

Since by definition

$$A^N(t) = \sum_{i=1}^{NS^N(t)} 1/\sqrt{N},$$

we can write

$$A^N(t) = \tilde{A}^N(S^N(t)) + \frac{1}{\sqrt{N}} \sum_{k=1}^{NS^N(t)} \alpha_k^N / \overline{\alpha}^N. \tag{3.2}$$

By (2.1), the last term on the right-hand side can be written as

$$\frac{1}{\sqrt{N}} \sum_{k=1}^{NS^N(t)} \alpha_k^N (\mu N - b_1 \sqrt{N}). \tag{3.3}$$

To simplify (3.3), we will use the relationship (which follows from the definition of $S^N(t)$)

$$\sum_{k=1}^{NS^N(t)} \alpha_k^N = t + O_N(t), \tag{3.4}$$

where $O_N(t) = -[t - \text{time of last arrival before } t]$ and $EO_N(t) = O(1/N)$, $EO_N^2(t) = O(1/N^2)$.

Representation of the departure process $D^N(\cdot)$. Using the fact that the service intervals are mutually independent and exponentially distributed with rate μ, we can decompose $D^N(\cdot)$ in "martingale" form as

$$\begin{aligned} D^N(t) &= \mu \int_0^t Q^N(s) ds + \tilde{D}^N(t) \\ &= -\mu \int_0^t X^N(s) ds + \sqrt{N} \mu t + \tilde{D}^N(t), \end{aligned} \tag{3.5}$$

where $\tilde{D}^N(\cdot)$ is a $\mathcal{B}_t^N$-martingale with quadratic variation

$$\langle \tilde{D}^N \rangle(t) = \mu \int_0^t Q^N(s) ds / \sqrt{N} = \mu \int_0^t \left[1 - \frac{X^N(s)}{\sqrt{N}}\right] ds. \tag{3.6}$$

Define $\tilde{A}_0^N(t) = -\tilde{A}^N(S^N(t))$. Note that $\tilde{A}_0^N(\cdot)$ is $\mathcal{B}_t^N$-measurable. Finally, using (3.2)–(3.5), (2.4) can be rewritten in the more useful form

$$\begin{aligned}X^N(t) &= X^N(0) + b_1 t - \mu\int_0^t X^N(s)ds + \tilde{A}_0^N(t)\\ &\quad + \tilde{D}^N(t) + Y^N(t) + \sqrt{N}O_N(t) = Y^N(t) + Z^N(t),\end{aligned} \tag{3.7}$$

where $Z^N(\cdot)$ is defined in the obvious way.

The first convergence theorem. We can now state the basic weak convergence theorem.

THEOREM 3.1. *Assume the conditions of this section and §2, and let* $\sup_N E|X^N(0)|^2 < \infty$. *Then* $\{X^N(\cdot), \tilde{A}_0^N(\cdot), \tilde{D}^N(\cdot), Y^N(\cdot)\}$ *is tight. Let* $(X(\cdot), W_1(\cdot), W_2(\cdot), Y(\cdot))$ *denote the limit of a weakly convergent subsequence. Then the limit processes have continuous paths w.p.*1 *and satisfy*

$$X(t) = X(0) + b_1 t - \mu\int_0^t X(s)ds + W_1(t) + W_2(t) + Y(t),\ X(t) \geq 0, \tag{3.8}$$

where the $W_i(\cdot)$ *are mutually independent Wiener processes with quadratic variations* $\langle W_1\rangle(t) = \sigma^2\mu t$, $\langle W_2\rangle(t) = \mu t$. *The other processes are nonanticipative with respect to the Wiener processes. Also* $Y(\cdot)$ *can increase only at those* t *for which* $X(t) = 0$. *The set* $\{X^N(t); N < \infty, t < \infty\}$ *is tight.*

Proof. Since the proof is essentially "classical" and closely related arguments are in [24], [17], [19], we only give the outline. It can easily be shown that $\{S^N(\cdot)\}$ converges weakly to the function with values μt, the limit of the mean number of arrivals on $[0, t]$ divided by N, and we omit the details.

Note that, for any bounded $\tilde{\mathcal{B}}_A^N(t)$-stopping time τ,

$$E[(\tilde{A}^N(\tau + t) - \tilde{A}^N(\tau))^2 | \tilde{\mathcal{B}}_A^N(\tau)] \leq \sigma_N^2 t, \tag{3.9a}$$

and the left-hand side converges to $\sigma^2 t$. For any bounded $\mathcal{B}_t^N$-stopping time τ,

$$\begin{aligned}E[(\tilde{D}^N(\tau+t) - \tilde{D}^N(\tau))^2|\mathcal{B}_\tau^N] &= \mu t - \mu E\left[\int_\tau^{\tau+t} X^N(s)ds/\sqrt{N}|\mathcal{B}_\tau^N\right]\\ &= E[\langle\tilde{D}^N\rangle(\tau+t) - \langle\tilde{D}^N\rangle(\tau)|\mathcal{B}_\tau^N],\\ E[(\tilde{A}_0^N(\tau+t) - \tilde{A}_0^N(\tau))(\tilde{D}^N(\tau+t) &- \tilde{D}^N(\tau))|\mathcal{B}_\tau^N] = 0.\end{aligned} \tag{3.9b}$$

Since $X^N(t) \leq \sqrt{N}$, (3.9a) and the first two lines of (3.9b) and the criterion (3.1) yield tightness of $\{\tilde{A}^N(\cdot), \tilde{D}^N(\cdot)\}$. (To get a bound on the expectation in (3.1) take expectations above and use Shwarz's inequality.)

The weak limits of $\{\tilde{D}^N(\cdot)\}$ are continuous because of the tightness and the fact that the "jumps" are of order $O(1/\sqrt{N})$. Next we show the continuity of the weak limits of $\{\tilde{A}^N(\cdot)\}$. To do this, it is sufficient to show that the jumps of $\tilde{A}^N(\cdot)$ go to zero as $N \to \infty$, in the sense that for any fixed $t < \infty$ and $\varepsilon > 0$

$$P\left\{\sup_{k\leq Nt}\frac{\alpha_k^N}{\bar{\alpha}^N\sqrt{N}} \geq \varepsilon\right\} \xrightarrow{N} 0. \tag{3.10}$$

Define the set

$$K_k^{N,\varepsilon} = \left\{ \omega : \frac{\alpha_k^N}{\bar{\alpha}^N \sqrt{N}} \geq \varepsilon \right\}.$$

By Chebychev's inequality, we write (and define δ_N)

$$P\{K_k^{N,\varepsilon}\} \leq \frac{1}{N\varepsilon^2} \int_{K_k^{N,\varepsilon}} \left(\frac{\alpha_k^N}{\bar{\alpha}^N} \right)^2 dP = \delta_N / N,$$

where by the uniform integrability of the set (2.2), $\delta_N \xrightarrow{N} 0$. Thus the left-hand side of (3.10) is bounded above by $(\delta_N/N)Nt$, which goes to zero as $N \to \infty$. Hence (3.10) holds. By the weak convergence of $\{S^N(\cdot)\}$ to the function with value μt at t and the continuity of the limits of $\tilde{A}^N(\cdot)$, the limits of $\tilde{A}_0^N(\cdot)$ have continuous paths w.p.1.

It will be shown below that the mean-square values of $X^N(t)$ are bounded uniformly in N, t. Assuming this fact in this paragraph, we see that the term $X^N(s)/\sqrt{N}$ disappears from (3.6) as $N \to \infty$. The limit of any weakly convergent subsequence of $\{\tilde{D}^N(\cdot)\}$ is a martingale (all the appropriate filtrations will be identified at the end of the proof). Since $\tilde{A}_0^N(\cdot)$ is a martingale and $S^N(\cdot)$ converges as stated, the weak limits of $\tilde{A}_0^N(\cdot)$ are also martingales. By the last line of (3.9b), the limit of any weakly convergent subsequence of $\{\tilde{A}_0^N(\cdot), \tilde{D}^N(\cdot)\}$ is orthogonal. By (3.9a) together with (3.6) (with $X^N(s)/\sqrt{N}$ dropped), (2.3), the convergence of $S^N(\cdot)$, and the continuity of the limits, the limit martingales have the quadratic variations asserted in the theorem; hence they are Wiener processes.

Boundedness of $\{EX^N(t)^2\}$. Suppose that $\sup_N E|X^N(0)|^2 < \infty$. Let $t_0^N = 0$, and let $\{t_i^N, i > 0\}$ denote the ordered sequence of event times (times of arrivals or departures). Then

$$\begin{aligned}[][X^N(t_{i+1}^N)]^2 - [X^N(t_i^N)]^2 &= 2X^N(t_i^N)(X^N(t_{i+1}^N) - X^N(t_i^N)) \\ &\quad + (X^N(t_{i+1}^N) - X^N(t_i^N))^2.\end{aligned}$$

Note that if a rejection of an arrival occurs at t_{i+1}^N, then $X^N(t_i^N)$ must equal 0. Thus $2X^N(t_i^N)(Y^N(t_{i+1}^N) - Y^N(t_i^N)) = 0$. Note that $X^N(t_{i+1}^N) - X^N(t_i^N) = \pm 1/\sqrt{N}$.

Using these observations, (3.7), and the martingale property, for any $t \geq 0$ and $\tau > 0$ the previous expression leads to

$$\begin{aligned} EX^N(t+\tau)^2 &= EX^N(t)^2 - 2\mu \int_t^{t+\tau} E(X^N(s))^2 ds + 2b_1 E \int_t^{t+\tau} X^N(s) ds \\ &\quad + g^N(t, t+\tau) + \delta_N, \end{aligned} \tag{3.11}$$

where $\delta_N \to 0$ uniformly in (t, τ) as $N \to \infty$ and $g^N(t, t+\tau)$ is bounded in absolute value by

$$\frac{3}{N} E\,(\text{number of arrivals and departures on } [t, t+\tau])\,.$$

This last expression is $O(\tau) + O(1/N)$. The $O(\cdot)$ functions might differ from case to case, but they will always be uniform in the sense that there is $k < \infty$ such that in all cases, for $\alpha \geq 0$, $|O(\alpha)| \leq k\alpha$. For any $c_1 \in (0, \mu)$ there is $c_2 > 0$ such that $2b_1|x| \leq c_2 + c_1 x^2$. Using this bound and the fact that $X^N(s) \geq 0$ in (3.11), redefining δ_N to account for the $O(1/N)$ and writing $m^N(t) = EX^N(t)^2$ yields that there is $c_3 > 0$ such that

$$m^N(t+\tau) \leq m^N(t) - \mu \int_t^{t+\tau} m^N(s)ds + c_3\tau + \delta_N, \tag{3.12a}$$

$$m^N(t+\tau) \geq m^N(t) - 2\mu \int_t^{t+\tau} m^N(s)ds - c_3\tau + \delta_N. \tag{3.12b}$$

The nonnegativity of $m^N(t)$ and (3.12a) yield

$$m^N(t+\tau) \leq m^N(t) + c_3\tau + \delta_N. \tag{3.12c}$$

Substituting $m^N(s) = m^N(t) + [m^N(s) - m^N(t)]$ in (3.12b) and using (3.12c) yield

$$m^N(t+\tau) \geq (1 - 2\mu\tau)m^N(t) + O(\tau) + \delta_N. \tag{3.12d}$$

Now, write $m^N(s) = m^N(t) + [m^N(s) - m^N(t)]$ in (3.12a) and use (3.12d) to get

$$m^N(t+\tau) \leq (1 - \mu\tau + O(\tau^2))m^N(\tau) + O(\tau) + \delta_N. \tag{3.13a}$$

Let $\Delta > 0$ be small and N large enough so that $|\delta_N| \leq \Delta$. Then by (3.13a) there is positive c_4 such that

$$m^N(t+\Delta) \leq m^N(t)\left[1 - \mu\Delta + O(\Delta^2)\right] + c_4\Delta. \tag{3.13b}$$

Now, letting $t = n\Delta$ in (3.13b) and iterating yield $\sup_n m^N(n\Delta) < \infty$ for large N. Combining this with (3.12c) yields that, for some $c_5 < \infty$ and large N,

$$\sup_t m^N(t) \leq c_5. \tag{3.14}$$

A similar proof yields that, for each $T < \infty$,

$$\sup_N E \max_{t \leq T} X^N(t)^2 < \infty. \tag{3.15}$$

Tightness of $\{X^N(\cdot)\}$. It follows from (3.15), the properties of $\tilde{A}_0^N(\cdot)$ and $\tilde{D}^N(\cdot)$, and the form (3.7) that $\sup_N E|Y^N(T)|^2 < \infty$ for each $T < \infty$. Thus, to prove tightness of $\{Y^N(\cdot)\}$ via (3.1) it is enough to show that

$$\lim_{\Delta \to 0} \sup_N \sup_{\tau \leq T} P\{Y^N(\tau + \Delta) - Y^N(\tau) \geq \varepsilon\} = 0 \tag{$*$}$$

for each $\varepsilon > 0, T < \infty$, where τ are stopping times. It is sufficient to use τ such that $X^N(\tau) = 0$. Recalling the definition of $Z^N(\cdot)$ in (3.7) and the definition of $Y^N(\cdot)$, we have

$$P\{Y^N(\tau+\Delta) - Y^N(\tau) \geq \varepsilon\} \leq P\left\{\sup_{s \leq \Delta}|Z^N(\tau+s) - Z^N(\tau)| \geq \varepsilon\right\}.$$

Now, the properties of $Z^N(\cdot)$ imply (*). Then $\{Y^N(\cdot)\}$ is tight. From here, it is straightforward to show tightness of $\{X^N(\cdot)\}$, and we omit the details.

The nonanticipativeness property. Let $t \geq 0, s > 0$; and let $\{t_i\}$ be real numbers no bigger than t. For an integer k, let $h(\cdot)$ be a real-valued, bounded, and continuous function of $\{X^N(t_i), \tilde{A}_0^N(t_i), \tilde{D}^N(t_i), Y^N(t_i), i \leq k\}$. Then

$$Eh(X^N(t_i), \tilde{A}_0^N(t_i), \tilde{D}^N(t_i), Y^N(t_i), i \leq k)[\tilde{D}^N(t+s) - \tilde{D}^N(t)] = 0.$$

Let $(X(\cdot), W_1(\cdot), W_2(\cdot), Y(\cdot))$ denote the limit of a weakly convergent subsequence. Then by the weak convergence

$$Eh(X(t_i), W_1(t_i), W_2(t_i), Y(t_i), i \leq k)[W_2(t+s) - W_2(t)] = 0. \tag{3.16}$$

Let $\mathcal{B}_t$ denote the σ-algebra induced by $\{X(s), W_1(s), W_2(s), Y(s), s \leq t\}$. Since $k, \{t_i\}, t, s, h(\cdot)$ are arbitrary, (3.16) implies that $W_2(\cdot)$ is a $\mathcal{B}_t$-Wiener process. A similar proof is used for $W_1(\cdot)$. This yields the nonanticipativeness assertion. □

DEFINITION. Stationary solution. *A solution of* (3.8) *is said to be* stationary *if the distribution of* $\{X(t+\cdot), W_1(t+\cdot) - W_1(t), W_2(t+\cdot) - W_2(t), Y(t+\cdot) - Y(t)\}$ *does not depend on* t. *For a sequence of real numbers* $\{t_N\}$, *define the "shifted" processes*

$$H^N(\cdot) = (X^N(t_N + \cdot), \tilde{A}_0^N(t_N + \cdot) - \tilde{A}_0^N(\cdot), \tilde{D}^N(t_N + \cdot) - \tilde{D}^N(t_N),$$

$$\tilde{Y}^N(t_N + \cdot) - Y^N(t_N)).$$

The following theorem will not be used in the sequel but is of interest if the process and the approximations are of concern over a long time period.

THEOREM 3.2. *Assume the conditions of Theorem* 3.1. *Let* t_N *be a sequence of real numbers tending to infinity. Then the set of "shifted" processes* $\{H^N(\cdot)\}$ *is tight. It converges weakly to the unique stationary solution* $(X(\cdot), W_1(\cdot), W_2(\cdot), Y(\cdot))$ *of* (3.8). *The process* $X(\cdot)$ *satisfying* (3.8) *is a strong Feller process.*

Proof. The tightness follows from Theorem 3.1 and the mean square boundedness of $\{X^N(s),\ s < \infty, N < \infty\}$. By Theorem 3.1, any subsequence of $\{H^N(\cdot)\}$ has a further subsequence that converges weakly to a solution of (3.8). Next, fix $T \in (0, \infty)$ and extract a weakly convergent subsequence of $\{X^N(t_N + \cdot), X^N(t_N - T + \cdot)\}$, with the limit denoted by $(\tilde{X}(\cdot), \tilde{X}_T(\cdot))$. Both of these processes satisfy equation (3.8) for some Wiener processes, and $\tilde{X}_T(T) = \tilde{X}(0)$. Also, for each initial condition x, the solution process $X(t)$ of (3.8) is bounded in mean, uniformly in t. It can be shown either directly or via a Girsanov transformation technique (starting with the system $X(t) = X(0) + W_1(t) + W_2(t) + Y(t)$) that the solution of (3.8) is a strong Feller process and that for $t > 0$ it has a transition density $p(x, t, y)$ that is nonzero for all $x \in [0, \infty)$ and $y \in [0, \infty)$. The last two sentences imply that there is a unique invariant measure $\pi(\cdot)$ and [3, Thm. 4]

$$\int f(y)p(x,t,y)dy \to \int f(y)\pi(dy) \tag{3.17}$$

for each x and bounded continuous real-valued $f(\cdot)$. The set of all possible values of $\{\tilde{X}_T(0), T < \infty\}$ is tight since $\{X^N(t); N < \infty, t < \infty\}$ is tight. Owing to this tightness and the arbitrariness of T, to show that $\tilde{X}(0)$ is the stationary random variable it is enough to show that the convergence in (3.17) is uniform in each compact x-set K. Given $K, t_1 > 0$, and $\delta > 0$, there is a compact set K_1 such that $\int_{K_1} p(x, t_1, y)dy \geq 1 - \delta$ for all $x \in K$. Now, let $t > t_1$ and use the representation

$$\int p(x, t_1, y_1)dy_1 \int f(y_2)p(y_1, t - t_1, y_2)dy_2, \qquad t > t_1,$$

of the left-hand side of (3.17), the arbitrariness of δ, and the pointwise convergence (3.17), to get the desired uniform convergence. □

4. A one-dimensional canonical control problem. In addition to the primary input source $A^N(\cdot)$ of §2, let there be an "exogenous" source of competing requests for the use of the N individual lines that arrive one at a time. The one-at-a-time assumption is not required, but it does save on the notation. In particular, define $G^N(t) =$ [number of requests from the exogenous source by time $t]/\sqrt{N}$. The interarrival times for this source might be correlated, and in this section we assume either

(a) that there is $b_0 > 0$ such that for any positive T_1

$$[G^N(t+T) - G^N(T)] \to b_0 t \tag{4.1a}$$

uniformly for $t \le T_1, T < \infty$, or, more generally;

(b) that $\{G^N(\cdot)\}$ is tight and converges weakly to a continuous process $G(\cdot)$. Also, suppose that the set

$$\{G^N(n+1) - G^N(n); N < \infty, n < \infty\} \tag{4.1b}$$

is uniformly integrable.

We also assume that $G^N(\cdot)$ is independent of $A^N(\cdot)$, $X^N(0)$, and the sequence of service intervals. Equation (4.1a) is similar to the "fluid" model used in [25], [26], [20], [21]. Commonly, the $G^N(\cdot)$ process is an "overflow" from another system and (4.1a) would not hold since $G^N(\cdot)$ would behave like our $Y^N(\cdot)$. The second case covers this. We note that (4.1a) might hold "approximately," if $G^N(\cdot)$ were the sum of inputs from many independent sources. The uniform integrability in condition (4.1b) holds for overflows from "trunk line" systems, as seen by the uniform integrability of the set (4.7) below.

The weak convergence assumed in (b) above implies that the mean rate of the exogenous source is $O(\sqrt{N})$. The basic control question is whether or not to accept any particular request from the exogenous sequence. Arrivals from the original sequence $A^N(\cdot)$ are always accepted if any line is available. If the order of the number of exogenous inputs per unit time were larger than $O(\sqrt{N})$, then the fraction rejected would go to unity as $N \to \infty$. Let $J^N(\cdot)$ denote the *acceptance process* defined as follows: If an exogenous arrival that occurs at time t is accepted, then $J^N(t) = 1$; otherwise $J^N(t) = 0$. Then the (scaled by $1/\sqrt{N}$) number of acceptances from the exogenous source by time t can be written as (note the definition of the Stieltjes integral)

$$\sum_{s\le t} J^N(s)[G^N(s) - G^N(s^-)] = \int_0^t J^N(s)dG^N(s) \equiv F^N(t).$$

Now we can write

$$\begin{aligned} X^N(t) &= X^N(0) + b_1 t - \mu\int_0^t X^N(s)ds \\ &\quad + \tilde{A}_0^N(t) + \tilde{D}^N(t) - F^N(t) + Y^N(t) + O_N(t)\sqrt{N}. \end{aligned} \tag{4.2}$$

The $O_N(\cdot)$ is the same as in (3.7).

A cost function. An average cost per unit time criterion will be dealt with in §5. Here, we formulate a "discounted" criterion. Recall that $Y^N(t)$ equals the (scaled) number of rejects from the primary source, due to arrivals when all lines are occupied. Let $c_1 > c_0 > 0$, $\beta > 0$, and define the cost

$$C^N(J^N, X^N(0)) = E\int_0^\infty e^{-\beta t}[c_1 dY^N(t) + c_0(1 - J^N(t))dG^N(t)]. \tag{4.3}$$

Thus c_1 (c_0, resp.) is the per unit cost of rejecting a request from the primary (exogenous, resp.) source. Redefine $\mathcal{B}_t^N$ so that it measures $\{G^N(s), F^N(s), A^N(s), D^N(S), X^N(0), s \le t\}$. Thus, $\mathcal{B}_t^N$ also measures $\{J^N(s), s \le t\}$. We say that $J^N(\cdot)$ is *admissible* if the following holds: $J^N(t) = 0$ if there is no exogeneous arrival at t; for t at which there is an exogeneous arrival, $J^N(t)$ depends only on the arrival and departure data up to and including t and on the acceptance data to t^-. Since the events occur discretely in time, $J^N(\cdot)$ is well defined.

THEOREM 4.1. *Assume the conditions of Theorem* 3.1 *and the assumptions on* $G^N(\cdot)$ *of this section. Let the* $J^N(\cdot)$ *be admissible. Then* $\{X^N(\cdot), \tilde{A}_0^N(\cdot), \tilde{D}^N(\cdot), Y^N(\cdot), F^N(\cdot), G^N(\cdot)\}$ *is tight. If* $(X(\cdot), W_1(\cdot), W_2(\cdot), Y(\cdot), F(\cdot), G(\cdot))$ *denotes the limit of a weakly convergent subsequence, then*

$$X(t) = X(0) + b_1 t - \mu\int_0^t X(s)ds + W_1(t) + W_2(t) - F(t) + Y(t), \tag{4.4}$$

where the $W_1(\cdot)$ *are as in Theorem* 3.1. *All other processes are continuous and nonanticipative with respect to the* $W_i(\cdot)$, *and* $F(\cdot)$ *has the form* (*under* (4.1a))

$$F(t) = b_0\int_0^t J(s)ds \tag{4.5a}$$

for some nonanticipative process $J(\cdot)$ *with* $J(t) \in [0,1]$. *More generally* (*under condition* (4.1b) *above*),

$$F(t) = \int_0^t J(s)dG(s). \tag{4.5b}$$

Also, if N *indexes the convergent subsequence, then* $C^N(J^N, X^N(0)) \to C(J, X(0))$, *where*

$$C(J, X(0)) = E\int_0^\infty e^{-\beta t}[c_1 dY(t) + c_0(1 - J(t))dG(t)]. \tag{4.6}$$

The set $\{EX^N(t)^2; N < \infty, t < \infty\}$ *is bounded.*

Remark. Note that $J(t)$ can be any value in $[0,1]$. The quantity $b_0 J(t)$ in (4.5a) can be viewed as the "local intensity" of the acceptance process for the exogenous requests at time t, in the limit. In (4.5b), $J(t)$ can be viewed as the "local probability" of acceptances at t when there are exogenous inputs. In typical applications where there is input control, the acceptance "local probability" is either zero or unity (i.e., reject or accept), but the theoretical development here requires that we allow the possibility of arbitrary values in $[0,1]$.

Proof. The boundedness of $\{EX^N(t)^2; N < \infty, t < \infty\}$ follows from Theorem 3.1, since the expectations are no larger when the exogenous inputs are introduced. Actually, all the assertions, except those concerning $F(\cdot)$ and the convergence of the costs, are proved in Theorem 3.1. $\{F^N(\cdot)\}$ is obviously tight by the assumed tightness of $\{G^N(\cdot)\}$. The nonanticipativeness property of $(X(\cdot), Y(\cdot), F(\cdot), G(\cdot))$ follows from the arguments used in connection with (3.16) and the independence assumptions put on $G^N(\cdot)$, since (3.16) continues to hold if $F(t_i), G(t_i)$ are added to the arguments of $h(\cdot)$. Equation (4.1a) implies that $F(\cdot)$ is absolutely continuous with respect to Lebesgue measure, with derivative bounded by b_0, hence the representation (4.5a). Analogously, (4.5b) holds since $F(\cdot)$ is always absolutely continuous with respect to $G(\cdot)$ with derivative in $[0,1]$.

If the set of increments (4.7)

$$\{Y^N(n+1) - Y^N(n); n < \infty, N < \infty\} \tag{4.7}$$

is uniformly integrable, then the weak convergence implies the convergence of the costs $C^N(J^N, X^N(0))$ to $C(J, X(0))$. To prove the uniform integrability, write

$$\begin{aligned} Y^N(n+1) \; - Y^N(n) &= (X^N(n+1) - X^N(n)) + \mu \int_n^{n+1} X^N(s)ds - b_1 \\ &\quad - (\tilde{A}_0^N(n+1) - \tilde{A}_0^N(n)) - (\tilde{D}^N(n+1) - \tilde{D}^N(n)) \\ &\quad + (F^N(n+1) - F^N(n)) - O_N(t)\sqrt{N}. \end{aligned}$$

Using the uniform integrability of the set (4.1b), the boundedness of the $EX^N(t)^2$, in N and t, and the fact that the sup over N and n of the squares of the last four terms on the right are bounded, we get the uniform integrability of (4.7). □

Remark. The above development can clearly be extended to multiple exogenous sources $G_1^N(\cdot), G_2^N(\cdot), \ldots$, each with its own cost and "acceptance" control.

Uniqueness. We say that the solution to system (4.4) *is unique in the weak sense* if the probability law of $(W_1(\cdot), W_2(\cdot), F(\cdot), X(0))$ determines the probability law of $(X(\cdot), W_1(\cdot), W_2(\cdot), F(\cdot))$. Weak sense uniqueness can easily be proved via the Girsanov measure transformation method [8, Chap. 4.4], for any $F(\cdot)$ of the form in (4.5b). To see this, let $W(\cdot)$ be a Wiener process and $G(\cdot)$ a nonnegative and nonanticipative (with respect to $W(\cdot)$) process with nondecreasing and continuous paths with $G(0) = 0$. Let $J(\cdot)$ be nonanticipative (with respect to $W(\cdot)$) and bounded. Consider the equation

$$X(t) = X(0) + W(t) - \int_0^t J(s)dG(s) + Y(t), \qquad X(t) \geq 0,$$

where $Y(\cdot)$ is the reflection term that is allowed to increase only when $X(t) = 0$. The equation has a weak sense unique solution. The Girsanov transformation method can now be used to add the drift terms $b_1 t$ and $-\mu \int_0^t X(s)ds$, exactly as done in [8, Chap. 4.4], and get the weak sense uniqueness for (4.4).

5. An ergodic control problem.

We continue with the controlled single trunk problem of the last section. Let $G^N(\cdot)$ be a process satisfying the conditions of §4. For $T > 0$, define the cost (for $x = X^N(0)$ = initial condition)

$$\gamma_T^N(J^N, x) = \frac{1}{T} E \int_0^T \left[c_1 dY^N(t) + c_0(1 - J^N(t))dG^N(t)\right]. \tag{5.1}$$

In this and in §§9 and 10, it will be shown that *stationary* controlled processes satisfying (4.4) provide very good approximations to the performance of $X^N(\cdot)$ for large

N and time. The trunk line systems tend to be of interest for a long time period, and virtually all of the current analyses assume some sort of stationarity. Here we concern ourselves with the ergodic cost problem for the single trunk case, but with an *arbitrary starting state.* Sections 9 and 10 are concerned with the ergodic control problem for the network. The very useful "occupation measure" method will be used, although we work with measures over the path space rather than over the value spaces of the state and control functions. We first introduce the needed notation.

DEFINITIONS. Stationary controlled process. *We say that the solution to* (4.4) *is* stationary *if the distribution of*

$$\begin{aligned} R_t(\cdot) = (&X(t+\cdot), W_1(t+\cdot) - W_1(t), W_2(t+\cdot) - W_2(t),\\ &F(t+\cdot) - F(t), Y(t+\cdot) - Y(t), G(t+\cdot) - G(t)) \end{aligned}$$

does not depend on t.

Assumptions and notation. Now we set up some notation for the functional occupation measure development. We need to define the canonical variables of the sample space of the above set of the six processes that are the components of $R_t(\cdot)$. We will use the 6-tuple $\rho(\cdot) = (\xi(\cdot), \psi_1(\cdot), \psi_2(\cdot), \phi(\cdot), y(\cdot), \alpha(\cdot))$ to denote the canonical element of $D^6[0,\infty)$, where each of the six components is a canonical element of $D[0,\infty)$. The element $\xi(\cdot)$ will represent the canonical sample path of either $X(\cdot)$ or $X^N(\cdot)$ or of appropriate "time-shifted" forms such as $X^N(t+\cdot)$. Similarly, $\psi_1(\cdot)$ will denote the canonical sample path of either $W_1(\cdot)$ or $\tilde{A}_0^N(\cdot)$ or of appropriate time-shifted and centered forms such as $\tilde{A}_0^N(t+\cdot) - \tilde{A}_0^N(t)$. Analogously, $\psi_2(\cdot), \phi(\cdot), y(\cdot)$, and $\alpha(\cdot)$ will denote (resp.) the canonical sample paths of either $W_2(\cdot)$ or $\tilde{D}^N(\cdot)$, either $F(\cdot)$ or $F^N(\cdot)$, either $Y(\cdot)$ or $Y^N(\cdot)$, and either $G^N(\cdot)$ or $G(\cdot)$, or appropriately time-shifted and centered forms of these processes. For $t \geq 0$, define the processes $R_t^N(\cdot)$ by

$$\begin{aligned} R_t^N(\cdot) = (&X^N(t+\cdot), \tilde{A}_0^N(t+\cdot) - \tilde{A}_0^N(t), \tilde{D}^N(t+\cdot) - \tilde{D}^N(t),\\ &F^N(t+\cdot) - F^N(t), Y^N(t+\cdot) - Y^N(t), G^N(t+\cdot) - G^N(t)). \end{aligned}$$

Effectively, $R_t^N(\cdot)$ are the original processes but shifted left by t. Let $P^{N,t}(\cdot)$ denote the measure that is induced by the process $R_t^N(\cdot)$, and define the *occupation measure* $P_T^N(\cdot)$:

$$P_T^N(\cdot) = \frac{1}{T}\int_0^T P^{N,t}(\cdot)dt.$$

Occupation measures on the path space provide a convenient method of dealing with approximations to ergodic cost problems in a variety of contexts [14].

The cost function. Let us write the cost function (5.1) in terms of $P_T^N(\cdot)$. Note that

$$\frac{1}{T}\int_0^T [F^N(t+1) - F^N(t)]dt = \frac{1}{T}F^N(T) + \delta_T^N, \tag{5.2a}$$

where

$$\delta_T^N = \frac{1}{T}\left[\int_T^{T+1} (F^N(t) - F^N(T))dt - \int_0^1 F^N(t)dt\right]. \tag{5.2b}$$

Thus, we can write

$$\begin{aligned}\gamma_T^N(J^N, X^N(0)) &= E\left[c_1Y^N(T) + c_0(G^N(T) - F^N(T))\right]/T \\ &= \frac{1}{T}E\int_0^T [c_1(Y^N(t+1) - Y^N(t)) + c_0(G^N(t+1) - G^N(t)) \\ &\quad - c_0(F^N(t+1) - F^N(t))]dt + \hat{\delta}_T^N,\end{aligned}$$

where the "error term" $\hat{\delta}_T^N$ can be obtained from the type of calculation done in connection with (5.2). Under the uniform integrability properties of (4.1b) and (4.7), $\hat{\delta}_T^N \to 0$ as $T \to 0$, uniformly in N. Now we can write (5.1) in terms of the occupation measure as

$$\gamma_T^N(J^N, X^N(0)) = \int [c_1 y(1) + c_0(\alpha(1) - \phi(1))]\, P_T^N(d\rho) + \hat{\delta}_T^N. \tag{5.3}$$

Thus, the asymptotic behavior of the costs depends on the asymptotic behavior of $\{P_T^N(\cdot)\}$. The next theorem tells us that the long-term averages are well approximated by those of a stationary limit problem. We note that the occupation measure method used in Theorem 5.1 will also be used in §§9 and 10 in essentially the same way. It is easier to understand in the context of the one-dimensional problem here. The next theorem justifies using the stationary limit problem to get approximations to the physical problems when the latter is of interest over a long time period.

THEOREM 5.1. *Assume the conditions of Theorem* 3.1, *as well as the assumptions on* $G^N(\cdot)$ *stated in* §4. *Let the* $J^N(\cdot)$ *be admissible. Then* $\{P_T^N(\cdot); N < \infty, T < \infty\}$ *is tight. Let* $\overline{P}(\cdot)$ *denote the limit of a weakly convergent subsequence* (*as* $N \to \infty, T \to \infty$). *There exists a stationary process, which we denote by* $\overline{R}(\cdot) = (X(\cdot), W_1(\cdot), W_2(\cdot), F(\cdot), Y(\cdot), G(\cdot))$, *that induces the measure* $\overline{P}(\cdot)$ *and satisfies* (4.4), *where* $F(\cdot)$ *has the representation* (4.5b) *for a nonanticipative process* $J(\cdot)$ *with values in* $[0,1]$. *The Wiener processes* $W_i(\cdot)$ *have the quadratic variation properties cited in Theorem* 3.1, *and the other processes are nonanticipative with respect to the Wiener processes. Also, as* $N \to \infty$ *and* $T \to \infty$ *through the convergent subsequence,*

$$\begin{aligned}\gamma_T^N(J^N, X^N(0)) \to \gamma(J) &\equiv \textstyle\int \overline{P}(d\rho)\, [c_1 y(1) + c_0(\alpha(1) - \phi(1))] \\ &= E\left[c_1 Y(1) + c_0\int_0^1 [1 - J(s)]dG(s)\right].\end{aligned} \tag{5.4}$$

Proof. By Theorem 4.1, the set of random variables $\{X^N(t); N < \infty, t < \infty\}$ is tight. By the proof in Theorem 4.1, this implies that the set of processes $\{R_t^N(\cdot); N < \infty, t < \infty\}$ is tight; i.e., the set of measures $\{P^{N,t}(\cdot); N < \infty, t < \infty\}$ is tight. The latter assertion implies the tightness of the set of occupation measures $\{P_T^N(\cdot); N < \infty, T < \infty\}$.

Now, abusing terminology, let N, T (both going to infinity) index a weakly convergent subsequence with limit denoted by $\overline{P}(\cdot)$. Let $\overline{R}(\cdot) = (\overline{X}(\cdot), \ldots, \overline{G}(\cdot))$ denote the process that induces $\overline{P}(\cdot)$. Note that the measure $P_T^N(\cdot)$ is that of a process $\tilde{R}_T^N(\cdot)$ that is constructed exactly as $R^N(\cdot) = R_0^N(\cdot)$ is but where one *randomizes* the initial time; i.e., $P\{\text{initial time} \in [t, t+\Delta]\} = \Delta/T$ for $0 \le t \le t + \Delta \le T$, and the initial time is independent of all other random variables in the system. By the tightness of the set $\{X^N(t); N < \infty, t < \infty\}$, the set of initial conditions for the "randomized initial time" processes $\{\tilde{R}_T^N(\cdot); N < \infty, T < \infty\}$ is tight. Then the fact that $\overline{R}(\cdot)$ satisfies (4.4) with the representation (4.5b) follows from the proofs of Theorems 3.1 and 4.1.

The paths of $\overline{R}(\cdot)$ are continuous w.p.1. Hence the functions with values $y(1), \alpha(1)$, and $\phi(1)$ are continuous in the Skorohod topology almost everywhere with respect to the measure of $\overline{R}(\cdot)$. Thus [1, Thm. 5.1] the convergence (5.4) follows from the weak convergence, the continuity of the paths of $\overline{R}(\cdot)$, the uniform integrability of the set $\{Y^N(n+1) - Y^N(n); N < \infty, n < \infty\}$ (see Theorem 4.1), and the assumed uniform integrability of $\{G^N(n+1) - G^N(n); N < \infty, n < \infty\}$. We need only prove the stationarity properties of $\overline{R}(\cdot)$.

For $g(\cdot) \in D^6[0,\infty)$ and $c > 0$, define the left shift $g_c(\cdot) = g(c + \cdot)$. For a Borel set $K \subset D^6[0,\infty)$, define the left shift $K_c = \{g(\cdot) : g_c(\cdot) \in K\}$. Then we can write

$$\begin{aligned} P_T^N(K_c) &= \frac{1}{T}\int_0^T P^{N,t}(K_c)dt = \frac{1}{T}\int_0^T P\{R_{t+c}^N(\cdot) \in K\}dt, \\ P_T^N(K_c) - P_T^N(K) &= \frac{1}{T}\int_T^{T+c} P\{R_{t+c}^N(\cdot) \in K\}dt - \frac{1}{T}\int_0^c P\{R_t^N(\cdot) \in K\}dt. \end{aligned} \tag{5.5}$$

Let $f(\cdot)$ be a bounded continuous real-valued function on $D^6[0,\infty)$. Then the "error" estimate (5.5) and the continuity of the limit process $\overline{R}(\cdot)$ imply that

$$\begin{aligned} Ef(\overline{R}(\cdot)) &= \lim_{N,T}\int f(\rho)P_T^N(d\rho) = \int f(\rho)\overline{P}(d\rho) \\ &= \lim_{N,T}\int f(\rho_c)P_T^N(d\rho) = Ef(\overline{R}_c(\cdot)). \end{aligned}$$

Since $c > 0$ is arbitrary, this last equation implies the stationarity. □

It can be shown that $\inf_J \gamma^N(J^N, x) \to \inf_J \gamma(J)$. The details are omitted since a harder problem is dealt with in Theorem 9.1.

6. Convergence and approximation of the controls: Discounted cost. We return to the discounted cost problem of §4. This section concerns convergence of the optimal costs for the physical problem to that for the limit problem. This topic is continued in §8, where the more complex network case of §7 is dealt with, and given in more detail.

DEFINITION. *A control $J(\cdot)$ for* (4.4) *is said to be* admissible *if it is a measurable process that takes values in* [0, 1] *and is nonanticipative with respect to the $W_i(\cdot)$. We were able to represent the limit $F(\cdot)$ in Theorems* 4.1 *and* 5.1 *in terms of an admissible control, since both $F(\cdot)$ and $G(\cdot)$ were nonanticipative with respect to the $W_i(\cdot)$. Define*

$$\overline{C}^N(x) = \inf\{C^N(J^N, x) : J^N(\cdot) \text{ admissible}\},$$

$$\overline{C}(x) = \inf\{C(J, x) : J(\cdot) \text{ admissible}\},$$

where $C^N(\cdot)$ and $C(\cdot)$ are defined by (4.3) *and* (4.6), *respectively. We wish to show next that*

$$\overline{C}^N(x) \to \overline{C}(x). \tag{6.1}$$

Note that no Markovian-type assumption is made on $G^N(\cdot)$ or on $G(\cdot)$.

THEOREM 6.1. *Assume the conditions of Theorem* 4.1. *Then* (6.1) *holds.*

Discussion of the proof. Theorem 4.1 implies that

$$\liminf_N \overline{C}^N(x) \geq \overline{C}(x).$$

Thus, we need only prove the "reverse" inequality

$$\limsup_N \overline{C}^N(x) \leq \overline{C}(x). \tag{6.2}$$

This is the usual problem in getting results like (6.1). In order to prove (6.2), one can proceed as follows. Let $\epsilon > 0$. A particular ϵ-optimal control (denoted by $J^\epsilon(\cdot)$) for the system $X(\cdot)$ of (4.4), (4.5b), and cost (4.6) is obtained. This control must be such that there is an adaptation, to be denoted by $J^{\epsilon,N}(\cdot)$, that can be used on the physical processes $X^N(\cdot)$. The adaptation should be such that, under $J^{\epsilon,N}(\cdot)$,

$$C^N(J^{\epsilon,N}, x) \to C(J^\epsilon, x).$$

Since $J^\epsilon(\cdot)$ is ϵ-optimal, we will then have

$$\limsup_N \bar{C}^N(x) \leq \lim_N C^N(J^{\epsilon,N}, x) = C(J^\epsilon, x) \leq \bar{C}(x) + \epsilon.$$

Since ϵ is arbitrary, this proves (6.2). The control $J^\epsilon(\cdot)$ is not to be considered to be a "practical" control. Its purpose is only to help prove (6.2).

A similar problem occurs in the proofs of the convergence of "computational approximations" to optimal stochastic control problems, and, indeed, the method of [13, Thm. 7.1] can be used to construct the desired $J^\epsilon(\cdot)$. In fact, the only difference between the requirements in [13, Thm. 7.1] and our requirements here is the need to take the input $G(\cdot)$ into account. The adaptation of the cited proof to our problem is not hard. But since an analogous problem for the more complicated controlled network will be dealt with in detail in §8 via a related approximation method, we will not pursue the matter further here but refer the reader to §8. This type of result is an important consequence of the heavy traffic analysis, since useful numerical methods are available to get the optimal value functions and controls for control problems for the limit models [13], [15], [16].

DEFINITION. *A control $J(\cdot)$ for $X(\cdot)$ is said to be a* feedback *control if there is a measurable function $u(\cdot)$ with values in* $[0, 1]$ *and such that $J(t) = u(X(t), t)$. If $u(\cdot)$ depends only on x, we say that it is a* state feedback *control. For feedback controls, we write the associated costs as $C^N(u, x), C(u, x)$. If $G(t)$ equals b_0t (as in* (4.5a)) *or is the "overflow" from another trunk line system, then one might expect $J(\cdot)$ to be of state feedback form and, more particularly, to be of the "threshold" form; namely, there is $B_0 > 0$ such that $J(t) = 1$ if $X(t) \geq B_0$ and $J(t) = 0$ otherwise. Such threshold policies are a common occurrence for queueing problems when the input can be controlled* [9], [22]. *For the network problem the optimal control will not be of the threshold type.*

For the next result, we need the following construction. Let $u(\cdot)$ be a state feedback control that is piecewise continuous in x. We can define a control $J^N(\cdot)$ from $u(\cdot)$ for use in (4.2) such that $C^N(J^N, x) \to C(J, x)$. If $u(\cdot)$ is an indicator function, then the adaptation is obvious; we use $J^N(t) = u(x^N(t))$. Otherwise, we use a randomization rule of the type that will be used in Theorem 8.3 below and where more detail will be given: In particular, use the following rule. Let there be an arrival from $G^N(\cdot)$ at t; then accept it with the conditional probability

$$P\{\text{accept} \mid \text{data}(t)\} = u(X^N(t^-)),$$

where data(t) is all the systems data up to and including time t, except for the acceptance or rejection of the request from $G^N(\cdot)$ at t. The weak convergence arguments can be carried through since $u(\cdot)$ is continuous w.p.1 with respect to the measure of $X(\cdot)$.

We can state the following extension of Theorem 6.1. Again, we omit the proof in view of the development in §8. The corollary asserts that a "nice" nearly optimal control for $x(\cdot), C(\cdot)$ is also nearly optimal for $X^N(\cdot), C^N(\cdot)$. If $u(\cdot)$ is of the threshold type, then its adaptation for use on the network does not require a randomization.

COROLLARY. *Assume the conditions of Theorem* 4.1. *Suppose that $u(\cdot)$ is a piecewise continuous state feedback control for $X(\cdot)$. Let $J^N(\cdot)$ denote its adaptation to $X^N(\cdot)$.*

Then $C^N(J^N, x) \to C(u, x)$. If $u(\cdot)$ is ε-optimal for (4.4), (4.6), *then $J^N(\cdot)$ will be 2ε-optimal for* (4.2), (4.3), *for large N.*

7. A simple controlled network. Description. The network will be defined and the basic weak convergence results of the past sections will be extended. It will be shown that the limit processes lead to well-defined control problems and that the costs converge to the cost for the limit problem. We consider a three-node network (i.e., a triangle). However, the methods and results hold for arbitrary networks provided that the basic heavy traffic conditions hold. The exogenous requests $G^N(\cdot)$ in the previous sections are now replaced by explicit requests from other trunk lines of the network. Numerical studies of the triangular network have shown how to devise very good and relatively easily implementable policies for arbitrary (hundreds of trunks) networks. Details of the numerical method and the simulation studies are given in [18].

Let the nodes be labelled A, B, C and the links $1, 2, 3$ with link 1 connecting (A, B) and links 2, 3, respectively, connecting (B, C) and (C, A). The trunk lines labelled $i = 1, 2, 3$ have $\beta_i N$ individual bidirectional lines, where $\beta_i > 0$ and N is a size parameter. The service time distribution of each individual line is exponential with rate $\mu > 0$. External arrivals appear at the nodes A, B, C. An arrival at A can be a request for service to either B or C. As far as the trunk line i is concerned, it does not matter at which end the service request originates. Thus, we use the following model for requests for service on trunk line i. Let $\{\alpha_k^{i,N}, k < \infty\}$ denote the sequence of interarrival times for trunk i, and suppose that they are independent and identically distributed and independent of all other service and interarrival times of the system. The set of all service and interarrival times are assumed to be independent.

Let $Q_r^{i,N}(t)$ denote the scaled (by $1/\sqrt{N}$) number of lines in the ith trunk that are used by rerouted requests at time t. We make the unrestrictive assumption that

$$\sup_N E|Q_r^{i,N}(0)| < \infty.$$

Set $\overline{\alpha}^{i,N} = E\alpha_k^{i,N}$, and assume (see (2.1)) that

$$(\overline{\alpha}^{i,N})^{-1} = \mu\beta_i N - b_i\sqrt{N}, \qquad i = 1, 2, 3, \tag{7.1}$$

for real b_i. This is the natural heavy traffic condition, since it implies that (modulo $O(\sqrt{N})$) the mean arrival rate to trunk i equals the mean service rate when all lines are occupied. Indeed, the heavy traffic analysis suggests that a "well-engineered" system will satisfy (7.1). Also, suppose (an unrestrictive assumption) that the sets

$$\{(\alpha_k^{i,N}/\overline{\alpha}^{i,N})^2; k < \infty, N < \infty\} \tag{7.2}$$

are uniformly integrable and that (see (2.3)) for some $\sigma_i^2 < \infty$

$$E[1 - \alpha_k^{i,N}/\overline{\alpha}^{i,N}]^2 = \text{var } \alpha_k^{i,N}/\overline{\alpha}^{i,N} \equiv (\sigma_i^N)^2 \to \sigma_i^2. \tag{7.3}$$

For $i = 1, 2, 3$, define the system variables:

$A^{i,N}(t) =$ [number of external arrivals to trunk i by t] $/\sqrt{N}$,
$D^{i,N}(t) =$ [number of service completions at trunk i by t] $/\sqrt{N}$,
$Q^{i,N}(t) =$ [number of lines of trunk i occupied at t] $/\sqrt{N}$,
$X^{i,N}(t) = \beta_i\sqrt{N} - Q^{i,N}(t)$,
$Y^{i,N}(t) =$ [number of external arrivals to i by t when i is full]/ $\sqrt{N}$.

$X^{i,N}(t)$ is the (scaled by $1/\sqrt{N}$) number of available circuits at trunk i at time t. We write $X^N(\cdot) = (X^{1,N}(\cdot), X^{2,N}(\cdot), X^{3,N}(\cdot))$ and similarly define the vectors for the other quantities defined above. Let (x^1, x^2, x^3) denote the canonical components of the vector x. Thus $\tilde{A}^N(\cdot)$ and $\tilde{D}^N(\cdot)$ are redefined to be the vectors of arrivals and departures, respectively.

Rerouting. Suppose that a request for service between A and B arrives but all $\beta_1 N$ lines of the connecting trunk 1 are occupied. We allow the possibility that the call can be routed between A and B by going through C, thus using one line on each trunk 2 and 3, if available. If not rerouted, the call is rejected from the system and disappears. The control problem is to determine when to reroute and when to reject. The same considerations apply to requests for service between A and C and between C and B.

An important restriction. We place one additional restriction on the system: There is $\Delta_0 > 0$ such that a requested rerouting from any trunk i to the *alternative pair* j, k will not be accepted if either $X^{j,N}(t) \leq \Delta_0$ or $X^{k,N}(t) \leq \Delta_0$. This restriction is not serious since Δ_0 can be made as small as desired, but it should be kept in mind since it will be exploited frequently and heavily in the sequel. It serves to prevent explosions of $Y^N(\cdot), F^N(\cdot)$ at the boundary that might be caused by the repeated rerouting of the same call. Note that this restriction is not a trunk reservation policy, since otherwise the controls are arbitrary.

An alternate assumption (which is not used here) would be to assign a small cost to each rerouting request. This would have a small effect on the limit but seems to require a more complex machinery.

Notation and the dynamical model. We use the convention that the superscript i indexing the trunk line is interpreted modulo 3. Thus, if $i = 2$, then $X^{(i+1),N} = X^{3,N}$ and $X^{(i+2),N} = X^{1,N}$. Let $J^{i,N}(t)$ denote the indicator function of the event that (an external) service request arrives at trunk i at time t, that the trunk is full, but that rerouting is accepted by the alternative pair. In order to avoid a minor notational complication, we suppose that only one event can occur at a time in the network (w.p.1). Otherwise, we need to define priorities in the timing. This is not hard to do, but it would complicate the notation. The control $J^N(\cdot) = (J^{1,N}(\cdot), \ldots)$ is said to be *admissible* if: (a) $J^{i,N}(t) = 0$ if there is no external arrival to trunk i at time t, $J^{i,N}(t)$ takes only the values 0 and 1 ; (b) $J^{i,N}(t)$ depends only on $\{A^N(s), D^N(s), s \leq t, J^N(s), s < t\}$; (c) $J^{i,N}(t) = 0$ if either $X^{(i+1),N}(t)$ or $X^{(i+2),N}(t)$ is less than Δ_0. Since the events occur singly and discretely in time, the $J^N(\cdot)$ are well defined.

Define the normalized overflow from trunk i that is actually rerouted:

$$F^{i,N}(t) = \int_0^t J^{i,N}(s)dY^{i,N}(s). \tag{7.4}$$

We have

$$Q^{1,N}(t) = Q^{1,N}(0) + A^{1,N}(t) - D^{1,N}(t) - Y^{1,N}(t) + F^{2,N}(t) + F^{3,N}(t),$$

since rerouted items from trunks 2 and 3 both use trunk 1. In general,

$$\begin{aligned} Q^{i,N}(t) = Q^{i,N}(0) &+ A^{i,N}(t) - D^{i,N}(t) - Y^{i,N}(t) \\ &+ F^{(i+1),N}(t) + F^{(i+2),N}(t). \end{aligned} \tag{7.5}$$

Let $\mathcal{B}_t^N$ denote the minimal σ-algebra that measures $\{X^{i,N}(s), A^{i,N}(s), D^{i,N}(s), J^{i,N}(s), s \leq t, i = 1, 2, 3\}$. Analogously to the definitions in §§3 and 4, define the processes

$$\tilde{A}^{i,N}(t) = \frac{1}{\sqrt{N}} \sum_{k=1}^{Nt} (1 - \alpha_k^{i,N}/\overline{\alpha}^{i,N}),$$

$$S^{i,N}(t) = \frac{1}{N} \times \max\left\{n : \sum_{k=1}^{n} \alpha_k^{i,N} \leq t\right\},$$

$$\tilde{A}_0^{i,N}(t) = -\tilde{A}^{i,N}(S^{i,N}(t)).$$

Write the martingale decomposition of $D^{i,N}(\cdot)$ analogously to the form (3.5):

$$\begin{aligned} D^{i,N}(t) &= \mu \int_0^t Q^{i,N}(s)ds + \tilde{D}^{i,N}(t) \\ &= -\mu \int_0^t X^{i,N}(s)ds + \sqrt{N}\beta_i \mu t + \tilde{D}^{i,N}(t), \end{aligned} \tag{7.6}$$

where $\tilde{D}^{i,N}(t)$ is a $\mathcal{B}_t^N$-martingale with quadratic variation

$$\begin{aligned} \langle \tilde{D}^{i,N} \rangle (t) &= \frac{\mu}{\sqrt{N}} \int_0^t Q^{i,N}(s)ds \\ &= \mu\beta_i t - \frac{\mu}{\sqrt{N}} \int_0^t X^{i,N}(s)ds. \end{aligned} \tag{7.7}$$

Now, analogously to (3.7), we have the system equations

$$\begin{aligned} X^{i,N}(t) = X^{i,N}(0) &+ b_i t - \mu \textstyle\int_0^t X^{i,N}(s)ds + (\tilde{A}_0^{i,N}(t) + \tilde{D}^{i,N}(t)) \\ &+ Y^{i,N}(t) - F^{(i+1),N}(t) - F^{(i+2),N}(t). \end{aligned} \tag{7.8}$$

Define the process

$$R^N(\cdot) = \{X^{i,N}(\cdot), \tilde{A}_0^{i,N}(\cdot), \tilde{D}^{i,N}(\cdot), Y^{i,N}(\cdot), F^{i,N}(\cdot); i = 1,2,3\}.$$

DEFINITION. *A measurable process $J(\cdot) = (J^1(\cdot), J^2(\cdot), J^3(\cdot))$ is said to be an* admissible control *for* (7.10) *below if the $J^i(t)$ are $[0,1]$-valued, nonanticipative with respect to the Wiener processes $W^j(\cdot)$ and $J^i(t) = 0$ if either $X^{(i+1)}(t)$ or $X^{(i+2)}(t)$ are less than Δ_0. The control is said to be* state feedback *if there is* measurable *$u(\cdot) = (u^1(\cdot), u^2(\cdot), u^3(\cdot))$ where $u^i(\cdot)$ is a $[0,1]$-valued function of $x^{(i+1)}, x^{(i+2)}$, and $J^i(t) = u^i(X^{(i+1)}(t), X^{(i+2)}(t))$. For simplicity, we often write this as $u^i(X(t))$. Since $u^i(\cdot)$ will occur only in the combination $u^i(X(t))dY^i(t)$, the x^i-dependence of $u^i(\cdot)$ is irrelevent.*

THEOREM 7.1. *Assume the conditions stated above in this section and let the $J^N(\cdot)$ be admissible and $\sup_N E|X^N(0)|^2 < \infty$. Then $\{R^N(\cdot)\}$ is tight and*

$$\sup_N \sup_t EX^{i,N}(t)^2 < \infty \quad \textit{for all } i. \tag{7.9}$$

Let $R(\cdot) = (X^i(\cdot), W_1^i(\cdot), W_2^i(\cdot), Y^i(\cdot), F^i(\cdot), i = 1,2,3)$ denote the limit of a weakly convergent subsequence of $\{R^N(\cdot)\}$. Write $W^i(t) = W_1^i(t) + W_2^i(t)$, $W(\cdot) = (W^1(\cdot), W^2(\cdot), W^3(\cdot))$. Then

$$X^i(t) = X^i(0) + b_i t - \mu \int_0^t X^i(s)ds + W^i(t) + Y^i(t) - F^{(i+1)}(t) - F^{(i+2)}(t). \tag{7.10}$$

The $W_j^i(\cdot)$ are mutually independent Wiener processes with quadratic variations

$$\begin{aligned} \langle W_1^i \rangle(t) &= \sigma_i^2 \beta_i \mu t, \\ \langle W_2^i \rangle(t) &= \beta_i \mu t. \end{aligned} \tag{7.11}$$

There is an admissible control $J(\cdot)$such that

$$F^i(t) = \int_0^t J^i(s)dY^i(s), \tag{7.12}$$

and the $X^i(\cdot)$ and $Y^i(\cdot)$ are nonanticipative with respect to the Wiener processes. The set

$$\{Y^{i,N}(n+1) - Y^{i,N}(n); n < \infty, N < \infty, i = 1,2,3\} \tag{7.13}$$

is uniformly integrable.

Proof. Equation (7.9) holds by Theorems 3.1 and 4.1 and the fact that the $X^{i,N}(t)$ are no greater than the values of the state when there is no rerouting.

The uniform integrability of (7.13) is proved by an argument like that below (4.7), using the fact that $J^{i,N}(t) = 0$ if either $X^{(i+1),N}(t)$ or $X^{(i+2),N}(t)$ are less than Δ_0. Now we elaborate this argument by working with a "bounding system." Let $\hat{X}^N(\cdot)$ and $\hat{Y}^N(\cdot)$, respectively, denote the state and reflection processes for the system (7.8) that is altered as follows. The state is reset to zero at $t = 1,2,3,\ldots$. Also, for each

$i, \hat{X}^{i,N}(t)$ is constrained to $[0, \Delta_0]$ by adding a reflection at Δ_0. (Thus, the F-terms do not appear in the dynamical equation.)

Note that $Y^{i,N}(n+1) - Y^{i,N}(n) \leq \hat{Y}^{i,N}(n+1) - \hat{Y}^{i,N}(n)$. We outline the proof of the uniform integrability for the altered system. Fix i and n; and define the stopping times $\sigma_0^N = n$ and for each $k \geq 0$, $\tau_{k+1}^N = \min\{t > \sigma_k^N : \hat{X}^{i,N}(t) = 0\} \wedge (n+1)$, $\sigma_{k+1}^N = \min\{t > \tau_{k+1}^N : \hat{X}^{i,N}(t) \geq \Delta_0\} \wedge (n+1)$. Define $M_n^{i,N} = \max\{k : \sigma_k^N < (n+1)\}$. Since $\hat{Y}^{i,N}(\cdot)$ cannot increase on the excursions from Δ_0 to next contact with 0, there is $K < \infty$ such that we can write

$$\begin{aligned}\hat{Y}^{i,N}(n+1) - \hat{Y}^{i,N}(n) \leq & -\sum_{k=1}^{\infty}[\tilde{A}_0^{i,N}(\sigma_k^N) - \tilde{A}_0^{i,N}(\tau_k^N)] \\ & + \text{a similar sum for } \tilde{D}^{i,N}(\cdot) + KM_n^{i,N} + K.\end{aligned}$$

The mean square value of the sums on the right are all bounded uniformly in n, i. Now, given $\delta_0 > 0$ it is not hard to show that there is $T_0 > 0$ such that with probability one

$$P\{\tau_{k+1}^N - \sigma_k^N < T_0 | \text{ data to } \sigma_k^N\} \leq 1 - \delta_0 \tag{7.14}$$

for all k, i, n and large N. By a recursion argument, this inequality can be used to show that each moment of $M_n^{i,N}$ is bounded uniformly in n, i, N and we omit the details. Thus the second moments of the set of increments in (7.13) are bounded uniformly.

The rest of the proof is like that of Theorems 3.1 and 4.1. Let τ be a (bounded) stopping time with respect to the filtration generated by the $\{\tilde{A}^{i,N}(t)\}$. Then from (7.3) we get

$$E[(\tilde{A}^{i,N}(\tau + t) - \tilde{A}^{i,N}(\tau))^2 | \tilde{A}^{i,N}(s), s \leq \tau] \leq (\sigma_i^N)^2 t$$

(and the left side converges to $\sigma_i^2 t$), which implies the tightness of $\{\tilde{A}^{i,N}(\cdot)\}$ via the criterion (3.1), analogous to the situation in Theorem 3.1. Continuity of the limits of $\{\tilde{A}^{i,N}(\cdot)\}$ follows from the uniform integrability of (7.2), again analogously to the situation in Theorem 3.1. The results for the $\tilde{D}^{i,N}(\cdot)$ sequence are also obtained essentially like that in Theorem 3.1. It can also be shown that $S^{i,N}(\cdot)$ converges weakly to the (deterministic) process with values $\mu\beta_i t$; i.e., the limit is just the limit of $(1/N)$ times the mean number of external arrivals to trunk i by time t. We have $dF^{(i+1),N}(t) = dF^{(i+2),N}(t) = 0$ for all t such that $X^{i,N}(t) \in [0, \Delta_0]$. This remark and a proof like that in Theorem 3.1 can be used to get the tightness of $Y^{i,N}(\cdot)$. The tightness of $F^{i,N}(\cdot)$ follows from this. These facts and the tightness of $\{\tilde{A}_0^N(\cdot)\}$ and $\{\tilde{D}^N(\cdot)\}$ imply the tightness of $\{X^{i,N}(\cdot)\}$. The representation (7.12) follows from the fact that $F^i(t+s) - F^i(t) \leq Y^i(t+s) - Y^i(t)$, for all $t, s \geq 0$.

Using calculations analogous to (3.9) and those leading to (3.16), we get that the $W_j^i(\cdot)$ are Wiener processes with the asserted variances as well as the nonanticipativeness assertion and the independence of the four groups of processes

$$W_1^1(\cdot), W_1^2(\cdot), W_1^3(\cdot), \{W_2^1(\cdot), W_2^2(\cdot), W_2^3(\cdot)\}.$$

The processes $\tilde{D}^{i,N}(\cdot)$ are not mutually independent since a rerouting involves the use of a line on two trunks for the same request. But the parts of, say, $\tilde{D}^{1,N}(\cdot), \tilde{D}^{2,N}(\cdot)$, which are due to the input $F^{3,N}(\cdot)$, will be shown in the next paragraph to have variance $O(1/\sqrt{N})$, and similarly for the other cases. This implies that the $\{W_2^i(\cdot), i = 1,2,3\}$ are also mutually independent.

To prove the desired assertion, first recall the definition of $Q_r^{i,N}(\cdot)$ above (7.1). We can decompose this process similarly to what was done in (3.5), (3.6). This yields

$$Q_r^{i,N}(t) = Q_r^{i,N}(0) + F^{(i+1),N}(t) + F^{(i+2),N}(t) - \mu \int_0^t Q_r^{i,N}(s)ds + \tilde{D}_r^{i,N}(t), \tag{7.15}$$

where $\tilde{D}_r^{i,N}(\cdot)$ is a martingale and is the contribution to $\tilde{D}^{i,N}(t)$ of the rerouting from other trunks. Its quadratic variation is

$$\frac{\mu}{\sqrt{N}} \int_0^t Q_r^{i,N}(s)ds. \tag{7.16}$$

By (7.15) and the uniform integrability properties of the increments of the $F^{j,N}(\cdot)$ (a consequence of (7.13)), we have that $\sup_{N,t} EQ_r^{i,N}(t) < \infty$. Thus the quadratic variation of $\tilde{D}_r^{i,N}(\cdot)$ has the asserted order and we are done. □

Note. The fact that the asymptotic effects of $Q_r^N(\cdot)$ are zero was a long-standing conjecture. But this seems to be the first proof.

8. Optimality results for the discounted cost problem. In this section, it is shown that the optimal costs for the network converge to that for the limit control problem. This is useful since numerical and analytical methods are available for the limit problem.

The cost function. We define the analogs of the costs (4.3) and (4.6) for the network. Let $X^N(0) = X(0) = x$ and $\beta > 0$: For admissible $J^N(\cdot)$ and $J(\cdot)$, set

$$C^N(J^N, x) = E\int_0^\infty e^{-\beta t} \sum_i (1 - J^{i,N}(t))dY^{i,N}(t), \tag{8.1}$$

$$C(J,x) = E\int_0^\infty e^{-\beta t} \sum_i (1 - J^i(t))dY^i(t),$$

$$\overline{C}^N(x) = \inf_{J^N} C^N(J^N, x), \quad \overline{C}(x) = \inf_J C(J,x), \tag{8.2}$$

where the infs are over the admissible controls. (8.1) is the normalized mean discounted total number of external arrivals that are rejected from the system.

We wish to show the analog of (6.1) for the network problem. As in §6, the main problem is proving (6.2). To do this, we will need to introduce a "comparison" control, as discussed in §6, and Theorem 8.1 is a preliminary result in that direction. The "comparison" controls that will be used in the convergence Theorem 8.3 are based on those constructed in Theorem 8.1.

DEFINITION. *Let $g^i(\cdot)$ be Lipschitz continuous $[0,1]$-valued functions of $(x^{(i+1)}, x^{(i+2)})$ that equal zero if either $x^{(i+1)} \le \Delta_0$ or $x^{(i+2)} \le \Delta_0$. For simplicity, we often write just $g^i(x)$. In fact, since this function only occurs in the form $g^i(X(t))dY^i(t)$, the x^i-dependence of $g^i(\cdot)$ is irrelevent. Let $\alpha^i \in [0,1]$ and define the control $J_\alpha^i(\cdot)$ by $J_\alpha^i(t) = \alpha^i g^i(X(t))$ and the transition probability under $J_\alpha(\cdot)$ by $P(x,t,\cdot|\alpha) = P\{X(t) \in \cdot | J_\alpha$ used, $X(0) = x\}$.*

THEOREM 8.1. *Assume the conditions of Theorem* 7.1 *and let $J(\cdot)$ be an admissible control for* (7.10). *Then a solution to* (7.10) *exists* (*in the strong sense*) *and* *is weak sense unique. For $t > 0, P(x,t,\cdot|\alpha)$ is mutually absolutely continuous with respect to Lebesgue measure and is continuous in x, α in the weak sense.*

Proof. The key to analyzing (7.10) is the fact that for each i at most one of the processes $Y^i(\cdot), F^{(i+1)}(\cdot)$, and $F^{(i+2)}(\cdot)$ can increase at any given time, analogous to the situation in Theorem 7.1. This observation will be used several times in the sequel, since it enables us to divide up time into disjoint intervals on each of which only one of these processes can change and to examine $X^i(\cdot)$ in "pieces."

The uniqueness can be established by a direct construction. For specificity, let $i = 1$ and $X^1(0) < \Delta_0$. Define $\sigma_0 = 0$, and for $n \geq 0$ set $\tau_{n+1} = \min\{t \geq \sigma_n : X^1(t) = 0\}$ and $\sigma_{n+1} = \min\{t \geq \tau_{n+1} : X^1(t) = \Delta_0\}$. The solution to (7.10) can be constructed by "pieces." On $[\sigma_n, \tau_{n+1}), X^1(\cdot)$ is a function of $X^1(\sigma_n)$ and the increments of $W^1(\cdot), F^2(\cdot), F^3(\cdot)$ on that interval and $dY^1(t) = 0$ there. On $[\tau_n, \sigma_n), J^2_\alpha(t) = J^3_\alpha(t) = 0$ since $X^1(t) < \Delta_0$. Thus, $dF^2(t) = dF^3(t) = 0$ on that interval, and $X^1(\cdot)$ is uncontrolled and uncoupled to $X^2(\cdot), X^3(\cdot)$ there. A continuation of this argument yields the existence and uniqueness, as asserted. The existence and uniqueness under $J_\alpha(\cdot)$ is proved in the same way.

Let $t > 0$. The fact that $P(x,t,\cdot|\alpha)$ is mutually absolutely continuous with respect to Lebesgue measure for each x and α follows from the arguments in [5, §7] for a related problem and the above observations concerning the examination of the process in "pieces."

The asserted uniform weak continuity of $P(x,t,\cdot|\alpha)$ can be proved by a weak convergence argument: Let $x_n \to x$ and $\alpha_n \to \alpha$, and use a weak convergence argument and the weak sense uniqueness of the solution to (7.10) under the given form of the control; we omit the details. □

THEOREM 8.2. *Assume the conditions of Theorem* 7.1, *and let $\{R^N(\cdot)\}$ converge weakly to $R(\cdot)$ (defined in Theorem* 7.1). *Then there is admissible $J(\cdot)$ such that*

$$C^N(J^N, x) \to C(J, x). \tag{8.3}$$

There exists an optimal admissible control for (7.10), (8.2). *Also, $\overline{C}(x)$ is continuous in x and in Δ_0 (for $\Delta_0 > 0$).*

Proof. Equation (8.3) is a consequence of Theorem 7.1. The existence of an optimal admissible control for (7.10), (8.2) can be shown by a weak convergence argument analogous to that used in Theorem 7.1 (but with system (7.10) and a minimizing sequence of controls used), and the details are omitted.

Perhaps the simplest way to prove the continuity of $\overline{C}(x)$ in x involves a change in the structure of the problem. For the physical network, the change is to allow the possibility of rejecting external inputs to a link i when $X^{i,N}(t) > 0$ and without attempting rerouting for such rejections. For $c_0 > 1$, assign a cost $c_0/\sqrt{N}$ per call rejected from i when $X^{i,N}(t) > 0$, for each i. The corresponding modification to model the limit (7.10) and cost (8.2) allows instantaneous *increases* in the components of the state, with the following cost: Let the impulsive increment move the state from x to y with $y^i \geq x^i$ for each i. Then we let the associated "impulsive" cost be $c_0\sum_i(y^i - x^i)$. Such a "reject" option would *never* be exercised in an optimal policy, so the modification does not alter the minimum costs. Now given x, let $x_n \to x$. A weak convergence argument can be used to show that there exists admissible $\tilde{J}(\cdot)$ such that $\overline{C}(x_n) \to C(\tilde{J}, x)$. Hence, $\liminf_n \overline{C}(x_n) \geq \overline{C}(x)$. Also, $\overline{C}(y) \leq \overline{C}(x)$ if $y^i \geq x^i$ for all i. These facts hold for both the original and the modified problems. The last three sentences imply that it is enough to show continuity in each component x^i separately and as x^i increases. But such continuity is guaranteed for the modified problem, since for that problem and $\delta^i > 0$

$$\overline{C}(x^1, x^2, x^3) \le \overline{C}(x^1 + \delta^1, x^2 + \delta^2, x^3 + \delta^3) + c_0(\delta^1 + \delta^2 + \delta^3),$$

$$\overline{C}(x^1, x^2, x^3) \ge \overline{C}(x^1 + \delta^1, x^2 + \delta^2, x^3 + \delta^3).$$

The continuity in Δ_0 can be shown by a related argument, and the details are omitted. □

The optimality theorem. Theorem 8.3 shows that the optimum cost for the network is well approximated by the optimum for the limit problem. This helps justify the use of the heavy traffic limit. We also note that the randomization technique used to get the comparison control in the theorem seems to be a basic technique in getting limit theorems for the trunk line type of problem.

For controls $J^i_\alpha(t) = \alpha^i g^i(X(t)), i = 1, 2, 3$, and $X(0) = x$, define the function

$$k^\delta(x, \alpha) = E\left[\int_0^\delta e^{-\beta t} \sum_i (1 - \alpha^i g^i(X(s))) dY^i(s) | X(0) = x\right].$$

THEOREM 8.3. *Under the conditions of Theorem* 7.1,

$$\overline{C}^N(x) \to \overline{C}(x). \tag{8.4}$$

Given $\varepsilon > 0$, suppose that there is a state feedback control $u(\cdot)$ for (7.10), (8.2), *which is ε-optimal and for which* (7.10) *has a unique weak sense solution. Write $u(x) = (u^1(x^2, x^3), u^2(x^1, x^3), u^3(x^1, x^2))$, and suppose that the $u^i(x)$ take values* 0 *or* 1 *and the boundary of the set where $u^i(x^{(i+1)}, x^{(i+2)}) = 0$ is piecewise C_2 and there are no cusps at the corners of the "pieces". Then $u(\cdot)$ is 2ε-optimal for* (7.8), (8.1) *for large N.*

Remark on the control. By an essentially classical argument of the type used for Itô equations, and working in "pieces" as in Theorem 8.1, it can be shown that (7.10) has a unique strong sense solution (in sense of [8]) if the $u^i(\cdot)$ are Lipschitz continuous. The discontinuous control described in the theorem statement is typical in applications, as seen from numerical results for the limit problem.

Numerical experiments support the conditions put on the ε-optimal control below (8.4). Indeed, the switching curve for a typical optimal control for a three-dimensional heavy traffic limit system is graphed in Fig. 1, where the upper region is "accept rerouting" and the lower is "reject rerouting." The graphed case is for an ergodic cost criterion, but similar results hold for the discounted cost case. For the graphed case, link 1 is full and the other axes are the state values for links two and three, respectively. The plotted contours are those of the "relative value function" from the dynamic programming equation and are used in the adaptation of the results for the three-dimensional case to the general network. Details of the numerics and algorithms will be given in a subsequent paper.

Remark on the proof. The proof of the first assertion will be done in several steps. First, an ε-optimal feedback control for (7.10), (8.2) of a special form is obtained. Then a piecewise constant approximation is derived. Then we get a continuous ε-optimal control for a time sampled form of (7.10), (8.2). Finally, this latter control is adapted for use on $X^N(\cdot)$. These controls are for theoretical use only and are not intended to be practical.

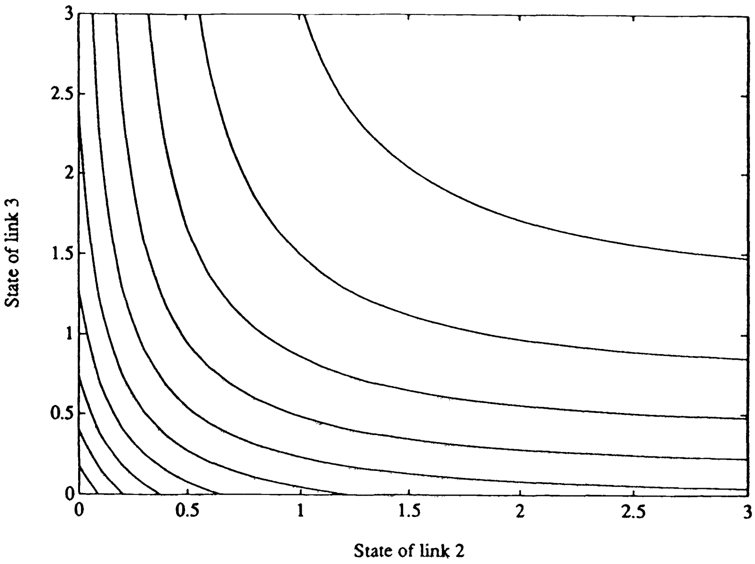


FIG. 1. *Decision curve for a particular three-dimensional system: Link* 1 *full.*

Proof. Equation (8.3) implies that

$$\liminf_N \overline{C}^N(x) \geq \overline{C}(x)$$

(for arbitrary $\varepsilon > 0$, simply let the $J^N(\cdot)$ in Theorem 8.2 be an optimal or ε-optimal policy). Thus, we need only prove

$$\limsup_N \overline{C}^N(x) \leq \overline{C}(x). \tag{8.5}$$

Let $\varepsilon > 0$. To prove (8.5), a special 3ε-optimal policy $J^\varepsilon(\cdot)$ for the limit problem will be constructed. $J^\varepsilon(\cdot)$ will be such that we can find an adaptation $J^{\varepsilon,N}(\cdot)$ for use in (7.8) such that

$$C^N(J^{\varepsilon,N}, x) \to C(J^\varepsilon, x). \tag{8.6}$$

Since $C(J^\varepsilon, x) \leq \overline{C}(x) + 3\varepsilon$ and $C^N(J^{\varepsilon,N}, x) \geq \overline{C}^N(x)$, (8.5) will follow.

Let $\delta^1 > 0$, and set $\Delta_1 = \Delta_0 + \delta^1$ and $X(0) = x$. In view of the continuity of $\overline{C}(x)$ in Δ_0 (Theorem 8.2), there is $\delta^1 > 0$ and an ε-optimal admissible policy $\tilde{J}(\cdot)$ such that $\tilde{J}^i(t) = 0$ if either $X^{(i+1)}(t) \leq \Delta_1$ or $X^{(i+2)}(t) \leq \Delta_1$. For each i, let $g^i(\cdot)$ be a function of the type introduced in Theorem 8.1 and taking the value unity when $x^{(i+1)} \geq \Delta_1$ and $x^{(i+2)} \geq \Delta_1$. Recall that $g^i(\cdot)$ is Lipschitz continuous, $[0,1]$-valued and has the value zero if $x^{(i+1)} \leq \Delta_0$ or $x^{(i+2)} \leq \Delta_0$. We suppose (without loss of generality) that the processes (7.10) for the various controls that will be used are defined on the same probability space and with the same Wiener processes $W^i(\cdot)$. The initial condition will always be $x = X(0)$. With $\tilde{J}(\cdot)$ given, let $\tilde{X}(\cdot), \tilde{F}(\cdot)$, and $\tilde{Y}(\cdot)$ denote the associated state, control, and reflection processes, with $\tilde{X}(0) = x$; i.e.,

$$\tilde{X}^i(t) = x^i + b_i t - \mu \int_0^t \tilde{X}^i(s)ds + W^i(t) + \tilde{Y}^i(t) - \tilde{F}^{(i+1)}(t) - \tilde{F}^{(i+2)}(t).$$

The control functions have the form

$$\tilde{F}^i(t) = \int_0^t \tilde{J}^i(s) g^i(\tilde{X}(s)) d\tilde{Y}^i(s).$$

Given $g(\cdot)$, let $C_g(x, J)$ denote the cost (8.2) for (7.10) when the control functions are retricted to the form

$$F^i(t) = \int_0^t J^i(s) g^i(X(s)) dY^i(s), \tag{8.7}$$

with $J(\cdot)$ admissible. Define $\overline{C}_g(x) = \inf_J C_g(J, x)$. Then $\overline{C}_g(x) \geq \overline{C}(x)$. Also, $C(\tilde{J}, x) = C_g(\tilde{J}, x) \leq \overline{C}(x) + \varepsilon$ by the definition of $\tilde{J}(\cdot)$.

In the next paragraph we define an approximation to $\tilde{J}(\cdot), \tilde{X}(\cdot)$. We will denote this new policy and associated control function, reflection, and state processes by $\tilde{F}_\delta(\cdot), \tilde{Y}_\delta(\cdot), \tilde{X}_\delta(\cdot)$, respectively. Thus the new approximating processes will satisfy

$$\begin{aligned} \tilde{X}^i_\delta(t) &= x^i + b_i t - \mu \int_0^t \tilde{X}^i_\delta(s) ds + W^i(t) + \tilde{Y}^i_\delta(t) - \tilde{F}^{(i+1)}_\delta(t) - \tilde{F}^{(i+2)}_\delta(t), \\ \tilde{F}^i_\delta(t) &= \int_0^t \tilde{J}^i_\delta(s) g^i(\tilde{X}_\delta(s)) d\tilde{Y}_\delta(s). \end{aligned} \tag{8.8}$$

If $\tilde{F}^i_\delta(t) \to \tilde{F}^i(t)$ as $\delta \to 0$, for each t and all i, then $\tilde{X}_\delta(t), \tilde{Y}_\delta(t)$ converge to $\tilde{X}(t), \tilde{Y}(t)$ for each t.

Define $\tilde{J}^i_\delta(\cdot)$ as follows. Given $\delta > 0$, define $\tilde{J}^i_\delta(\cdot)$ recursively on the intervals $[n\delta, n\delta + \delta)$. On $[0, \delta)$, set $\tilde{J}^i_\delta(s) = 0$. For $n > 0$ and on $[n\delta, n\delta + \delta)$, set $\tilde{J}^i_\delta(s) = 1$ until either (whichever comes first): (a) $t = n\delta + \delta$, or (b) the first time t that $\tilde{F}^i_\delta(t) = \tilde{F}^i(t)$. As $\delta \downarrow 0$, an argument by contradiction can be used to show that for each i and t

$$\tilde{F}^i_\delta(t) \uparrow \tilde{F}^i(t), \text{ as } \delta \downarrow 0. \tag{8.9}$$

A contradiction can be shown if one supposes that the convergence (8.9) does not hold for some i, since in that case $\lim_\delta \tilde{J}^i_\delta(t)$ would equal unity for $t \geq \inf\{s : \lim_\delta \tilde{F}^i_\delta(s) < \tilde{F}^i(s)\}$. By (8.9), $\tilde{X}^i_\delta(\cdot), \tilde{Y}^i_\delta(\cdot)$ converge to $\tilde{X}^i(\cdot), \tilde{Y}^i(\cdot)$. This convergence, (8.9), and the uniform integrability of (7.13) imply that $C_g(\tilde{J}_\delta, x) \to C_g(\tilde{J}, x)$ as $\delta \to 0$. Now, given this result, and letting δ be small, we see that the fraction of intervals on which some $\tilde{J}^i_\delta(\cdot)$ takes both values 1 and 0 goes to zero as $\delta \to 0$. Thus, we can suppose that $\tilde{J}_\delta(t)$ equals identically either 1 or 0 on the intervals $[n\delta, n\delta + \delta)$ and that $C_g(\tilde{J}_\delta, x) \leq C_g(\tilde{J}, x) + \varepsilon$ and that (8.9) holds. The value of δ will be fixed henceforth in this proof.

Let $C^\delta_g(J, x)$ denote the cost for input functions of the form in (8.7) but where the $J(\cdot)$ process is constant on each interval $[n\delta, n\delta + \delta)$. Let $\overline{C}^\delta_g(x)$ denote the infimum over this class. Note that $\overline{C}^\delta_g(x) \geq \overline{C}(x)$, $\overline{C}^\delta_g(x) \leq \overline{C}(x) + 2\varepsilon$. If there is a measurable function $u(\cdot)$ such that the control on $[n\delta, n\delta + \delta)$ can be written as $u(X(n\delta))$, then we abuse notation and call the control *state feedback* and write the costs as $C^\delta_g(u, x)$.

The optimization problem for cost $C^\delta_g(J, x)$ can be reduced to a *discrete time* problem by working with the samples $\{\tilde{X}_\delta(n\delta), n = 0, 1, \ldots\}$ and an appropriate discrete time form of the cost, and we now do this.

For the discrete time-approximating problem, we can suppose that the dynamic programming equation for the discrete time problem is

$$\overline{C}_g^\delta(x) = \min_\alpha \left[\int e^{-\beta\delta} \overline{C}_g^\delta(y) P(x,\delta,dy|\alpha) + k^\delta(x,\alpha) \right] \tag{8.10}$$

where $k^\delta(\cdot)$ is defined above the theorem. A weak convergence argument can be used to get the continuity of $k^\delta(\cdot)$. The continuity of $\overline{C}_g^\delta(\cdot)$ can be proved by the method used in Theorem 8.2. It can be shown that there is a continuous state feedback control $u^\varepsilon(\cdot)$ that is ε-optimal with respect to all controls for the problem whose dynamic programming equation is (8.10). Thus $C_g^\delta(u^\varepsilon, x) \leq \overline{C}(x) + 3\varepsilon$.

Now we are prepared to adapt $u^\varepsilon(\cdot) = (u^{1,\varepsilon}(\cdot), u^{2,\varepsilon}(\cdot), u^{3,\varepsilon}(\cdot))$ to the physical model (7.8). The adapted control law will be called $J^N(\cdot)$ and the associated state and reflection processes called $X^N(\cdot), Y^N(\cdot)$, respectively. The $J^N(\cdot)$ will be a *randomized* control. It will take values 0 or 1 and will be determined by the following conditional probability law : For $t \in (n\delta, n\delta + \delta]$, we use (and define $\tilde{u}(\cdot)$)

(8.11)

$$P\{J^{i,N}(t) = 1 | A^N(s), D^N(s), s \leq t, J^N(s), s < t; dA^{i,N}(t) > 0, X^{i,N}(t^-) = 0\}$$

$$= u^{i,\varepsilon}(X^N(n\delta)) g^i(X^N(t^-)) \equiv \tilde{u}^{i,N}(t).$$

Note that the conditioning event implies that there is an arrival to i at t and no available capacity there. Since the events occur one at a time and are separated in time, the conditional expectation is well defined.

Next we put the control terms $F^{i,N}(\cdot)$ into a manageable form by showing that the effects of the randomization disappear as $N \to \infty$. Let $\tau_n^{i,N}$ denote the time of the nth arrival to trunk i from the external sequence $A^{i,N}(\cdot)$. Let $I_n^{i,N}$ denote the indicator of the set where $X^{i,N}(\tau_n^{i,N-}) = 0$, and set $J_n^{i,N} = J^{i,N}(\tau_n^{i,N})$. Then we can write (possibly modulo $1/\sqrt{N}$)

$$Y^{i,N}(t) = \frac{1}{\sqrt{N}} \sum_{n=1}^{NS^{i,N}(t)} I_n^{i,N},$$

$$F^{i,N}(t) = \int_0^t J^{i,N}(s) dY^{i,N}(s) = \frac{1}{\sqrt{N}} \sum_{n=1}^{NS^{i,N}(t)} J_n^{i,N} I_n^{i,N}$$

$$= \frac{1}{\sqrt{N}} \sum_{n=1}^{NS^{i,N}(t)} \tilde{u}^{i,N}(\tau_n^{i,N}) I_n^{i,N} + \frac{1}{\sqrt{N}} \sum_{n=1}^{NS^{i,N}(t)} (J_n^{i,N} - \tilde{u}^{i,N}(\tau_n^{i,N})) I_n^{i,N}.$$

The first sum on the last line equals $\int_0^t \tilde{u}^{i,N}(s) dY^{i,N}(s)$. Owing to the definition of $J_n^{i,N}$ via a conditional probability, the second sum is a martingale (when a discrete index is used in place of $NS^{i,N}(t)$) and its variance is bounded by

$$\frac{\text{const}}{\sqrt{N}} E|Y^{i,N}(t)|.$$

Thus, the evolution equation for $X^n(\cdot)$ under (8.11) can be written as (7.8) but with $F^{i,N}(t)$ replaced by $\int_0^t \tilde{u}^{i,N}(s) dY^{i,N}(s)$ plus a "noise" term that goes to zero as $N \to \infty$.

Now, take a weakly convergent subsequence of the set $\{R^N(\cdot)\}$ and let $R(\cdot) = (X(\cdot), \ldots, F(\cdot))$ denote the limit processes. Then the limit satisfies (7.10) where on $[n\delta, n\delta + \delta)$

$$F^i(t) - F^i(n\delta) = \int_{n\delta}^{t} u^{i,\varepsilon}(X(n\delta))g^i(X(s))dY^i(s).$$

Owing to the continuity of $u^{i,\varepsilon}(\cdot)$ and Lipschitz continuity of $g^i(\cdot)$, the solution to (7.10) with these input functions is weak sense unique. This uniqueness implies that $R^N(\cdot) \Rightarrow R(\cdot)$, as $N \to \infty$. Also by the weak convergence and uniform integrability of (7.13),

$$C^N(J^N, x) \to C_g^\delta(u^\varepsilon, x) \leq \overline{C}(x) + 3\varepsilon.$$

Since $\overline{C}^N(x) \leq C^N(J^N, x)$ and ε is arbitrary, we have (8.5). The proof of the last assertion of the theorem uses a weak convergence argument and depends on the local properties of the Wiener process and [1, Theorem 5.1] to deal with the discontinuous control, and we omit the details. □

9. The ergodic cost function.

Now we treat the ergodic cost problem for the network and show that the optimal costs for the network converge to the optimal cost for the ergodic control problem for the limit process. The method of proof is somewhat indirect, since there does not seem to be any available technique for dealing with controlled reflection directions with the ergodic cost criterion. Owing to the many details required, the major details are in the Appendix. Again, numerical methods of the "Markov chain approximation type" are available.

The cost function. Define the cost (the mean number of rejections from the entire system per unit time for (7.8))

$$\gamma_T^N(J^N, x) = \frac{1}{T}E\int_0^T \sum_i (1 - J^{i,N}(s))dY^{i,N}(s), \qquad X^N(0) = x, \tag{9.1}$$

$$\bar{\gamma}_T^N(x) = \inf_{J^N} \gamma_T^N(J^N, x),$$

$$\gamma^N(J^N, x) = \limsup_T \gamma_T^N(J^N, x),$$

$$\overline{\gamma}^N(x) = \inf_{J^N} \gamma^N(J^N, x).$$

For (7.10) and $X(0) = x$, define the costs

$$\begin{gathered}\gamma_T(J, x) = \frac{1}{T}E\int_0^T \sum_i (1 - J^i(s))dY^i(s), \\ \gamma(J, x) = \limsup_T \gamma_T(J, x), \quad \overline{\gamma}(x) = \inf_J \gamma(J, x),\end{gathered} \tag{9.2}$$

where all the infs are over the admissible controls. The occupation measure arguments of Theorem 5.1 can be adapted to the current problem. Let $J^N(\cdot)$ be a sequence of admissible controls for (7.8). Analogously to what was done in §5, define

$$R_t^N(\cdot) = (X^{i,N}(t+\cdot), \tilde{A}_0^{i,N}(t+\cdot) - \tilde{A}_0^{i,N}(t), \tilde{D}^{i,N}(t+\cdot) - \tilde{D}^{i,N}(t),$$

$$F^{i,N}(t+\cdot) - F^{i,N}(t), Y^{i,N}(t+\cdot) - Y^{i,N}(t), i = 1,2,3)$$

and let $P^{N,t}(\cdot)$ denote the measure of $R_t^N(\cdot)$. Define the occupation measure $P_T^N(\cdot)$ from $P^{N,t}(\cdot)$ as (5.2a). We have the following theorem.

THEOREM 9.1. *Assume the conditions of Theorem* 7.1. *Then* $\{P_T^N(\cdot); N < \infty, T < \infty\}$ *is tight. Let* $\overline{P}(\cdot)$ *denote the limit of a weakly convergent subsequence. Then* $\overline{P}(\cdot)$ *is the measure induced by a process*

$$R(\cdot) = (X^i(\cdot), W_1^i(\cdot), W_2^i(\cdot), F^i(\cdot), Y^i(\cdot), i = 1,2,3),$$

which satisfies (7.10) *with the representation* (7.12), *with admissible* $J(\cdot)$. *The* $W_k^i(\cdot)$ *are as in Theorem* 7.1, *and we write* $W^i(\cdot) = W_1^i(\cdot) + W_2^i(\cdot)$. *The limit process is stationary. Let* N, T *index the weakly convergent subsequence. Then*

$$\gamma_T^N(J^N, x) \to \gamma(J), \text{ as } N \to \infty, T \to \infty, \tag{9.3}$$

where $\gamma(J)$ *is the cost for the limit stationary process.*

Remark on the proof. The proof follows very closely the lines of the occupation measure argument of Theorem 5.1 together with the details in Theorem 7.1 concerning the weak convergence, and the details are omitted.

Convergence of the optimal costs. Theorem 9.1 yields

$$\liminf_{N,T} \overline{\gamma}_T^N(x) \geq \overline{\gamma}(x). \tag{9.4}$$

Thus, in order to prove

$$\overline{\gamma}_T^N(x) \to \overline{\gamma}(x), \text{ as } N \to \infty, T \to \infty, \tag{9.5}$$

one needs the analog of (8.5) and, in particular, some "comparison" control, which serves the same purpose as that constructed in Theorem 8.3. Suppose that for each $\varepsilon > 0$ there is an ε-optimal state feedback control $u^\varepsilon(\cdot)$ that has an "adaptation" $J^{\varepsilon,N}(\cdot)$ such that $\gamma_T^N(J^{\varepsilon,N}, x) \to \gamma(u^\varepsilon, x)$. Then, as in Theorem 8.3, one has

$$\limsup_{N,T} \overline{\gamma}_T^N(x) \leq \overline{\gamma}(x) \tag{9.6}$$

and (9.5) follows.

Getting the comparison control $u^\varepsilon(\cdot)$ seems to be much harder for the ergodic cost problem, due to difficulties in dealing with approximations of stationary solutions to (7.10). Consequently, an indirect approach is used in the Appendix where the control inputs are approximated by nicer functions and classical methods are used to get (9.5). The reader who is willing to accept the existence of a smooth ε-optimal feedback comparison control can omit the Appendix. The thrust of the arguments in the appendix is essentially to replace the reflection by a "barrier," adjust the control accordingly, and then apply the methods of [8] for the unreflected problem.

10. Appendix. A modified limit system and the ergodic problem: Proof of (9.5), (9.6).

In this section, we prove (9.6) and hence obtain the result (9.5). To do this we first approximate the input functions $F^{i,N}(\cdot), F^i(\cdot)$ for (7.8) and (7.10), respectively, by functions that are easier to handle. Essentially, a natural "barrier" method will be used. Then we prove that the associated costs provide good approximations. Finally, we prove the desired result for the approximating system using results that are known for the ergodic problem for nondegenerate diffusions. Most of this section will be devoted to defining the approximation and discussing its most important properties. The approximations will be such that the associated optimal costs provide upper and lower bounds for $\overline{\gamma}(x)$. The bounds will be arbitrarily close. Then a "nice" approximating control for the bounding system can be used to complete the proof of (9.6).

For large $k_0 > 0$ and small $\delta_{k_0} > 0$, define the real-valued function $k_1(\cdot)$ on $[0, \infty) : k_1(t) = k_0$ for $t \in [0, \delta_{k_0}], k_1(t) = 0$ for $t > \delta_{k_0}$. Let $\delta_{k_0} < \Delta_0$. Recall the notation $W^i(t) = W_1^i(t) + W_2^i(t)$ and $W(\cdot) = (W^i(\cdot), i = 1, 2, 3)$. Let $\Sigma = \text{cov } W(1)$.

Recall that if $J(\cdot)$ is an admissible control, then we require $J^i(t) = 0$ if either $X^{(i+1)}(t)$ or $X^{(i+2)}(t)$ are less than Δ_0. The approximating system to (7.10) is

$$(10.1)\qquad \begin{aligned} X^i(t) = X^i(0) + b_i t - \mu \int_0^t X^i(s)ds + Y^i(t) + W^i(t) \\ + \int_0^t k_1(X^i(s))ds - \int_0^t k_1(X^{(i+1)}(s))J^{(i+1)}(s)ds \\ - \int_0^t k_1(X^{(i+2)}(s))J^{(i+2)}(s)ds. \end{aligned}$$

The motivation for the approximation appears in the paragraph below (10.3). Next we put (10.1) into a more convenient vector form. Define $k(x) = (k_1(x^1), k_1(x^2), k_1(x^3))$ and the matrix

$$K(x) = \begin{bmatrix} 0 & k_1(x^2) & k_1(x^3) \\ k_1(x^1) & 0 & k(x^3) \\ k_1(x^1) & k_1(x^2) & 0 \end{bmatrix}.$$

Then we can write (10.1) in the compact vector form

$$(10.2)\qquad \begin{aligned} X(t) = X(0) + bt - \mu \int_0^t X(s)ds + Y(t) + W(t) \\ + \int_0^t k(X(s))ds - \int_0^t K(X(s))J(s)ds. \end{aligned}$$

In (10.2), the control $J^i(t)$ might be nonzero even if $X^i(t) \neq 0$. Of course, if $X^i(t) > \delta_{k_0}$, then the value of $J^i(t)$ is irrelevent since $k_1(X^i(t)) = 0$ there. Define the costs for the approximating problem, for $X(0) = x$,

$$(10.3)\qquad \begin{aligned} \gamma_T(k_0, J, x) &= \frac{1}{T} E \left[\sum_i \int_0^T (1 - J^i(s)) k_1(X^i(s))ds \right], \\ \gamma(k_0, J, x) &= \limsup_T \gamma_T(k_0, J, x). \end{aligned}$$

If $J(\cdot)$ is of state feedback form (i.e., there is measurable $[0,1]^3$-valued $u(\cdot)$, such that $J(t) = u(X(t))$ for some $u(\cdot)$), we write the cost as $\gamma(k_0, u, x)$. Define $\overline{\gamma}(k_0, x) = \inf_J \gamma(k_0, J, x)$ where the inf is over the admissible controls. If they do not depend on the initial condition x, then we drop the x-argument from $\gamma(k_0, J, x)$ or $\overline{\gamma}(k_0, x)$.

Motivation. The basic motivation for the approximation (10.2) is that for appropriately large k_0 and small δ_{k_o}, the reflection term $Y(\cdot)$ there is small and the reflection term in (7.10) is effectively replaced by the integral $\int_0^t k(X(s))ds$. Since the reflection term can be made as small as desired by making k_0 large (Theorem 10.5), ignoring it in the cost function (10.3) is unimportant. Loosely speaking, the approximation is equivalent to not accepting external requests on the direct trunk and requesting rerouting starting "just before" a trunk fills up, with "intensity" determined by $k_1(\cdot)$, and not requesting rerouting for arrivals to a full trunk. With these approximations, the acceptances and decisions do not take place on the boundary but just before the boundary. This "smoothing" greatly facilitates the analysis.

Theorems 10.1–10.4 concern regularity, stationarity, and ergodic properties of (10.2). Theorem 10.5 is used to show that (10.2) well approximates (7.10) for large k_0 and small δ_{k_0}. Theorem 10.6 establishes that $\bar{\gamma}(k_0, x)$ is also the inf over all state feedback controls, and Theorem 10.7 yields the final convergence result. It is of general interest to know that some problems with controlled reflections can be approximated in this way.

The control problem with system (10.2) and cost (10.3). The control problem (10.2), (10.3) is relatively easy to study via existing "Girsanov measure transformation" methods, in essentially the same way as done for the "unreflected" problem, and we now show how to do this. Let $W_0(\cdot)$ be a Wiener process with covariance matrix Σt, define the solution to the *uncontrolled* problem by the equation

$$X(t) = X(0) + bt - \mu \int_0^t X(s)ds + W_0(t) + Y(t) + \int_0^t k_1(X(s))ds, \tag{10.4}$$

and let P_x^0 denote the measure of the $(X(\cdot), Y(\cdot))$ in (10.4) when $X(0) = x$. This process (10.4) is very simple since it is actually composed of three mutually independent one-dimensional processes. For all x and all $t > 0$, the transition function $P^0(x, t, \cdot) \equiv P_x^0\{X(t) \in \cdot\}$ is mutually absolutely continuous with respect to Lebesque measure. The process is a strong Markov and strong Feller process and the solution to (10.4) is weak sense unique. If $k_1(\cdot)$ were Lipschitz continuous, then the solution would be strong sense unique. Girsanov transformation methods can be used since we need only shift the drift and not the singular component $Y(\cdot)$ to get the control term in (10.2).

Following a well-known procedure for the unreflected problem, the controlled system (10.2) will be defined from the solution of (10.4) via a Girsanov measure transformation. For an admissible (with respect to $W_0(\cdot)$) control $J(\cdot)$, define

$$\psi_0^t(J) = -\int_0^t [K(X(s))J(s)]'\Sigma^{-1}dW_0(s)$$
$$-\frac{1}{2}\int_0^t [K(X(s))J(s)]'\Sigma^{-1}[K(X(s))J(s)]ds,$$

$$\zeta_0^t(J) = \exp \psi_0^t(J).$$

For admissible (with respect to $W_0(\cdot)$) $J(\cdot)$, define the measure P_x^J via its restrictions to functions on the intervals $[0,t]$ for each t by the Radon–Nikodym derivatives that take the values $dP_x^J/dP_x^0 = \zeta_0^t(J)$ at such points. Then P_x^J is induced by a process $(X(\cdot),Y(\cdot))$ solving (10.2) for appropriate $W(\cdot)$ [8, Chapter IV.4]. If $J(t)=u(X(t))$ for a state feedback $u(\cdot)$, we write $\zeta_0^t(u)$ and P_x^u in lieu of $\zeta_0^t(J)$ and P_x^J, respectively. The following result holds due to the properties of the Girsanov transformation, just as for the unreflected system, since all of the properties of the solution to (10.4) that were cited in the last paragraph (except for the strong sense uniqueness) carry over to the solution of (10.2).

THEOREM 10.1. *Let $u(\cdot)$ be a state feedback control. The process $X(\cdot)$ with transition function $P^u(x,t,\cdot) = P_x^u\{X(t) \in\}$ is a strong Markov and strong Feller process. For all x and all $t > 0, P^u(x,t,\cdot)$ is mutually absolutely continuous with respect to Lebesgue measure, and it is weakly continuous in x. Also,* (10.2) *has a weak sense unique solution.*

Due to Theorem 10.1, the analysis of the ergodic control problem is virtually identical to that for the unreflected problem in [12]. In fact, the only properties used in [12] are those asserted in Theorem 10.1 and certain stability properties, which will be stated in the next theorem. Theorems 10.3, 10.4, and 10.6 essentially rewrite the results of [12] in the terms of the problem of this paper. For future reference, note that the control term called $b^u(x)$ in [12] is our $K(x)u(x)$.

THEOREM 10.2. *Let $E|X(0)|^2<\infty$. Then there is $M_0<\infty$ such that for* (10.2) *and admissible $J(\cdot)$,*

$$\sup_t E^J|X(t)|^2 \le M_0 \tag{10.5}$$

and M_0 can be chosen independently of $J(\cdot), k_0, \delta_{k_0}$. For each state feedback $u(\cdot)$, there is a unique invariant measure $\pi_u(\cdot)$, which is mutually absolutely continuous with respect to Lebesgue measure. Also, for each Borel set A

$$P^u(x,t,A) \to \pi_u(A) \tag{10.6}$$

as $t\to\infty$.

In addition, for (10.2) *and $J(\cdot)$ admissible*

$$\sup_{\substack{J,i,n,k_0,\delta_{k_0}\\ u,x}} E_x^u\left[\int_n^{n+1} k_1(X^i(s))ds + [Y^i(n+1)-Y^i(n)]\right]^2 < \infty. \tag{10.7}$$

For each state feedback control $u(\cdot)$, $X(\cdot)$ is recurrent in the sense that for any measurable set K of nonzero Lebesgue measure

$$E\int_0^\infty I_K(X(s))ds = \infty.$$

Proof. Define

$$V(x) = \sum_i (x^i-\delta_{k_0})^2 I_{\{x^i\ge\delta_{k_0}\}}.$$

Let $J(\cdot)$ be admissible for (10.2). Since $-V_x'(X(s))K(X(s))J(s)\le 0$ for all s, Itô's lemma yields

$$V(X(t)) \leq V(X(0)) + \int_0^t V_x'(X(s))[b - \mu X(s)]ds + (\text{trace}\,\Sigma)t$$
$$+ \int_0^t V_x'(X(s))[dY(s) + dW(s)].$$

Since $V_x'(X(s))dY(s) \equiv 0$, (10.5) follows from the last inequality via the Bellman–Gronwall ineqality. (Note that the same proof can be used to show that all moments of $X(t)$ are bounded uniformly in t.) The assertions below (10.5) up to and including (10.6) follow from the stability and Theorem 10.1, exactly as in [12, Thm. 3.1], which requires only the stability and strong Markov and Feller properties. The proof of (10.7) is the same as that used to prove uniform integrability of (7.13) using the fact that $\delta_{k_0} < \Delta_0$. The recurrence property is a consequence of (10.5), the weak continuity of $P^u(x,t,\cdot)$ in x, and the fact that $P^u(x,t,\cdot)$ is absolutely continuous with respect to Lebesgue measure. □

THEOREM 10.3. *Let the sequence of state feedback controls $u_n(\cdot)$ converge to a state feedback control $u(\cdot)$ in the sense that*

$$\int_A u_n(x)dx \to \int_A u(x)dx \tag{10.8}$$

for each Borel A. Then

$$\exp \zeta_0^t(u_n) \to \exp \zeta_0^t(u) \tag{10.9}$$

in L_1 (measure P_x^0) for each t and x. Also, for each Borel set A,

$$P^{u_n}(x,t,A) \to P^u(x,t,A), \tag{10.10}$$

$$\pi_{u_n}(A) \to \pi_u(A). \tag{10.11}$$

Proof. The proof is that of [12, Thm. 4.3], which only requires the strong Feller and strong Markov and stability properties. Hence it only needs the assertions of Theorem 10.1. The only part of the proof of [8, Thm. 4.3] that is not in that paper is a reference to [2, Thm. IV-3] for the proof of (10.9). But the proof of (10.9) in [2] also requires only the strong Feller and strong Markov properties. □

THEOREM 10.4. *Let $E|X(0)|^2 < \infty$. Let $u_n(\cdot)$ and $u(\cdot)$ be state feedback controls. Then $\gamma(k_0, u, X(0))$ does not depend on $X(0)$. Under* (10.8),

$$\gamma(k_0, u_n) \to \gamma(k_0, u). \tag{10.12}$$

There exists an optimal admissible state feedback control; i.e., there is a state feedback control $\overline{u}(\cdot)$ such that $\overline{\gamma}(k_0) \equiv \inf_{u(\cdot)\ \text{adm.}} \gamma(k_0, u) = \gamma(k_0, \overline{u})$.

For any $\varepsilon > 0$, there is a Lipschitz continuous ε-optimal state feedback control $u_\varepsilon(\cdot)$; i.e.,

$$\gamma(k_0, u_\varepsilon) \leq \overline{\gamma}(k_0) + \varepsilon.$$

Proof. The proof is that of [12, Thm. 4.4], and we just review its structure. For a state feedback control $u(\cdot)$, the integrand in the first line of (10.3) can be written as $\int_0^t c^u(X(s))ds$, where

$$c^u(x) = \sum_i (1 - u^i(x))k_1(x^i).$$

In [8, Thm. 4.4], F_n is used for our $c^u(\cdot)$, where the n there refers to the use of a control $u_n(\cdot)$. Then, for each state feedback control $u(\cdot), \gamma(k_0, u) = \int c^u(x)\pi_u(dx)$. Let $u_n(\cdot), u(\cdot)$ be any state feedback controls satisfying (10.8). Then the proof of the assertion (10.12) is in the proof of Theorem 4.4 in [12]. That proof is self-contained and uses only assertions analogous to those in Theorems 10.3 and 10.2 (up to (10.6)), except for a reference to [2, Prop. IV-4] for a proof of a proposition analogous to the convergence

$$\int_0^t c^{u_n}(X(s))ds \to \int_0^t c^u(X(s))ds$$

in probability (P_x^0-measure) for each t, x. This proof of this latter fact in [2, Prop. IV-4] requires only the strong Feller and strong Markov properties, hence it needs only the assertions in Theorems 10.1 and 10.2. Now let $\{u_n(\cdot)\}$ be a minimizing sequence of admissible state feedback controls. There is $\overline{u}(\cdot)$ such that (choose a subsequence, if necessary) (10.8) holds for $\overline{u}(\cdot)$ replacing $u(\cdot)$. The last assertion of the theorem then follows from (10.12), since for any state feedback $u(\cdot)$ there are $u_n(\cdot)$, arbitrarily smooth, such that (10.8) holds with $\overline{u}(\cdot)$ replacing $u(\cdot)$. □

THEOREM 10.5. *Let $\delta_{k_0} \to 0$ as $k_0 \to \infty$ such that $k_0(\exp -q_i k_0 \delta_{k_0}) \to 0$, where $q_i = 2/E[W^i(1)]^2$. Then for $z \in (0, \Delta_0)$ and as $k_0 \to \infty$ ((10.13) defines $G_0(k_0, z)$)*

$$P^J\{X^i(t) \text{ reaches } \Delta_0 \text{ before } 0|X^i(0) = z\} \equiv G_0(k_0, z) \to 1, \tag{10.13}$$

$$\sup_{n\geq 1, i} E[Y^i(n+1) - Y^i(n)] \to 0, \tag{10.14}$$

uniformly in $J(\cdot)$.

Proof. Equation (10.13) follows from a direct calculation of the value. In fact, $J(\cdot)$ is not relevant since $X^i(t) \in [0, \Delta_0]$ in the calculation. For simplicity, suppose that we have transformed the measure so that the system on the spacial interval of concern $[0, \Delta_0]$ (with absorbtion on the boundary) can be written as $dz = k_1(z)dt + dW^1$, $\text{var}W^1(1) = 2/q_i$. Let $G(\cdot)$ denote the probability $G_0(\cdot)$ for the new system. The measures of the transformed and the original systems are mutually absolutely continuous, and the Radon–Nikodyn is bounded in mean square, uniformly in $J(\cdot), k_0, \delta_{k_0}$. Then, if (10.13) holds for the transformed system, it will hold for the original system. Then $G(k_0, z)$ satisfies the differential equation $q_i k_1(z) G_z(k_0, z) + G_{zz}(k_0, z) = 0, z \in (0, \Delta_0)$, with the boundary condition $G(k_0, 0) = 0, G(k_0, \Delta_0) = 1$. Define $U(k_0, z) = G_x(k_0, z)$. Then since $k_1(z) = k_0$ for $z \leq \delta_{k_0}$ and it equals zero otherwise:

$$\begin{aligned} U(k_0, z) &= U(k_0, 0) \exp - \int_0^z q_i k_1(y) dy \\ &= U(k_0, 0) \exp -q_i k_0 z \qquad (z \leq \delta_{k_0}) \\ &= U(k_0, 0) \exp -q_i k_0 \delta_{k_0} \qquad (z > \delta_{k_0}), \end{aligned}$$

and

$$G(k_0, \delta_{k_0}) = \frac{(1 - e^{-q_i k_0 \delta_{k_0}})}{(1 - e^{-q_i k_0 \delta_{k_0}}) + q_i k_0 (\Delta_0 - \delta_{k_0}) e^{-q_i k_0 \delta_{k_0}}}, \tag{10.15}$$

from which (10.13) follows for $z = \delta_{k_0}$. A similar calculation yields (10.13), for all $z \in (0, \Delta_0)$.

Let $T_1 > 0$. Then (10.13) can be used to show that the probability that $X^i(t) = 0$ somewhere on the interval $[T, T + T_1]$ goes to zero uniformly in $T, J(\cdot)$ (provided T is bounded away from zero if $X^i(0) = 0$ with a positive probability). This, together with the uniform integrability of $\{Y^i(n+1) - Y^i(n), n < \infty\}$ which is implied by (10.7) yields (10.14). □

A maximum principle. For our purposes, it is necessary to know that $\overline{\gamma}(k_0, x)$ is also the infimum over all state feedback controls (in which case, it would not depend on x). Such a result was not explicitly shown in [12] that worked only with state feedback controls, but it is readily obtainable from the results of that paper; we give the required adaptation here. Let $\bar{u}(\cdot)$ be an optimal state feedback control. Let $\psi^{\overline{u}}(\cdot)$ be a measurable $I\!R^3$-valued function of x, which will be defined further below. We adjust the terminology to conform with that in [8] when possible. For a state feedback control $v(\cdot)$, define ($c^u(\cdot)$ was defined in Theorem 10.4)

$$e^{\overline{u},v}(x) = [c^{\overline{u}}(x) - c^v(x)] + \psi^{\overline{u}}(x) K(x) [\overline{u}(x) - v(x)]. \tag{10.16}$$

This is the same as the $e^{u,v}(x)$ in [8, (6.3)] with $u = \overline{u}$.

In [12, Thm. 6.1] it is shown that there exists a Borel $\psi^{\overline{u}}(\cdot)$ such that the condition that $e^{\overline{u},v}(x) \le 0$ for almost all x for each admissible state feedback $v(\cdot)$ is both necessary and sufficient for $\overline{u}(\cdot)$ to be optimal for the system (10.2) and cost (10.3) in the class of state feedback controls. The proofs in [8] are for the unreflected diffusion. But, as noted earlier, the development in that paper also holds for the approximating models (10.2). The proof of the cited theorem can be adapted to show that $\bar{u}(\cdot)$ is optimal with respect to all admissible controls, and now we outline the required alterations. Henceforth $\psi^{\overline{u}}(\cdot)$ will denote the cited function. The reader who is willing to accept that the infima of the costs over feedback and general admissible controls are equal can skip the next theorem.

THEOREM 10.6. *$\overline{\gamma}(k_0)$ is the infimum of $\overline{\gamma}(k_0, J, x_0)$ either over all admissible controls or over all state feedback controls, for each initial condition $X(0) = x_0$.*

Proof. Only an outline of the adaptation of the proofs in [12] will be given. Let $J(\cdot)$ be admissible with $X(0) = x_0$, and let $t_n \to \infty$ be a sequence of real numbers such that $\gamma_{t_n}(k_0, J, x_0) \to \gamma(k_0, J, x_0)$. Let $\overline{u}(\cdot)$ be optimal for (10.2),(10.3) in the class of state feedback controls. Then, as noted above the theorem, $e^{\overline{u},v}(x) \le 0$ for almost all x for each admissible state feedback $v(\cdot)$.

Define the "centered" cost rate

$$\tilde{c}^{\overline{u}}(x) = c^{\overline{u}}(x) - \gamma(k_0, \overline{u}).$$

Following a procedure very close to that which led to the fourth equation from the bottom of [12, p. 343] (where our $\tilde{c}^u$ is their $\tilde{k}^u$), for admissible $J(\cdot)$ we have

$$\lim_{n\to\infty} \frac{1}{n} E^J \int_0^{t_n} \left\{ \tilde{c}^{\overline{u}}(X(s)) + \psi^{\overline{u}}(X(s)) K(X(s)) [\overline{u}(X(s)) - J(s)] \right\} ds = 0. \tag{10.17}$$

Analogous to the definition of $\tilde{c}^{\overline{u}}(x)$, define the "centered" cost rate

$$\tilde{c}^J(s) = \sum_i (1 - J^i(s)) k_1(X^i(s)) - \gamma(k_0, J, x_0)$$

and note that $E^J \int_0^{t_n} \tilde{c}^J(s)ds/t_n \to 0$ by the centering of $\tilde{c}^J(\cdot)$ and the definitions of $\gamma(k_0, J, x_0)$ and t_n. Then by (10.17), we can write

(10.18)
$$\lim_{n\to\infty} \frac{1}{t_n} E^J \int_0^{t_n} \left\{ \tilde{c}^{\overline{u}}(X(s)) - \tilde{c}^J(s) + \psi^{\overline{u}}(X(s))K(X(s))[\overline{u}(X(s)) - J(s)] \right\} ds = 0].$$

There is a measurable function $w(\cdot)$ mapping $[0,\infty)^3 \times [0,\infty)$ into $[0,1]^3$ such that with probability one for each s not in a set of measure zero,

$$w(X(s), s) = E^J[J(s)|X(s)].$$

For almost all $s, w(\cdot,\dot{}s)$ is defined for almost all x, since $X(s)$ has a positive density with respect to Lebesgue measure. Fix the values on the exceptional set so that it is defined for all x for almost all s.

Using the definition of $e^{\overline{u},v}(x)$ for each s and with $v(\cdot) = w(\cdot, s)$, (10.18) can be written as

$$\lim_{n\to\infty} \frac{1}{t_n} E^J \int_0^{t_n} [e^{\overline{u},v(s)}(X(s)) - \gamma(k_0, \overline{u}) + \gamma(k_0, J, x_0)]ds = 0.$$

By the optimality of $\overline{u}(\cdot)$ in the class of feedback controls and the above part of the proof, for almost all s $e^{\overline{u},w(s)}(x) \le 0$ for almost all x. Thus $e^{\overline{u},w(s)}(X(s)) \le 0$ w.p.1 for almost all s. Hence the last displayed equation yields

$$\gamma(k_0, \overline{u}) \le \gamma(k_0, J, x_0),$$

which proves that $\overline{u}(\cdot)$ is optimal with respect to all admissible controls. □

The convergence of the costs. We are finally prepared to prove (9.6). The system (10.2) will provide an upper bound to the cost $\overline{\gamma}(x)$. First, we need to introduce a "lower bounding" system. This will be of the form (10.2) but with (effectively) a slightly larger state space. Let $\delta/\mu > \delta_{k_0}$, let $\overline{\delta}$ denote the vector with components δ, and define the "δ-perturbed" form of (10.2) for state feedback $u(\cdot)$:

(10.19) $$dX = bdt + \overline{\delta}dt - \mu X dt + dY + dW + k(X)dt - K(X)u(X)dt, \qquad X^i(t) \ge 0.$$

Let $\gamma(k_0, \delta, u)$ denote the cost (10.3) for this system. We note the fact that the ergodic cost under state feedback controls does not depend on the initial condition. Set $\overline{\gamma}(k_0, \delta) = \inf \gamma(k_0, \delta, u)$, where the inf is over all state feedback controls. By Theorem 10.6, the inf is the same over admissible controls.

THEOREM 10.7. *Let* $\sup_N E|X_N(0)|^2 < \infty$. *Then* $\overline{\gamma}(x)$ *does not depend on* x *and* (*writing* $\overline{\gamma}(x) = \overline{\gamma}$)

(10.20) $$\overline{\gamma}_T^N(x) \to \overline{\gamma},$$

as $N \to \infty, T \to \infty$.

Proof. First, we get a "lower boundary" system. Let $\overline{E}^u$ denote the expectation operator for functionals of the stationary process (10.2), under state feedback control $u(\cdot)$. We have

(10.21) $$\overline{\gamma}(k_0, \delta) \le \overline{\gamma} \le \overline{\gamma}(k_0) + \sum_i \overline{E}^{\overline{u}(k_0)} Y^i(1),$$

where $\overline{u}(k_0, \cdot)$ is the optimal control for cost $\gamma(k_0, u)$ and system (10.2). The right side of (10.21) should be clear. To see the left-hand side, proceed as follows: First, define the system (10.2) on the spacial set $[-\delta/\mu, \infty)$ instead of on $[0, \infty)$; i.e., the reflection is at $-\delta/\mu$. Let $\Delta_0 - \delta/\mu$ replace Δ_0. Write the resulting system as

$$\begin{aligned} d\tilde{X}^i =& b_i dt + dW^i + d\tilde{Y}^i - \mu\tilde{X}^i dt + k_1(\tilde{X}^i + \delta/\mu)dt \\ &-k_1(\tilde{X}^{(i+1)} + \delta/\mu)\tilde{J}^{(i+1)}dt - k_1(\tilde{X}^{(i+2)} + \delta/\mu)\tilde{J}^{(i+2)}dt. \end{aligned}$$

Recall that $\delta/\mu > \delta_0$. Then clearly the inf of the cost $\gamma(k_0, \delta, u)$ for the new system is no greater than $\bar{\gamma}$. (Loosely speaking, this is true since trunk i of the new system "rejects and requests rerouting" at rate $k_1(x^i + \delta/\mu)$; i.e., it "rejects and requests rerouting" only when the state value is negative, and the reflection term is not counted in the cost.) Now change variables $x^i = \tilde{x}^i + \delta/\mu$ to get (10.19) and the cost $\gamma(k_0, \delta, u)$.

It is also true that

$$\bar{\gamma}(k_0, \delta) - \bar{\gamma}(k_0) \to 0 \tag{10.22}$$

as $\delta \to 0$, uniformly in k_0. In fact, the difference is bounded above by 3δ. (A remark on the network analog of this appears after the proof.) To see this bound, add a nondecreasing control function $H(\cdot)$ satisfying $H^i(t) \le \delta t, H^i(0) = 0$ to (10.2), and use the same cost $\gamma(k_0, J, x)$. The best that we can do with the additional control is to reduce the minimum ergodic cost by 3δ.

Now that we have a bounding system (10.2), we can proceed to get a "nice" ε-optimal control for this system. Let $\varepsilon > 0$, and let k_0 and δ_{k_0} satisfy the requirements of Theorem 10.5. By Theorem 10.5, we can choose k_0 large enough such that $\sum_i \overline{E}^u Y^i(1) \le \varepsilon$ for all $u(\cdot)$. Suppose that $|\bar{\gamma}(k_0, \delta) - \bar{\gamma}(k_0)| \le \varepsilon$, and let $u_\varepsilon(\cdot) = (u^1_\varepsilon(\cdot), u^2_\varepsilon(\cdot), u^3_\varepsilon(\cdot))$ be a state feedback Lipschitz continuous ε-optimal for cost $\gamma(k_0, u)$. Such a control exists by Theorem 10.4 together with the fact (Theorem 10.6) that $\bar{\gamma}(k_0)$ is the infimum over both all admissible controls and the state feedback controls.

Now we adapt $u_\varepsilon(\cdot)$ to the physical network by "randomizing" as done in Theorem 8.3. Recall that we supposed that only one arrival/departure event can occur in the network at a time (w.p.1). Define $\text{data}^N(t) = \{A^N(s), D^N(s),\ s \le t,$ rerouting decisions for $s < t.\}$ Then adapt $u_\varepsilon(\cdot)$ as follows. Let N be large enough so that $k_0 < \sqrt{N}$. Suppose that there is an external arrival to trunk i at time t. If the trunk is full, reject it with no request for rerouting. Otherwise, reject from i and request rerouting with the conditional probability

$$P\{\text{reject from } i \text{ and request rerouting}|\text{data}^N(t)\} = k_1(X^{i,N}(t^-))/\sqrt{N}.$$

The request is accepted on the alternative route with conditional probability

$$P\{\text{request from } i \text{ accepted on alternate route}|\text{data}^N(t)\} = u^i_\varepsilon(X^N(t^-)).$$

Since the events are separated in time, the conditional probability is well defined.

The rest of the details are similar to those in Theorems 8.3 and 9.1, and only a few comments will be made. Analogous to the case at the end of the proof of Theorem 8.3 where the rerouting policy was also randomized, we can write $(1/\sqrt{N})\times$(number of rerouting requests made by trunk i by time t) as the sum of the compensator or conditional mean value, which is (mod $O(1/\sqrt{N})$)

$$\int_0^t k_1(X^{i,N}(s))ds,$$

and a martingale ("noise") term. Similarly, $1/\sqrt{N}\times$(number of rerouting requests made by trunk i by t that are accepted by the alternative path) can be written as

$$\int_0^t k_1(X^{i,N}(s))u_\varepsilon^i(X(s))ds,$$

plus a "noise" term. As in Theorem 8.3, the variances of these "noise" terms go to zero as $N \to 0$.

Let $X^N(0) = x$, and define

(10.23)
$$\gamma_T^N(k_0, u_\varepsilon, x) = \frac{1}{T}E\left[\int_0^t \sum_i (1 - u_\varepsilon^i(X^N(s)))k_1(X^{i,N}(s))ds + \sum_i Y^{i,N}(T)\right].$$

Then, obviously,

$$\gamma_T^N(k_0, u_\varepsilon, x) \geq \bar{\gamma}_T^N(x). \tag{10.24}$$

By an occupation measure argument of the type required for Theorem 5.1, we get that

$$\gamma_T^N(k_0, u^\varepsilon, x) \to \gamma(k_0, u^\varepsilon) \tag{10.25}$$

as $N \to \infty$, $T \to \infty$. Now (10.21), (10.24), (10.25), and the fact that the right-hand sum in (10.21) and the difference in (10.22) are both less than ε give

$$\limsup_{N,T} \bar{\gamma}_T^N(x) \leq \bar{\gamma} + 3\varepsilon.$$

This together with (9.4) yields the theorem, since ε is arbitrary. □

Note. It is worth adding a comment on the analog of (10.19), (10.22) for the network. Adding δ to b_i is equivalent to cutting out external inputs to each trunk randomly so that the effective (external) arrival rate to each is reduced by $\delta\sqrt{N}$. This reduces the cost due to subsequent losses to the three-trunk network by at most 3δ. An alternative interpretation is provided by simply increasing the number of lines in each trunk by $\delta\sqrt{N}/\gamma$ and using the definition $X^{i,N}(t) = [\beta_i N + \delta\sqrt{N}/\mu -$number occupied $]/\sqrt{N}$. Then the mean total loss is also reduced by at most 3δ. If $\delta/\mu > \delta_0$, then the rejections/requests for rerouting from trunk i can occur only when the original $\beta_i N$ lines are fully occupied. This implies that $\bar{\gamma}(k_0, \delta) \leq \bar{\gamma}$.

REFERENCES

[1] P. Billingsley, *Convergence of Probability Measures*, John Wiley, New York, 1968.

[2] J. M. Bismut, *Théorie probabiliste du contrôle des diffusions*, Mem. Amer. Math. Soc. 167, American Mathematical Society, Providence, RI, 1976.

[3] J. L. Doob, *Asymptotic properties of Markov transition probabilities*, Trans. Amer. Math. Soc., 63 (1948), pp. 393–421.

[4] S. N. Ethier and T. G. Kurtz, *Markov Processes: Characterization and Convergence*, Wiley, New York, 1986.

[5] J. M. Harrison and R. J. Williams, *Brownian models of open queueing networks with homogeneous customer populations*, Stochastics 22 (1987), pp. 77–115.

[6] P. J. Hunt, *An asymptotic analysis of a single link network*, preprint, University of Cambridge, 1989.

[7] ———, *On the asymptotic behavior of loss networks*, Technical report, University of Cambridge, 1989. Smith's prize essay.

[8] N. Ikeda and S. Watanabe, *Stochastic Differential Equations and Diffusion Processes*, First Ed., North-Holland, Amsterdam, 1981.

[9] S. Stidham Jr., *Optimal control of admission to a queueing system*, IEEE Trans. on Automatic Control, AC-30 (1985), pp. 705–713.

[10] F. P. Kelly, *Routing and capacity allocation in networks with trunk reservation*, Math. Oper. Res., 15 (1990), pp. 771–793.

[11] ———, *Loss networks*, Ann. Appl. Probab., 1 (1991), pp. 319–377.

[12] H. J. Kushner, *Optimality conditions for the average cost per unit time problem with a diffusion model*, SIAM J. Control Optim., 16 (1978), pp. 330–346.

[13] ———, *Numerical methods for stochastic control problems in continuous time*, SIAM J. Control Optim., 28 (1990), pp. 999–1048.

[14] ———, *Weak Convergence Methods and Singularly Perturbed Stochastic Control and Filtering Problems*, volume 3 of Systems and Control, Birkhäuser, Boston, 1990.

[15] H. J. Kushner and P. Dupuis, *Numerical Methods for Stochastic Control Problems in Continuous Time*, Springer, New York and Berlin, 1992.

[16] H. J. Kushner and L. F. Martins, *Numerical methods for stochastic singular control problems*, SIAM J. Control Optim., 29 (1991), pp. 1443–1475.

[17] H. J. Kushner and K. M. Ramachandran, *Optimal and approximately optimal control policies for queues in heavy traffic*, SIAM J. Control Optim., 27 (1989), pp. 1293–1318.

[18] H. J. Kushner and J. Yang, *Numerical methods for controlled routing in large trunk line systems via stochastic control theory*, Technical report, Brown University, Lefschetz Center for Dynamical Systems, 1992; to appear in the ORSA Journal on Computing.

[19] L. F. Martins and H. J. Kushner, *Routing and singular control for queueing networks in heavy traffic*, SIAM J. Control Optim., 28 (1990), pp. 1209–1233.

[20] D. Mitra and R. J. Gibbens, *State dependent routing on symmetric loss networks with trunk reservations: Asymptotics, optimal design*, Ann. Oper. Res., 35 (1992), pp. 3–30.

[21] D. Mitra, R. J. Gibbens, and B. D. Huang, *Analysis and optimal design of aggregated-least-busy-alternative routing on symmetric loss networks with trunk reservation*, in Teletraffic and Datatraffic in a Period of Change, A. Jensen and V.B. Jensen, eds., North-Holland–Elsevier, Amsterdam, 1991.

[22] V. Nguyen, *On the optimality of trunk reservation in overflow processes*, Probab. Engrg. Inform. Sci., 5 (1991), pp. 369–390.

[23] T. J. Ott and K. R. Krishnan, *Separable routing: A scheme for state dependent routing of circuit switched telephone traffic*, Ann. Oper. Res., 35 (1992), pp. 43–68.

[24] M. I. Reiman, *Open queueing networks in heavy traffic*, Math. Oper. Res., 9 (1984), pp. 441–458.

[25] ———, *Asymptotically optimal trunk reservation for large trunk groups*, Proc. 28th Conference on Decision and Control, New York, IEEE, 1989.

[26] ———, *Optimal trunk reservations for a critically loaded line*, in Teletraffic and Datatraffic in a Period of Change, A. Jensen and V.B. Jensen, eds., North-Holland–Elsevier, Amsterdam, 1991.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 33, No. 3, pp. 804–833, May 1995

© 1995 Society for Industrial and Applied Mathematics

008

ON THE STABILIZATION IN FINITE TIME OF LOCALLY CONTROLLABLE SYSTEMS BY MEANS OF CONTINUOUS TIME-VARYING FEEDBACK LAW*

JEAN-MICHEL CORON†

This paper is dedicated to Henry Hermes for his 60th birthday.

Abstract. It is proven that, if, for any positive time T, there exists an open-loop control $u(a,t)$ depending continuously on the initial data a, vanishing for $a=0$, and steering a small neighborhood of 0 into 0 in time T, then the control system can be locally stabilized in small time by means of continuous time-varying feedback law, provided that the dimension of the state space is at least 4 and the strong accessibility rank condition holds.

Key words. stabilization, time-varying feedback, controllability

AMS subject classifications. 93D15, 93C10

1. Introduction and statements of the main results.

Let Σ be the control system

$$\dot{x} = f(x,u), \tag{1.1}$$

where $x \in \mathbb{R}^n$ is the state, $u \in \mathbb{R}^m$ is the control, and $f : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}^n$ satisfies

$$f(0,0) = 0. \tag{1.2}$$

Using small controls we want to steer to 0 states close to 0 and always stay close to 0. This is clearly possible by means of open-loop controls if 0 is locally reachable in small time and with small control, i.e., for any point a in a small neighborhood of 0 in $\mathbb{R}^n$ there exists a small open-loop control $u_a(\cdot)$ that steers this point a to 0 in small time. Unfortunately, for "real" systems it is well known that, for example, due to random perturbations, open-loop controls can lead to very bad practical results; see, e.g., [So2, Chap. 1, §4]. To take care of this problem one usually tries to find a closed-loop control or feedback law that asymptotically stabilizes the system; indeed, compared to open-loop control, such a feedback law has the advantage of compensating automatically for these perturbations—at least if they are small. A classical result is that for a linear system this local reachability implies that Σ can be asymptotically stabilized by means of stationary feedback law; see, e.g., [So2; Thm. 7, p. 134]. Let us recall that M. Kawski has proved in [Kaw1] (see also [DMK]) that this result still holds if f is nonlinear provided that f is analytic, $n=2$, $m=1$, and $f(x,u) = f_0(x)+uf_1(x)$. Unfortunately, it has been shown by R. Brockett in [Br] (see also [Su1, Appendix]) that this is no longer true in the general case, even if f is analytic. The goal of this paper is to show that the situation is much better if, following E. Sontag and H. Sussmann [SS] (see also [Sa] and [Co1]), one allows the feedback law to depend on time. Indeed, we will see that "many" sufficient conditions for this local reachability

*Received by the editors November 18, 1992; accepted for publication November 18, 1993.

†Centre de Mathématiques et de Leurs Applications, Centre National de la Recherche Scientifique, Unité de Recherche Associée 1611, École Normale Supérieure de Cachan, 61, avenue du Président Wilson, 94235 Cachan, France.

(including, e.g., the Hermes condition [He2] or [Su3]) imply that, if the dimension of the state space is at least 4, then Σ can be locally asymptotically—and even in finite time—stabilized by means of a continuous periodic time-varying feedback law.

In order to state precisely our main results let us first introduce three definitions.

DEFINITION 1.1. *The origin (of $\mathbb{R}^n$) is locally reachable (for Σ) in small time and with small control if, for any positive real number T, there exist $u : \mathbb{R}^n \to L^1((0,T);\mathbb{R}^m)$ and a positive real number ε such that*

$$\|u(a)\|_\infty := \text{ess. sup}\{|u(a)(t)|; t \in (0,T)\} \to 0 \quad as\, a \to 0, \tag{1.3}$$

$$(\dot{x} = f(x, u(x(0))(t)) \quad and \quad |x(0)| < \varepsilon) \Rightarrow x(T) = 0. \tag{1.4}$$

If, moreover, u can be chosen in $C^0(\mathbb{R}^n; L^1((0,T);\mathbb{R}^m))$, we say that 0 is locally continuously reachable in small time and with small control.

Of course, in (1.4) x is any *maximal* solution, and by $x(T) = 0$ we mean that $x(T)$ *exists* and is equal to zero. We use these standard conventions throughout this paper (even for vector fields that are only continuous with respect to x). Let us note that, following a method of M. Kawski [Kaw2] (see also [Hel]), we have proved in [Co2, Lem. 3.1 and §5] that "many" sufficient conditions for locally reachability of 0 in small time and with small control imply that 0 is locally *continuously* reachable in small time and with small control. In particular, this is the case for the Hermes condition [He2] or [Su3] and its generalization due to H. J. Sussmann [Su4, Thm. 7.3]; this is, in fact, also the case for the Bianchini and Stefani condition [BS, Cor., p. 970] which extends [Su4, Thm. 7.3]. We conjecture that, if f is analytic, 0 is locally continuously reachable in small time and with small control if 0 is locally reachable in small time and with small control.

DEFINITION 1.2. *System Σ is locally smoothly stabilizable in small time by means of periodic time-varying feedback law if, for any positive real number T, there exists u in $C^0(\mathbb{R}^n \times \mathbb{R}; \mathbb{R}^m)$ of class C^∞ on $(\mathbb{R}^n\backslash\{0\}) \times \mathbb{R}$, vanishing on $\{0\} \times \mathbb{R}$, T-periodic in time, and such that, for some positive real number ε,*

$$((\dot{x} = f(x, u(x,t)) \quad and \quad x(s) = 0) \Rightarrow (x(\tau) = 0 \quad \forall \tau \geq s)) \quad \forall s \in \mathbb{R}, \tag{1.5}$$

$$(\dot{x} = f(x, u(x,t)) \quad and \quad |x(\tau)| < \varepsilon) \Rightarrow (x(\tau + T) = 0) \quad \forall \tau \in \mathbb{R}. \tag{1.6}$$

In particular (see Lemma 2.15) 0 is a uniformly locally asymptotically stable point for $\dot{x} = f(x, u(x,t))$.

Let us recall that 0 is a uniformly locally asymptotically stable point for $\dot{x} = X(x,t)$ with X in $C^0(\mathbb{R}^n \times \mathbb{R}; \mathbb{R}^n)$ if it is a uniformly locally stable point (i.e., $\forall \varepsilon > 0, \exists \delta > 0$ s.t., $\forall s \in \mathbb{R}$, $\dot{x} = X(x,t)$ and $|x(s)| < \delta$ imply $|x(\tau)| < \varepsilon \forall \tau > s$) and a uniformly locally attractive point with respect to time and state (i.e., $\exists \delta > 0$ s.t., $\forall \varepsilon > 0 \, \exists M > 0$ s.t., $\forall s \in \mathbb{R}$, $\dot{x} = X(x,t)$ and $|x(s)| < \delta$ imply $|x(\tau)| < \varepsilon \forall \tau > s+M$). Note that "uniformly locally attractive" follows directly from (1.5) and (1.6); for "uniformly locally stable," see Lemma 2.15. Our last definition is as follows.

DEFINITION 1.3 [Co3]. *System Σ satisfies the strong jet accessibility rank condition at $(\bar{x}, \bar{u})$ if*

$$a(\bar{x},\bar{u}) := \text{Span}\left\{h(\bar{x}); h \in \left\{\frac{\partial^{|\alpha|} f}{\partial u^\alpha}(\cdot,\bar{u}), \alpha \in \mathbb{N}^m, |\alpha| \geq 1\right\} \cup \text{Br}(f,\bar{u})\right\} = \mathbb{R}^n, \tag{1.7}$$

where $\mathrm{Br}(f,\bar{u})$ *is the set of iterated Lie brackets of vector fields in* $\{(\partial^{|\alpha|}f/\partial u^{\alpha})(\cdot,\bar{u});\alpha\in\mathbb{N}^m\}$.

Let us point out that, in general, $f(\bar{x},\bar{u})$ does not belong to $a(\bar{x},\bar{u})$: we do not consider $\partial^{|\alpha|}f/\partial u^{\alpha}$ as an iterated Lie bracket. Let us also note that the sufficient conditions for local reachability of 0 in small time and with small control mentioned above imply that the strong jet accessibility rank condition at $(0,0)$ holds. Finally, note that, if Σ is locally smoothly stabilizable in small time by means of periodic time-varying feedback law, then $0\in\mathbb{R}^n$ is locally continuously reachable in small time and with small control. Our main result is that the converse holds if $n\geq 4$ and if Σ satisfies the strong jet accessibility rank condition at $(0,0)$, i.e., we have the following theorem.

THEOREM 1.4. *Assume* $n\geq 4$ *and*

(1.8) *0 is locally continuously reachable in small time and with small control*,

(1.9) Σ *satisfies the strong jet accessibility rank condition at* $(0,0)$.

Then Σ *is locally smoothly stabilizable in small time by means of periodic time-varying feedback law.*

This theorem is proved in §2. We do not know if the assumption $n\geq 4$ can be removed. It is proved in [Co4] that this is the case if $n=1$ and f is analytic—in this case the stabilizing feedback law can be chosen independent of time. In §3 we give other cases where the assumption $n\geq 4$ can be removed; in particular, we prove the following propositions.

PROPOSITION 1.5. *If* $f(x,u)=\sum_{i=1}^{m}u_if_i(x)$ *for some functions* f_i *in* $C^\infty(\mathbb{R}^n;\mathbb{R}^n)$ *with* $i\in[1,m]$, *and if* (1.9) *holds, then* Σ *is locally smoothly stabilizable in small time by means of periodic time-varying feedback law.*

PROPOSITION 1.6. *Assume that* (1.8) *and* (1.9) *hold. Assume that, with* $x=(x_1,x_2)\in\mathbb{R}^{n-1}\times\mathbb{R}$ *and* $u=(u_1,u_2)\in\mathbb{R}^{m-1}\times\mathbb{R}$, $f(x,u)=(f_1(x_1,u_1),u_2)\in\mathbb{R}^{n-1}\times\mathbb{R}$ *for some function* f_1 *in* $C^\infty(\mathbb{R}^{n-1}\times\mathbb{R}^{m-1};\mathbb{R}^{n-1})$. *Then* Σ *is locally smoothly stabilizable in small time by means of periodic time-varying feedback law.*

Remark 1.7. (a) Theorem 1.4 is related to the previous result [Su1] by Sussmann, where, roughly speaking, it is proved that controllability implies that the system can be steered to the origin by means of discontinuous—stationary—feedback law. Note that [Su1] has the advantage of leading to a global result, which is not the case with our theorem; on the other hand, our feedback law is continuous and 0 is uniformly locally asymptotically stable for the closed-loop system, which gives some kind of robustness, since, by a theorem of J. Kurzweil [Ku], it implies the existence of a Lyapunov function. (b) One easily checks that the usual strong accessibility subspace of Σ at $\bar{x}$ (see, e.g., [SJ, p. 101] or [So1, p. 549]) contains $a(\bar{x},\bar{u})$ for all $\bar{u}$ in $\mathbb{R}^m$ and that, if f is a polynomial with respect to u or if f is analytic with respect to x and u, these inclusions are all equalities; hence, by a theorem of H. Sussmann and V. Jurdjevic [SJ], (1.8) implies (1.9) if f is analytic. (c) For f as in Proposition 1.6, (1.9) implies (1.8) (see [Co2]), and it has already been proved in [Co3] that for such an f, (1.9) implies that Σ is uniformly locally asymptotically stabilizable by means of periodic time-varying feedback law of class C^∞. (d) This paper is a detailed and improved version of [Co2], where a weaker but related result [Co2, Thm. 1.8] has been stated with a sketch of proof. (e) The stabilization problem is a field of research that expands very quickly; for a survey on this subject see [Ba].

2. Proof of Theorem 1.4. Let I be an interval of $\mathbb{R}$. By a trajectory of the control system Σ on I we mean $(\gamma, u) \in C^\infty(I; \mathbb{R}^n \times \mathbb{R}^m)$ satisfying $\dot\gamma(t) = f(\gamma(t), u(t))$ for all t in I. The linearized control system around (γ, u) is $\dot\xi = A(t)\xi + B(t)w$, where the state is $\xi \in \mathbb{R}^n$, the control is $w \in \mathbb{R}^m$, and $A(t) = \partial f/\partial x(\gamma(t), u(t)) \in \mathcal{L}(\mathbb{R}^n; \mathbb{R}^n)$, $B(t) = \partial f/\partial u(\gamma(t), u(t)) \in \mathcal{L}(\mathbb{R}^m; \mathbb{R}^n)$ for all t in I. We first introduce the following definition.

DEFINITION 2.1. *The trajectory (γ, u) is supple on $S \subset I$ if, for all s in S,*

$$\text{Span}\{((d/dt) - A(t))^i B(t)|_{t=s} w;\ w \in \mathbb{R}^m, i \geq 0\} = \mathbb{R}^n. \tag{2.1}$$

In (2.1) we use the classical convention $(d/dt - A(t))^0 B(t) = B(t)$. Let us recall that L. Silverman and H. Meadows have shown in [SM] that (2.1) implies that the linearized control system around (γ, u) is controllable with impulsive controls at time s (in the sense of [Kai, p. 614]). Let T be a positive real number. For u in $C^0(\mathbb{R}^n \times [0,T]; \mathbb{R}^m)$ and a in $\mathbb{R}^n$, let $x(a, \cdot; u)$ be the maximal solution of $\partial x/\partial t = f(x, u(a,t))$, $x(a, 0; u) = a$. Also, let C^* be the set of $u \in C^0(\mathbb{R}^n \times [0,T]; \mathbb{R}^m)$ of class C^∞ on $(\mathbb{R}^n \backslash \{0\}) \times [0,T]$ and vanishing on $\{0\} \times [0,T]$. Unless otherwise specified we assume (1.8) and (1.9). To try to make the arguments clearer we first sketch the four steps of the proof. For simplicity, in this sketch of proof we omit some details that are important for taking care of the uniqueness property (1.5) (note that without (1.5) one does not have stability).

Step 1. Using (1.8), (1.9), and [Co2] or [Co3], one proves that there exist ε_1 in $(0, +\infty)$ and u_1 in C^*, vanishing on $\mathbb{R}^n \times \{T\}$, such that

$$|a| \leq \varepsilon_1 \Rightarrow x(a, T; u_1) = 0, \tag{2.2}$$

$$0 < |a| \leq \varepsilon_1 \Rightarrow (x(a, \cdot; u_1), u_1(a, \cdot)) \text{ is supple on } [0, T]. \tag{2.3}$$

Step 2. Let Γ be a closed submanifold of $\mathbb{R}^n \backslash \{0\}$ of dimension 1 such that $\Gamma \subset \{x \in \mathbb{R}^n; 0 < |x| < \varepsilon_1\}$ Perturbing u_1 in a suitable way, one obtains a map u_2 in C^*, vanishing on $\mathbb{R}^n \times \{T\}$, such that

$$|a| \leq \varepsilon_1 \Rightarrow x(a, T; u_2) = 0, \tag{2.4}$$

$$0 < |a| \leq \varepsilon_1 \Rightarrow (x(a, \cdot; u_2), u_2(a, \cdot)) \text{ is supple on } [0, T), \tag{2.5}$$

$$a \in \Gamma \to x(t, a; u_2) \in \mathbb{R}^n \text{ is an embedding of } \Gamma \text{ into } \mathbb{R}^n \backslash \{0\} \quad \forall t \in [0, T]. \tag{2.6}$$

Here one uses the assumption $n \geq 4$ and proceeds as in the classical proof of the Whitney embedding theorem (see, e.g., [GG, Chap. II, §5]). Let us emphasize that we use this assumption only in this step.

Step 3. From Step 2 one deduces the existence of u_3^* in C^*, vanishing on $\mathbb{R}^n \times \{T\}$, and an open neighborhood $\mathcal{N}^*$ of Γ in $\mathbb{R}^n \backslash \{0\}$ such that

$$a \in \mathcal{N}^* \Rightarrow x\,(a, T; u_3^*) = 0, \tag{2.7}$$

$$a \in \mathcal{N}^* \to x(a, t; u_3^*) \text{ is an embedding of } \mathcal{N}^* \text{ into } \mathbb{R}^n \backslash \{0\} \quad \forall t \in [0, T). \tag{2.8}$$

This embedding property allows one to transform, on $\{(x(a,t;u_3),t); a \in \mathcal{N}, t \in [0,T)\}$, the open-loop control u_3^* into a feedback law u_3. So (see, in particular, (2.7) and note that u_3^* vanishes on $\mathbb{R}^n \times \{T\}$) there exist u_3 in C^* and an open neighborhood $\mathcal{N}$ of Γ in $\mathbb{R}^n \backslash \{0\}$ such that

$$(x(0) \in \mathcal{N} \text{ and } \dot{x} = f(x, u_3(x,t))) \Rightarrow (x(T) = 0). \tag{2.9}$$

One can also impose, for all τ in $[0,T]$,

$$(\dot{x} = f(x, u_3(x,t)) \text{ and } x(\tau) = 0) \Rightarrow (x(t) = 0 \quad \forall t \in [\tau, T]). \tag{2.10}$$

Step 4. In this last step one shows the existence of a closed submanifold of $\mathbb{R}^n \backslash \{0\}$ of dimension 1 included in $\{x \in \mathbb{R}^n; 0 < |x| < \varepsilon_1\}$ such that, for any neighborhood $\mathcal{N}$ of Γ in $\mathbb{R}^n \backslash \{0\}$, there exists u_4 in C^* such that, for some ε_4 in $(0, +\infty)$,

$$(\dot{x} = f(x, u_4(x,t)) \text{ and } |x(0)| < \varepsilon_4) \Rightarrow (x(T) \in \mathcal{N} \cup \{0\}), \tag{2.11}$$

$$((\dot{x} = f(x, u_4(x,t)) \text{ and } x(\tau) = 0) \Rightarrow (x(t) = 0) \quad \forall t \in [\tau, T]))\forall \tau \in [0,T]. \tag{2.12}$$

Finally, let $u : \mathbb{R}^n \times \mathbb{R} \to \mathbb{R}^m$ be equal to u_4 on $\mathbb{R}^n \times [0,T]$, $2T$-periodic with respect to time, and such that $u(x,t) = u_3(x, t-T)$ for all (x,t) in $\mathbb{R}^n \times (T, 2T)$. Then u vanishes on $\{0\} \times \mathbb{R}$, is continuous on $\mathbb{R}^n \times (\mathbb{R} \backslash \mathbb{Z}T)$, is of class C^∞ on $(\mathbb{R}^n \backslash \{0\}) \times (\mathbb{R} \backslash \mathbb{Z}T)$, and satisfies

$$(\dot{x} = f(x, u(x,t)) \text{ and } |x(0)| < \varepsilon_4) \Rightarrow (x(2T) = 0), \tag{2.13}$$

$$(\dot{x} = f(x, u(x,t)) \text{ and } x(\tau) = 0) \Rightarrow (x(t) = 0, \quad \forall t \geq \tau) \quad \forall \tau \in \mathbb{R}, \tag{2.14}$$

which implies, as we will see, that (1.6) holds, with $4T$ instead of T and $\varepsilon > 0$ small enough, and that 0 is uniformly locally asymptotically stable for $\dot{x} = f(x, u(x,t))$. Since T is arbitrary, Theorem 1.4 is proved (modulo a problem of regularity of u at (x,t) in $\mathbb{R}^n \times \mathbb{Z}T$ that we will fix).

We now give the proofs of Steps 1 to 4.

Proof of Step 1. We prove the existence of u_1. For a positive real number ε, let $B_\varepsilon = \{x \in \mathbb{R}^n; |x| < \varepsilon\}$ and $B'_\varepsilon = B_\varepsilon \backslash \{0\}$. Let T be a positive real number. The goal of this step is to prove the following proposition.

PROPOSITION 2.2. *There exist u_1 in C^*, vanishing on $\mathbb{R}^n \times \{T\}$, and ε_1 in $(0, +\infty)$ such that* (2.2) *and* (2.3) *hold, and*

$$u_1/|a|^2 \in L^\infty(B'_1 \times [0, \tau]) \quad \forall \tau \in [0, T). \tag{2.15}$$

By (2.15) we mean that, for all τ in $[0,T)$, there exists M in $[0, +\infty)$ such that $|u_1(a,t)| \leq M|a|^2$ for all (a,t) in $B'_1 \times [0, \tau]$. To prove Proposition 2.2 one first proves the following lemma.

LEMMA 2.3. *There exists u_0 in $C^\infty(\mathbb{R}^n \times [0,T); \mathbb{R}^m \cap C^0(\mathbb{R}^n \times [0,T]; \mathbb{R}^m)$, vanishing on $\mathbb{R}^n \times \{T\}$ and on a neighborhood of $\{0\} \times [0,T)$ in $\mathbb{R}^n \times [0,T)$, such that, for some ε_0 in $(0, +\infty)$, $x(a, T; u_0) = 0$ for all a in $\mathbb{R}^n$ with $|a| \leq \varepsilon_0$.*

Proof. Without loss of generality we may assume that $T = 1$. Let i be a positive integer. Since $0 \in \mathbb{R}^n$ is locally continuously reachable in small time and with small control, there exist v_i in $C^0(\mathbb{R}^n; L^1((1-2^{1-i}, 1-2^{-i}); \mathbb{R}^m))$ and a positive real number δ_i in $(0, 1/i)$ such that

$$\|v_i(a)\|_\infty \leq 1/i \qquad \forall a \in \mathbb{R}^n, \tag{2.16}$$

$$\|v_i(a)\|_\infty \to 0 \quad \text{as } a \to 0, \tag{2.17}$$

$$(\dot{x} = f(x, v_i(x(1-2^{1-i}))(t)) \text{ and } |x(1-2^{1-i})| \leq \delta_i) \Rightarrow (x(1-2^{-i}) = 0). \tag{2.18}$$

Let K be the set of u in $\mathbb{R}^m$ of norm less than or equal to 2, and let, for a in $\mathbb{R}^n$ and v in $L^1((1-2^{1-i}, 1-2^{-i}); K)$, $P(a, v) = y(a, 1-2^{-i}; v) \in \mathbb{R}^n$, where y is defined by $\partial y/\partial t = f(y, v(t))$, $y(a, 1-2^{1-i}) = a$. Since K is compact, the map P is defined on an open subset of $\mathbb{R}^n \times L^1((1-2^{1-i}, 1-2^{-i}); K)$— $L^1((1-2^{1-i}, 1-2^{-i}); K)$ being equipped with the L^1 norm—and is continuous on this subset; so, by the continuity of v_i, (2.16), (2.17), and (2.18), there exists w_i in $C^\infty(\mathbb{R}^n \times [1-2^{1-i}, 1-2^{-i}]; \mathbb{R}^m)$ such that w_i vanishes on a neighborhood of $\{0\} \times [1-2^{1-i}, 1-2^{-i}]$ in $\mathbb{R}^n \times [1-2^{1-i}, 1-2^{-i}]$ and

$$\|w_i\|_\infty \leq 2/i, \tag{2.19}$$

$$\partial^\alpha w_i = 0 \quad \text{on} \quad \mathbb{R}^n \times \{1-2^{1-i}, 1-2^{-i}\} \qquad \forall \alpha \in \mathbb{N}^{n+1}, \tag{2.20}$$

$$(\dot{x} = f(x, w_i(x(1-2^{1-i}), t)), |x(1-2^{1-i})| \leq \delta_i) \Rightarrow (|x(1-2^{-i})| < \delta_{i+1}). \tag{2.21}$$

We now define w on $B_{\delta_1} \times [0, 1-2^{-i}]$ by induction on $i \geq 2$ by requiring $w = w_1$ on $B_{\delta_1} \times [0, 1/2]$ and $w(a, t) = w_i(x(a, 1-2^{1-i}; \bar{u}_0), t)$ if t is in $[1-2^{1-i}, 1-2^{-i}]$. One easily checks that w is well defined on $B_{\delta_1} \times [0, 1)$ and is of class C^∞ on this set. Using (2.19) one deduces that there exists u_0 in $C^\infty(\mathbb{R}^n \times [0, 1); \mathbb{R}^m) \cap C^0(\mathbb{R}^n \times [0, 1]; \mathbb{R}^m)$ vanishing on $\mathbb{R}^n \times \{T\}$ and equal to w on $B_{\delta_1/2} \times [0, 1)$. Then u_0 satisfies all the properties mentioned in Lemma 2.3—with $\varepsilon_0 = \delta_1/2$. □

Our next lemma is as follows.

LEMMA 2.4. *There exists* $\bar{u}_0$ *in* $C^\infty(\mathbb{R}^n \times [0, T]; \mathbb{R}^m)$ *and* $\bar{\varepsilon}_0$ *in* $(0, +\infty)$ *such that*

$$|\bar{u}_0(a, t)| \leq |a|^2 \qquad \forall (a, t) \in \mathbb{R}^n \times [0, T], \tag{2.22}$$

$$(x(a, \cdot; \bar{u}_0), \bar{u}_0(a, \cdot)) \text{ is supple on } [0, T) \qquad \forall a \in B'_{\bar{\varepsilon}_0}, \tag{2.23}$$

$$\partial^\alpha \bar{u}_0 = 0 \quad \text{on} \quad \mathbb{R}^n \times \{T\} \qquad \forall \alpha \in \mathbb{N}^{n+1}. \tag{2.24}$$

Proof. Let us first recall the definition of saturation introduced in [Co3, Def. 1.2]. Let p be a positive integer, let X be in $C^\infty(\mathbb{R}^p \times \mathbb{R}^m; \mathbb{R}^p)$, let N be an open subset of $\mathbb{R}^p$, and let u be in $C^\infty(N; \mathbb{R}^m)$. We say that u saturates X on N if

$$a^X(\bar{y}, u(\bar{y})) = \operatorname{Span}\left\{\left(ad^k_{X_u} \frac{\partial X}{\partial u_i}(\cdot, u(\cdot))\right)(\bar{y}); i \in [1, m], k \geq 0\right\} \qquad \forall \bar{y} \in N \tag{2.25}$$

with $X_u(y) = X(y, u(y))$, and where a^X is defined by replacing f by X in the definition of a given in (1.7). Note that the right-hand side of (2.25) is the strong accessibility algebra evaluated at time 0 of the linearized control system around the trajectory $(\gamma, u(\gamma))$ of X where γ is defined by $\dot{\gamma} = X(\gamma, u(\gamma)), \gamma(0) = \bar{y}$. The right-hand side of (2.25) is always included in the left-hand side of (2.25).

We take $p = 2n+1$ and, for all $(x, a, s, u) \in \mathbb{R}^n \times \mathbb{R}^n \times \mathbb{R} \times \mathbb{R}^m \simeq \mathbb{R}^{2n+1} \times \mathbb{R}^m$, let $X_0(x, a, s, u) = (f(x, u), 0, 1) \in \mathbb{R}^n \times \mathbb{R}^n \times \mathbb{R} \simeq \mathbb{R}^{2n+1}$. For the control system $\dot{y} = X_0(y, u)$, the state is $y = (x, a, s)$ in $\mathbb{R}^n \times \mathbb{R}^n \times \mathbb{R} \simeq \mathbb{R}^{2n+1}$ and the control is u in $\mathbb{R}^m$. We take $N = \mathbb{R}^n \times (\mathbb{R}^n \backslash \{0\}) \times (-\infty, T)$. Let h in $C^\infty(\mathbb{R}^n \times \mathbb{R}^n \times \mathbb{R}; \mathbb{R}^n \times \mathbb{R})$ be defined by $h(x, a, s) = (a, s)$ for all (x, a, s) in $\mathbb{R}^n \times \mathbb{R}^n \times \mathbb{R}$. Note that

$$(2.26) \qquad h'(x, a, s)X_0(x, a, s, u) \neq 0 \qquad \forall (x, a, s, u) \in \mathbb{R}^n \times \mathbb{R}^n \times \mathbb{R} \times \mathbb{R}^m.$$

Let $Y = (\mathbb{R}^n \backslash \{0\}) \times (-\infty, T)$ and let Ω be the set of u in $C^\infty(Y; \mathbb{R}^m)$ such that

$$(2.27) \quad |\partial^\alpha u(a, s)| < |a|^2(T-s)/T \qquad \forall (a, s, \alpha) \in Y \times \mathbb{N}^{n+1} \text{ with } |\alpha||a|(T-s) \leq 1.$$

We equip $C^\infty(Y; \mathbb{R}^m)$ with the C^∞-topology defined in [Co3, §1]. For convenience let us recall the definition of this topology. Let $\mathcal{O}$ be an open subset of $\mathbb{R}^p$ and $\bar{v}$ be in $C^\infty(\mathcal{O}; \mathbb{R}^q)$; V is a neighborhood of $\bar{v}$ if there exist two maps, A in $C^0(\mathcal{O}; (0, +\infty))$ and ε in $C^0(\mathcal{O}; (0, +\infty))$, such that all v in $C^\infty(\mathcal{O}; \mathbb{R}^q)$, satisfying $|\partial^\alpha(v - \bar{v})(x)| \leq \varepsilon(x)$ for all (x, α) in $\mathcal{O} \times \mathbb{N}^p$ such that $|\alpha| \leq A(x)$, are in V. Let us remark that Ω is an open neighborhood of 0 in $C^\infty(Y; \mathbb{R}^m)$. Applying [Co3, Thm. 1.3]—note in particular (2.26)—we get the existence of $\bar{u}$ in Ω such that

$$(2.28) \qquad \bar{u} \circ h \text{ saturates } X_0 \text{ on } N.$$

Let $\bar{u}_0 : \mathbb{R}^n \times [0, T] \to \mathbb{R}^m$ be defined by $\bar{u}_0 = \bar{u}$ on Y and $\bar{u}_0 = 0$ on $(\{0\} \times [0, T]) \cup (\mathbb{R}^n \times \{T\})$. By (2.27) $\bar{u}_0$ is of class C^∞ on $\mathbb{R}^n \times [0, T]$ and satisfies (2.22) and (2.24). The existence of $\bar{\varepsilon}_0$ in $(0, +\infty)$ such that (2.23) holds follows from (1.2), (1.9), (2.22), and (2.28). □

From Lemmas 2.3 and 2.4 we will deduce the following lemma.

LEMMA 2.5. *There exists $\bar{u}_1$ in C^*, vanishing on $\mathbb{R}^n \times \{T\}$, and $\bar{\varepsilon}_1$ in $(0, +\infty)$ such that* (2.2) *and* (2.15) *hold with $\bar{u}_1$ and $\bar{\varepsilon}_1$ instead of u_1 and ε_1 and such that*

$$(2.29) \qquad (x(a, \cdot; \bar{u}_1), \bar{u}_1(a, \cdot)) \textit{ is supple on } \{0, T\} \qquad \forall a \in B'_{\bar{\varepsilon}_1}.$$

Before proving Lemma 2.5 let us first explain how to get Proposition 2.2 from Lemma 2.5. Again we apply [Co3, Thm. 1.3] to $(\dot{x}, \dot{a}, \dot{s}) = X_0(x, a, s, u)$ with X_0 still defined as above; h is also unchanged but we define N, Y, and Ω in the following ways: N is now $\mathbb{R}^n \times (\mathbb{R}^n \times \{0\}) \times (0, T)$, $Y = (\mathbb{R}^n \times \{0\}) \times (0, T)$, and Ω is an open neighborhood of $\bar{u}_1$—restricted to Y—in $C^\infty(Y; \mathbb{R}^m)$. We will say that a property P holds if Ω is small enough, if there exists an open neighborhood Ω_1 of $\bar{u}_1$ (for the C^∞-topology) such that P holds if $\Omega \subset \Omega_1$. For example, if Ω is small enough, any u in Ω extended by $\bar{u}_1$ on $(\{0\} \times [0, T]) \cup (\mathbb{R}^n \times \{0, T\})$ is in C^*. Indeed, in this case it suffices for us to choose Ω_1 to be the set of u in $C^\infty(Y; \mathbb{R}^m)$ such that

$$(2.30) \quad |\partial^\alpha(u - \bar{u}_1)(a, s)| < |a|s(T-s) \quad \forall (a, s, \alpha) \in Y \times \mathbb{N}^{n+1} \text{ with } |\alpha||a|s(T-s) \leq 1.$$

Hence, if Ω is small enough, we have, with a slight abuse of notation, $\Omega \subset C^*$. We now assume that $\Omega \subset C^*$. For u in Ω let $y : \mathbb{R}^n \times [0, T] \to \mathbb{R}^n$ be defined by

$\partial y/\partial t = f(y, u(a,t))$, $y(a,T) = 0$. The domain of definition of y is an open subset of $\mathbb{R}^n \times [0,T]$ containing $(\{0\} \times [0,T]) \cup (\mathbb{R}^n \times \{T\})$. One easily checks that, if Ω is small enough, for any u in Ω there exists an open neighborhood ω of 0 in $\mathbb{R}^n$ such that, with φ defined by $\varphi(a) = y(a,0)$, φ is an homeomorphism from ω onto $B_{\bar{\varepsilon}_1/2}$ and φ is a diffeomorphism of class C^∞ from $\omega\backslash\{0\}$ onto $B'_{\bar{\varepsilon}_1/2}$. Let $u_1 \in C^*$, vanishing on $\mathbb{R}^n \times \{T\}$, be such that $u_1(a,t) = u(\varphi^{-1}(a), t)$ for all (a,t) in $B_{\tilde{\varepsilon}_1} \times (0,T)$ with $\tilde{\varepsilon}_1 = \bar{\varepsilon}_1/4$. Then we have (2.2) and, since $\bar{u}_1$ satisfies (2.15), u_1 also satisfies (2.15) if Ω is small enough. Moreover, by (2.29), if Ω is small enough, we have

$$(x(a,\cdot;u_1)); u_1(a,\cdot)) \text{ is supple on } \{0,T\} \quad \forall a \in B'_{\tilde{\varepsilon}_1}. \tag{2.31}$$

Finally, applying [Co3, Thm. 1.3], we get the existence of u in Ω such that $u \circ h$ saturates X_0 on N, which, with (1.2), (1.9), and the fact that u vanishes on $\{0\} \times [0,T]$, implies that, for $|a|$ small enough but not zero, $(y(a,\cdot), u(a,\cdot))$ is supple on $(0,T)$ and, therefore, if $|a|$ is small enough but not zero,

$$(x(a,\cdot;u_1), u_1(a,\cdot)) \text{ is supple on } (0,T). \tag{2.32}$$

Finally, (2.32) and (2.31) imply (2.3) if ε_1 is small enough. This ends the proof of Proposition 2.2 if Lemma 2.5 is proved.

Let us prove Lemma 2.5. We now take $N = \mathbb{R}^n \times (\mathbb{R}^n\backslash\{0\}) \times \mathbb{R}$ and consider the system $(\dot{x}, \dot{a}, \dot{s}) = X_1(x,a,s,u) := (f(x,(s-2T)u), 0, 1) \in \mathbb{R}^n \times \mathbb{R}^n \times \mathbb{R}$. We take $Y = (\mathbb{R}^n\backslash\{0\}) \times \mathbb{R}$, $h(x,a,s) = (a,s)$. By [Co3, Thm. 1.3] there exists v in $C^\infty((\mathbb{R}^n\backslash\{0\}) \times \mathbb{R}; \mathbb{R}^m)$ such that

$$v \circ h \, \text{saturates} \, X_1 \quad \text{on} \, N. \tag{2.33}$$

Let $\tilde{u} \in C^\infty((\mathbb{R}^n\backslash\{0\}) \times \mathbb{R}; \mathbb{R}^m)$ be defined by $\tilde{u}(a,s) = (s-2T)v(a,s)$, then (2.33) gives

$$\tilde{u} \circ h \text{ saturates } X_0 \quad \text{on} \, N. \tag{2.34}$$

Note that

$$\tilde{u}(a, 2T) = 0. \tag{2.35}$$

From (1.9), (2.34), and (2.35) we get that

$$(z(a,\cdot), \tilde{u}(a,\cdot)) \text{ is supple on } \{2T\} \quad \forall a \in \mathbb{R}^n\backslash\{0\}, \tag{2.36}$$

where z is defined by $\partial z/\partial t = f(z, \tilde{u}(a,\cdot))$, $z(a,2T) = 0$. Let $\tilde{\delta}$ be in $C^0(\mathbb{R}^n\backslash\{0\}; (0,+\infty))$. Using Lemma 2.3 with $[T,2T]$ instead of $[0,T]$ and modifying u_0 in a small neighborhood of $(\mathbb{R}^n\backslash\{0\}) \times \{T, 2T\}$ in $(\mathbb{R}^n\backslash\{0\}) \times [T,2T]$, we get the existence of $\tilde{u}_0$ in $C^0(\mathbb{R}^n \times [T,2T]; \mathbb{R}^m)$ of class C^∞ on $(\mathbb{R}^n \times [T,2T]) \backslash\{(0,2T)\}$ such that for some positive real number $\tilde{\varepsilon}_0$,

$$\partial^\alpha(\tilde{u}_0 - \tilde{u}) = 0 \quad \text{on} \quad (\mathbb{R}^n\backslash\{0\}) \times \{2T\} \quad \forall \alpha \in \mathbb{N}^{n+1}, \tag{2.37}$$

(2.38)
$\tilde{u}_0$ vanishes on a neighborhood of$(\{0\} \times [T,2T)) \cup (\mathbb{R}^n \times \{T\})$in $\mathbb{R}^n \times [T,2T)$,

$$0<|a|\le \tilde{\varepsilon}_0 \Rightarrow |\tilde{x}(a,T)-a|<\tilde{\delta}(a), \tag{2.39}$$

where $\tilde{x}$ is defined by $\partial\tilde{x}/\partial t = f(\tilde{x},\tilde{u}_0(a,t))$ and $\tilde{x}(a,2T)=0$. Now let $\bar{u}_0$ and $\bar{\varepsilon}_0$ be as in Lemma 2.4. Using (2.22), (2.23), and Lemma A.1 in Appendix A (with $T_1=0, T_2=T/3, T_3=2T/3, T_4=T, \Lambda=B'_{\bar{\varepsilon}_0}$) we get the following lemma.

LEMMA 2.6. *There exist $\bar{\delta}$ in $C^0(\mathbb{R}^n\backslash\{0\};(0,+\infty))$ and $\tilde{H}\in C^0(\tilde{\omega}\times[0,T];\mathbb{R}^m)$, with $\tilde{\omega}=\{(a,b)\in\mathbb{R}^n\times\mathbb{R}^n; 0<|a|<\bar{\varepsilon}_0/2, |b|<\bar{\delta}(a)\}\cup\{(0,0)\}$, such that*

$$\tilde{H}\in C^\infty((\tilde{\omega}\backslash\{0,0\})\times[0,T];\mathbb{R}^m), \tag{2.40}$$

$$\tilde{H}(a,b,t)=\bar{u}_0(a,t) \quad \forall (a,b)\in\tilde{\omega} \quad \forall t\in[0,T/3]\cup[2T/3,T], \tag{2.41}$$

$$\tilde{H}(a,0,t)=\bar{u}_0(a,t) \quad \forall a\in B_{\bar{\varepsilon}_0/2} \quad \forall t\in[0,T], \tag{2.42}$$

$$|\tilde{H}(a,b,t)|\le 2|a|^2 \quad \forall (a,b,t)\in\tilde{\omega}\times[0,T], \tag{2.43}$$

$$z(a,b,T)=x(a,T;\bar{u}_0)+b \quad \forall (a,b)\in\tilde{\omega}, \tag{2.44}$$

where z is defined by $\partial z/\partial t = f(z,\tilde{H}(a,b,t))$ and $z(a,b,0)=a$.

Let $\theta(a)=x(a,T;\bar{u}_0)$ and let $\tilde{\delta}\in C^0(\mathbb{R}^n\backslash\{0\};(0,+\infty))$ be such that $\tilde{\delta}(\theta(a))\le\bar{\delta}(a)$ if $|a|$ is small enough but positive. Using (2.24), (2.35), (2.37), (2.38), (2.39), and Lemma 2.6 we get that there exists $\bar{u}_1$ in $C^\infty((\mathbb{R}^n\backslash\{0\})\times[0;2T];\mathbb{R}^m)\cap C^0(\mathbb{R}^n\times[0,2T];\mathbb{R}^m)$, which vanishes on $\mathbb{R}^n\times\{2T\}$ and satifies, if $|a|$ is small enough,

$$\bar{u}_1(a,t)=\tilde{H}(a,\tilde{x}(\theta(a),T)-\theta(a),t) \quad \forall t\in[0,T], \tag{2.45}$$

$$\bar{u}_1(a,t)=\tilde{u}_0(\tilde{x}(\theta(a),T),t) \quad \forall t\in(T,2T). \tag{2.46}$$

Using (2.44), (2.45), and (2.46) we get that, if $|a|$ is small enough, $x(a,2T;\bar{u}_1)=0$. Using (2.23), (2.36), (2.37), (2.45), and (2.46) we get that, if $|a|$ is small enough but positive, $x(a,\cdot;\bar{u}_1)$ is supple on $\{0,2T\}$. Finally, using (2.38), (2.43), (2.45), and (2.46), we get that $\bar{u}_1/|a|^2$ is in $L^\infty(B'_1\times[0,\tau])$ for all τ in $[0,2T)$. So $\bar{u}_1$ has the properties required in Lemma 2.5 provided that $\bar{\varepsilon}_1$ is small enough and that one replaces $2T$ with T. Since T is arbitrary in $(0,+\infty)$, Lemma 2.5 is proved. This ends the proof of Proposition 2.2.

Let us end this step with some comments. It follows from our proof of Proposition 2.2 that one can use the weaker result [Co2, Thm. 2.1] instead of [Co3, Thm. 1.3]. Let us recall that this result is related to prior works by M. Gromov [G, Chap. 2, §3.8(E)] and E. Sontag [So1]; moreover, it follows from a recent work [So3] of E. Sontag that this result can be deduced from a theorem on observability of H. Sussmann [Su2] when f is analytic.

Proof of Step 2. Let u_1 and ε_1 be as in Proposition 2.2. Let Γ be a C^∞-closed submanifold of $\mathbb{R}^n\backslash\{0\}$ of dimension 1 such that $\Gamma\subset B'_{\varepsilon_1}$. The goal of this step it to prove the following proposition.

PROPOSITION 2.7. *Assume $n\ge 4$. Then there exists u_2 in C^*, vanishing on $\mathbb{R}^n\times\{T\}$, satisfying* (2.4), (2.5), (2.6), *and, with u_2 instead of u_1,* (2.15).

In this proposition, as well as in the remaining part of this paper, when we refer to embedding we mean embedding of class C^∞. Our proof of Proposition 2.7 is inspired from the classical proof of the Whitney embedding theorem (see, e.g., [GG, Chap. II, §5]). Let Ω be an open neighborhood of $u_{1|B'_{\varepsilon_1}\times(0,T)}$ in $C^\infty(B'_{\varepsilon_1}\times(0,T);\mathbb{R}^m)$. For v in Ω let $\varphi_v : B'_{\varepsilon_1}\to\mathbb{R}^n$ be defined by $\varphi_v(a)=y(0)$, where y is defined by $\dot y = f(y,v(a,t)), y(T)=0$. If Ω is small enough,

$$\text{(2.47)}\qquad \varphi_v \text{ is a diffeomorphism from } B'_{\varepsilon_1} \text{ onto } B'_{\varepsilon_1}\quad \forall v\in\Omega.$$

Let us recall that this means the existence of a neighborhood Ω_1 of $u_{1|B'_{\varepsilon_1}\times(0,T)}$ in $C^\infty(B'_{\varepsilon_1}\times(0,T);\mathbb{R}^m)$ such that (2.47) holds if $\Omega\subset\Omega_1$. From now on we assume that (2.47) holds. Let $u_v:\mathbb{R}^n\times[0,T]\to\mathbb{R}^m$ be defined by $u_v(a,t)=v(\varphi_v^{-1}(a),t)$ if (a,t) is in $B'_{\varepsilon_1}\times(0,T)$, $u_v(a,t)=u_1(a,t)$ if (a,t) is in $B'_{\varepsilon_1}\times(0,T)$. Again, if Ω is small enough, u_v is in C^*, vanishes on $\mathbb{R}^n\times\{T\}$, and satisfies

$$\text{(2.48)}\qquad (x(a,\cdot;u_v),u_v(a,\cdot)) \text{ is supple on } [0,T]\quad \forall a\in B_{\varepsilon_1}.$$

From now on we assume that all these properties hold. Let us recall that a residual subset of Ω is the countable intersection of open dense subsets of Ω. Moreover, since Ω is an open subset of the Baire space $C^\infty(B'_{\varepsilon_1}\times(0,T);\mathbb{R}^m)$, it is a Baire space; therefore, any residual subset of Ω is dense in Ω. So Proposition 2.7 is a consequence of the following lemma, the proof of which is given in Appendix B.

LEMMA 2.8. *If $n\geq 3$, $\Omega_i := \{v\in\Omega; a\in\Gamma\to x(a,t;u_v)$ is an immersion $\forall t\in(0,T)\}$ and $\Omega_0 := \{v\in\Omega; x(a,t;u_v)\neq 0\ \forall a\in\Gamma, \forall t\in(0,T)\}$ are residual subsets of Ω. If $n\geq 4$, $\Omega_1 := \{v\in\Omega; a\in\Gamma\to x(a,t;u_v)$ is one-to-one $\forall t\in(0,T)\}$ is a residual subset of Ω.*

Remark 2.9. (a) Proposition 2.7 remains true if Γ is a closed submanifold of $\mathbb{R}^n\backslash\{0\}$ included in B'_{ε_1} of dimension d provided that $n\geq 2(d+1)$. The proof is similar (use also [GG, Chap. II, Thm. 5.4]). (b) One can also prove Proposition 2.7 by using the method proposed by M. Gromov in [G, Chap. 2, §3.2 (E′)]; but note that our proof does not use the Nash (Newton–Moser) process [G, Chap. 2, §3.2]. (c) Our proof still works if one replaces $[0,T]$ by $[0,T)$ in (2.3).

Proof of Step 3. We assume that Γ is as in Step 2 and that the conclusion of Proposition 2.7 holds (we do not assume $n\geq 4$). The goal of this step is to prove the following proposition.

PROPOSITION 2.10. *There exist a time-varying feedback law u_3 in $C^*(\mathbb{R}^n\times[0,T];\mathbb{R}^m)$ and a neighborhood $\mathcal{N}$ of Γ in $\mathbb{R}^n\backslash\{0\}$ such that (2.9) and (2.10) hold, and*

$$\text{(2.49)}\qquad \partial^\alpha u_3 = 0 \quad on \quad (\mathbb{R}^n\backslash\{0\})\times\{T\}\qquad \forall\alpha\in\mathbb{N}^{n+1}.$$

Let us assume, for the moment, the following lemma.

LEMMA 2.11. *There exist u_3^* in $C^\infty((\mathbb{R}^n\backslash\{0\})\times[0,T];\mathbb{R}^m)\cap C^0(\mathbb{R}^n\times[0,T];\mathbb{R}^m)$ and an open neighborhood $\mathcal{N}^*$ of Γ in $\mathbb{R}^n\backslash\{0\}$ such that* (2.7) *and* (2.8) *hold and*

$$\text{(2.50)}\qquad u_3^* = 0 \quad on \quad (\{0\}\times[0,T])\cup(\mathbb{R}^n\times\{T\}),$$

$$\text{(2.51)}\qquad u_3^*/|a|^2 \in L^\infty(B'_1\times[0,\tau])\quad \forall\tau\in[0,T).$$

Let us prove Proposition 2.10. From (1.2) and (2.51) we get the existence of δ in $(0,\infty)$ such that

$$\text{(2.52)}\qquad \forall\tau\in[0,T)\quad \exists\varepsilon>0 \text{ such that } |x(a,t;u_3^*)|\geq\delta|a|\quad \forall a\in B_\varepsilon\quad \forall t\in[0,\tau].$$

Let $\mathcal{N}_2^* \subset \mathcal{N}_1^* \subset \mathcal{N}^*$ be two bounded open neighborhoods of Γ in $\mathbb{R}^n \backslash \{0\}$ such that, in $\mathbb{R}^n$,

$$\overline{\mathcal{N}_2^*} \subset \mathcal{N}_1^* \cup \{0\}, \quad \overline{\mathcal{N}_1^*} \subset \mathcal{N}^* \cup \{0\}. \tag{2.53}$$

For $i \in \{1,2\}$ let $\mathcal{O}_i = \{(x(a,t;u_3^*),t); a \in \mathcal{N}_i^*, t \in [0,T)\}$. Note that by (2.8) and (2.53) $\mathcal{O}_1$ and $\mathcal{O}_2$ are open in $\mathbb{R}^n \times [0,T]$, $\overline{\mathcal{O}_2}$ is included in $\mathcal{O}_1 \cup (\{0\} \times [0,T])$, and $\overline{\mathcal{O}_1}$ is included in $\mathcal{O}_1 \cup (\{0\} \times [0,T])$. Once more using (2.8) we see that there exists a (unique) $\bar{u}_3$ in $C^\infty(\mathcal{O}_1;\mathbb{R}^n)$ such that

$$\bar{u}_3(x(a,t;u_3^*),t) = u_3^*(a,t) \quad \forall (a,t) \in \mathcal{N}_1^* \times [0,T). \tag{2.54}$$

From (2.7), (2.8), (2.50), and (2.53) one gets that

$$\mathrm{Sup}\{|\bar{u}_3(x,t)|; (x,t) \in \mathcal{O}_1, |x| \leq \varepsilon\} \to 0 \quad \text{as} \quad \varepsilon \to 0. \tag{2.55}$$

Moreover, it follows from (2.52) that

$$\forall \tau \in [0,T) \quad \exists \varepsilon > 0 \text{ such that } |\bar{u}_3(x,t)| \leq |x| \quad \forall (x,t) \in \mathcal{O}_1 \cap (B_\varepsilon \times [0,\tau]). \tag{2.56}$$

Finally, using standard arguments relying on partitions of unity, we get the existence of u_3 in $C^*(\mathbb{R}^n \times [0,T];\mathbb{R}^m)$ such that

$$u_3 = 0 \text{ on a neighborhood of } (\mathbb{R}^n \backslash \{0\}) \times \{T\} \text{ in } (\mathbb{R}^n \backslash \{0\}) \times [0,T], \tag{2.57}$$

$$u_3 = \bar{u}_3 \quad \text{on} \quad \mathcal{O}_2, \tag{2.58}$$

$$\forall \tau \in [0,T) \quad \exists \varepsilon > 0 \text{ such that } |\bar{u}_3(x,t)| \leq |x| \quad \forall (x,t) \in B_\varepsilon \times [0,\tau]. \tag{2.59}$$

Then (2.57) implies (2.49); (2.58) implies (2.9) with $\mathcal{N} = \mathcal{N}_2^*$; (1.2) and (2.59) imply (2.10).

Now, only Lemma 2.11 needs to be proven. Note that in order to prove Theorem 2.7 only the case where Γ is diffeomorphic to $\mathbb{R}$ is useful (see Step 4). So, for simplicity, we study only this case, but note that our proof can be easily adapted to treat the general case. So there is a proper embedding $\gamma : \mathbb{R} \to \mathbb{R}^n \backslash \{0\}$ such that (recall that Γ is bounded) $\gamma(\mathbb{R}) = \Gamma$ and $\gamma(s) \to 0$ as $s \to 0$. Let $\pi \in C^\infty(\mathbb{R}^n \times [0,T);\mathbb{R}^n)$ be such that

$$y \in \mathbb{R}^n \to \pi(y,t) \text{ is an embedding into } \mathbb{R}^n \backslash \{0\} \quad \forall t \in [0,T), \tag{2.60}$$

$$\mathrm{Sup}\{|\pi((z,s),t)|; z \in \mathbb{R}^{n-1}, t \in [0,T)\} \to 0 \quad \text{as} \quad |s| \to +\infty, \tag{2.61}$$

$$\mathrm{Sup}\{|\pi(y,t)|; y \in \mathbb{R}^n\} \to 0 \quad \text{as} \quad t \to T, \tag{2.62}$$

$$\pi((0,s),t) = x(\gamma(s),t;u_2) \quad \forall (s,t) \in \mathbb{R} \times [0,T), \tag{2.63}$$

where u_2 satisfies the properties required in Proposition 2.7. The existence of π can be proved by slightly modifying the proof of the tubular neighborhood theorem given in [GG, Chap. II, §7]. The control system $\dot{x} = f(x,u)$ gives, by pullback using π, a time-varying control system on $\mathbb{R}^n$ $\dot{y} = \bar{F}(y,t,u)$, where the state is $y \in \mathbb{R}^n$, the control $u \in \mathbb{R}^m$, and $\bar{F} \in C^\infty(\mathbb{R}^n \times [0,T) \times \mathbb{R}^m, \mathbb{R}^n)$. Let $F : \mathbb{R}^n \times [0,T) \times \mathbb{R}^m \to \mathbb{R}^n$ be defined by $F(y,t,u) = \bar{F}(y,t,u_2(\gamma(s),t)+u)$ with $y = (z,s) \in \mathbb{R}^{n-1} \times \mathbb{R} \simeq \mathbb{R}^n$. Then $F \in C^\infty(\mathbb{R}^n \times [0,T) \times \mathbb{R}^m; \mathbb{R}^n)$ and, by (2.63), $F((0,s),t,0) = 0\ \forall (s,t)$ in $\mathbb{R} \times [0,T)$. Let $A(s,t) = \partial F/\partial y((0,s),t,0)$, $B(s,t) = \partial F/\partial u((0,s),t,0)$. Since u_2 satisfies (2.3)—with u_1 instead of u_2—we have

(2.64)
$$\operatorname{Span}\left\{ \left(\frac{\partial}{\partial t} - A(s,t)\right)^i B(s,t)\Big|_{t=\bar{t}} w; w \in \mathbb{R}^m, i \geq 0 \right\} = \mathbb{R}^n \quad \forall (s,\bar{t}) \in \mathbb{R} \times [0,T).$$

Then Lemma 2.11 follows from the following lemma, which is proved in Appendix C.

LEMMA 2.12. *Let F in $C^\infty(\mathbb{R}^n \times [0,T) \times \mathbb{R}^m; \mathbb{R}^m)$, vanishing on $(\{0\} \times \mathbb{R}) \times \mathbb{R} \times \{0\}$, be such that* (2.64) *holds. Then, for any ε in $C^0(\mathbb{R} \times [0,T); (0,+\infty))$ there exist an open neighborhood $\mathcal{N}$ of $\{0\} \times \mathbb{R} \subset \mathbb{R}^{n-1} \times \mathbb{R} \simeq \mathbb{R}^n$ and u in $C^\infty(\mathbb{R}^n \times [0,T); \mathbb{R}^m)$ such that*

$$(2.65) \qquad |u(z,s,t)| \leq \varepsilon(s,t) \quad \forall (z,s,t) \in \mathbb{R}^{n-1} \times \mathbb{R} \times [0,T),$$

$$(2.66) \qquad a \in \mathcal{N} \to y(a,t;u) \textit{ is an embedding of } \mathcal{N} \textit{ into } \mathbb{R}^n \quad \forall t \in [0,T),$$

where $y(a,t;u)$ is defined by $\partial y/\partial t = F(y,t,u(a,t))$, $y(a,0;u) = a$.

Remark 2.13. (a) Proposition 2.10 again holds regardless of the dimension of Γ. (b) It follows from its proof that Proposition 2.10 remains true if, instead of (2.5), one assumes that, for a sequence $(t_i; i \in \mathbb{N})$ such that $t_i \in [0,T)$ $\forall i$ and $t_i \to T$ as $i \to +\infty$, we have $(x(a,\cdot;u_2),u_2(a,\cdot))$ is supple on $\{t_i; i \in \mathbb{N}\}$ for all a in Γ.

Proof of Step 4. In this step we end the proof of Theorem 1.4. Let us assume for the moment that we have the following proposition, where (1.8), (1.9), and $n \geq 4$ are not assumed to be true (but where (1.2) is still assumed to be true).

PROPOSITION 2.14. *Assume*

$$(2.67) \qquad \exists \alpha \in \mathbb{N}^m \textit{ such that } \frac{\partial^{|\alpha|} f}{\partial u^\alpha}(0,0) \neq 0.$$

Let ε_1 be in $(0,+\infty)$. Then there exists a proper embedding $\gamma : \mathbb{R} \to \mathbb{R}^n \backslash \{0\}$ such that $\Gamma := \gamma(\mathbb{R})$ is included in B'_{ε_1}, and for any u_0 in $C^*(\mathbb{R}^n \times [0,T]; \mathbb{R}^m)$ and any neighborhood $\mathcal{N}$ of Γ in $\mathbb{R}^n \backslash \{0\}$ there exists u_4 in $C^*(\mathbb{R}^n \times [0,T]; \mathbb{R}^m)$ and ε_4 in $(0,+\infty)$ such that (2.11) *and* (2.12) *hold and*

$$(2.68) \qquad \partial^\alpha u_4 = 0 \textit{ on } (\mathbb{R}^n \backslash \{0\}) \times \{0\} \quad \forall \alpha \in \mathbb{N}^{n+1},$$

$$(2.69) \qquad \partial^\alpha u_4 = \partial^\alpha u_0 \textit{ on } (\mathbb{R}^n \backslash \{0\}) \times \{T\} \quad \forall \alpha \in \mathbb{N}^{n+1}.$$

Now let us assume again that (1.8), (1.9), and $n \geq 4$ hold. Let ε_1 be as in Proposition 2.2. Note that (1.9) implies (2.67), hence we may apply Proposition 2.14. Let $\Gamma = \gamma(\mathbb{R})$ be as in this proposition. Then let u_3 and $\mathcal{N}$ be as in Proposition 2.10. Let u_0 in C^* be such that

$$(2.70) \qquad \partial^\alpha u_0(x,T) = \partial^\alpha u_3(x,0) \quad \forall (x,\alpha) \in (\mathbb{R}^n \backslash \{0\}) \times \mathbb{N}^{n+1}.$$

Finally, let u_4 and ε_4 be as in Proposition 2.14 and $u : \mathbb{R}^n \times \mathbb{R} \to \mathbb{R}^m$ be 2T-periodic with respect to time, equal to u_4 on $\mathbb{R}^n \times [0, T]$, and such that

$$u(x,t) = u_3(x, t-T) \quad \forall (x,t) \in \mathbb{R}^n \times (T, 2T). \tag{2.71}$$

Then (see, in particular, (2.49), (2.68), (2.69), (2.70), and (2.71)) u is continuous on $\mathbb{R}^n \times \mathbb{R}$, of class C^∞ on $(\mathbb{R}^n \backslash \{0\}) \times \mathbb{R}$, and vanishes on $\{0\} \times \mathbb{R}$. Moreover, using (2.9), (2.10), (2.11), and (2.71), we have (2.13). Finally, note that, by (2.10), (2.12), and (2.71), we have (2.14). Since T is an arbitrary positive real number, Theorem 1.4 follows from the following lemma.

LEMMA 2.15. *Let X in $C^0(\mathbb{R}^n \times \mathbb{R}; \mathbb{R}^n)$, vanishing on $\{0\} \times \mathbb{R}$ and T-periodic in time, be such that, for some positive real number ε_0,*

$$((\dot{x} = X(x,t) \text{ and } x(\tau) = 0) \Rightarrow (x(t) = 0 \quad \forall t \geq \tau)) \quad \forall \tau \in \mathbb{R}, \tag{2.72}$$

$$(\dot{x} = X(x,t) \text{ and } |x(0)| < \varepsilon_0) \Rightarrow (x(T) = 0). \tag{2.73}$$

Then 0 is a uniformly locally asymptotically stable point for $\dot{x} = X(x,t)$ and, for some positive real number ε_1,

$$((\dot{x} = X(x,t) \text{ and } |x(\tau)| < \varepsilon_1) \Rightarrow (x(\tau + 2T) = 0)) \quad \forall \tau \in \mathbb{R}. \tag{2.74}$$

Proof. One first notices that, for any positive real number ε, there exists $\delta(\varepsilon)$ in $(0, \min(\varepsilon, \varepsilon_0))$ such that for any $0 \leq s \leq \tau \leq T$ and any maximal solution of $\dot{x} = X(x,t)$, $|x(s)| < \delta(\varepsilon)$ implies that $|x(\tau)| < \varepsilon$. Indeed , if this is not the case, there exist a positive real number ε_2, two sequences of real numbers $(s_n; n \in \mathbb{N})$, $(\tau_n; n \in \mathbb{N})$ with $0 \leq s_n \leq \tau_n \leq T$, and a sequence $(x_n; n \in \mathbb{N})$ of solutions of $\dot{x}_n = X(x_n, t)$ such that $|x(s_n)| \leq 1/n$, $|x(t)| \leq \varepsilon_2$ for all t in $[s_n, \tau_n]$, and $|x(\tau_n)| = \varepsilon_2$. Letting n go to ∞ we get, by the Ascoli theorem, the existence of two real numbers s and τ with $0 \leq s \leq \tau \leq T$, and a solution of $\dot{x} = X(x,t)$ such that $x(s) = 0$ and $|x(\tau)| = \varepsilon_2$, which contradicts (2.73). Then one easily sees that (2.74) holds with $\varepsilon_1 = \delta(\varepsilon_0)$ and that, for any positive real number ε and any real number s, $\dot{x} = X(x,t)$ and $|x(s)| < \delta(\delta(\varepsilon))$ implies that $|x(\tau)| < \varepsilon)$ for all $\tau \geq s$. □

We now prove Proposition 2.14. It follows from (2.67) that there exist c in $\mathbb{R}^m \backslash \{0\}$ and an i in $\mathbb{N}$ such that, with $\bar{f}(x,v) = f(x, vc)$ for $(x,v) \in \mathbb{R}^n \times \mathbb{R}$, $\partial^i \bar{f}/\partial v^i(0,0) \neq 0$. Hence, replacing, if necessary, $\dot{x} = f(x,u)$ by $\dot{x} = \bar{f}(x,v)$, we may assume without loss of generality that $m = 1$. Let p be the integer defined by $\partial^i f/\partial u^i(0,0) = 0$ for all i in $[0, p-1]$, $\partial^p f/\partial u^p(0,0) \neq 0$. Clearly, without loss of generality, we may assume $\partial^p f/\partial u^p(0,0) = e_n$. Let us define f_1 and f_2 by $f(x,u) = (f_1(x,u), f_2(x,u)) \in \mathbb{R}^{n-1} \times \mathbb{R} \simeq \mathbb{R}^n$. Hence, by the Malgrange preparation theorem (see, e.g., [GG, Chap. IV, §2]), there exist an open neighborhood V of (0,0) in $\mathbb{R}^n \times \mathbb{R}$, p maps $a_0, a_1, \ldots, a_{p-1}$ in $C^\infty(V; \mathbb{R})$, and a map θ in $C^\infty(V; \mathbb{R})$ such that

$$f_2(x,u) = \theta(x,u) \left(u^p + \sum_{i=0}^{p-1} a_i(x,u) u^i \right) \quad \forall (x,u) \in V, \tag{2.75}$$

$$\theta(0,0) = 1, \tag{2.76}$$

$$a_i(0,u) = 0 \quad \forall i \in [0,p-1] \;\; \forall u \in \mathbb{R} \;\text{ with }\; (0,u) \in V. \tag{2.77}$$

Let $u_5 : \mathbb{R}^n \longrightarrow \mathbb{R}$ be defined by $u_5(x) = |x|^{1/(p+1)}$. Then, from (2.75), (2.76), and (2.77), we get the existence of ε_5 in $(0, \varepsilon_1/2)$ such that

$$f_2(x, u(x)) \geq |x|^{p/(p+1)}/2 \quad \forall x \in B_{\varepsilon_5}. \tag{2.78}$$

Let $F : B_{\varepsilon_5} \longrightarrow \mathbb{R}^{n-1}$ be defined by $F(x) = f_1(x, u(x))/f_2(x, u(x))$ for all x in B'_{ε_5}, $F(0) = 0$. Then F is continuous, and straightforward computations show that, for some $C_5 > 0$,

$$|F(y_1, z) - F(y_2, z)| \leq C_5 |y_1 - y_2| / |z|^{p/(p+1)} \tag{2.79}$$

for all (y_1, z) in $\mathbb{R}^{n-1} \times \mathbb{R} \simeq \mathbb{R}^n$ and all (y_2, z) in $\mathbb{R}^{n-1} \times \mathbb{R} \simeq \mathbb{R}^n$, with $(y_1, z) \in B_{\varepsilon_5}$ and $(y_2, z) \in B_{\varepsilon_5}$. Note that

$$\int_{-1}^{1} dz/|z|^{p/p+1} < +\infty. \tag{2.80}$$

From (2.79) and (2.80) it follows (see, e.g., [F, Chap. 1, §1, Thms. 1 and 3]) that the Cauchy problem $dy/dz = F(y, z)$, $y(z_0) = y_0$, (y, z) in $\mathbb{R}^{n-1} \times \mathbb{R} \simeq \mathbb{R}^n$, $(y, z) \in B_{\varepsilon_5}$, has locally one and only one solution if (y_0, z_0) is in B_{ε_5}. Let ε_6 be a positive real number such that $dy/dz = F(y, z)$, $y(0) = 0$, $z \in \{-\varepsilon_6, \varepsilon_6\}$, $(y, z) \in B_{\varepsilon_5}$ has one and only one solution. Let $\bar{y}$ be this solution. We choose ε_6 small enough so that there exists a proper embedding γ of $\mathbb{R}$ into $\mathbb{R}^n \backslash \{0\}$ such that $\Gamma := \gamma(\mathbb{R})$ is included in B'_{ε_1} and $\{(\bar{y}(z), z); z \in (0, \varepsilon_6]\}$ is included in Γ. Since Proposition 2.14 is a local result we may assume, without loss of generality, that for some constant M in $(0, +\infty)$

$$|f(x, u)| \leq M \quad \forall (x, u) \in \mathbb{R}^n \times \mathbb{R}^m. \tag{2.81}$$

Note that by (2.78) we have , for z in $\{-\varepsilon_6, \varepsilon_6\}$, $f_2((\bar{y}(z), z), u_5(\bar{y}(z), z)) \geq |z|^{p/p+1}/2$; therefore, the Cauchy problem $dz/dt = f_2((\bar{y}(z), z), u_5(\bar{y}(z), z))$, $z(0) = 0$ has one and only one solution on $\{-\varepsilon_6/M, \varepsilon_6/M\}$ such that $tz(t) > 0$ for all t in $\{-\varepsilon_6/M, \varepsilon_6/M\} \backslash \{0\}$. Hence there exists one and only one ϕ in $C^0(\mathbb{R}^n \times \mathbb{R}; \mathbb{R}^n)$ such that $\partial \phi / \partial t = f(\phi, u_5(\phi))$, $\phi(x, 0) = x$, and, for all (x, τ) in $\mathbb{R}^n \times \mathbb{R}$,

$$(\phi(x, \tau) = 0) \Rightarrow (\phi_n(x, t) \cdot (t - \tau) > 0 \;\text{ if }\; 0 < |t - \tau| \leq \varepsilon_6/M), \tag{2.82}$$

where ϕ_n denotes the last component of ϕ.

Now, let $\mathcal{N}$ be an open neighborhood of Γ in $\mathbb{R}^n \backslash \{0\}$. Let $(d_i; i \in \mathbb{N} \cup \{\infty\})$ be a sequence of maps in $C^1((0, \varepsilon_6]; (0, +\infty))$ such that

$$d_i(z) < d_{i+1}(z), \; d_i(z) < d_\infty(z) \quad \forall (i, z) \in \mathbb{N} \times (0, \varepsilon_6], \tag{2.83}$$

$$\mathcal{N}_i := \{(y, z) \in \mathbb{R}^{n-1} \times (0, \varepsilon_6]; |y - \bar{y}(z)|^2 \leq d_i(z)\} \subset \mathcal{N} \quad \forall i \in \mathbb{N} \cup \{\infty\}, \tag{2.84}$$

$$d'_i(z) > 2C_5 d_i(z)/z^{p/(p+1)} \quad \forall (i, z) \in (\mathbb{N} \cup \{\infty\}) \times (0, \varepsilon_6]. \tag{2.85}$$

The existence of such a sequence can be easily established by letting $d_i(z) = \bar{d}_i(z) \exp\int_0^z 2C_5/\tau^{p/(p+1)}d\tau$ ($\bar{d}_i'(z) > 0$ for all (i,z) in $(\mathbb{N} \cup \{\infty\}) \times (0,\varepsilon_6]$ is equivalent to (2.85)). Let $T_0 = T - (\varepsilon_6/2M)$, $T_i = T - (T - T_0)/2^i$ for all i in $\mathbb{N}\backslash\{0\}$, and $r_i = (\varepsilon_6/2) + (T_i - T_0)/M$ for all i in $\mathbb{N}$. Then we have, with the convention $1/0 = +\infty$, the following lemma.

LEMMA 2.16. *Let i be a nonnegative integer. Then there exists a positive real number δ_i in $(0, 1/i)$ such that for any ε in $(0, +\infty)$ there exists $\tilde{u}$ in $C^\infty(\mathbb{R}^n \times [T_i, T_{i+1}]; \mathbb{R}^m)$ such that*

$$(2.86) \qquad |\tilde{u}(x,t)| \le |x|^{1/p+1} \quad \forall(x,t) \in \mathbb{R}^n \times [T_i, T_{i+1}],$$

$$(2.87) \qquad \partial^\alpha \tilde{u} = 0 \ \text{ on } \ \mathbb{R}^n \times \{T_i, T_{i+1}\} \quad \forall\alpha \in \mathbb{N}^{n+1},$$

$$(2.88) \ (\dot{x} = f(x, \tilde{u}(x,t)), x(T_i) \in (\mathcal{N}_i \cup B_{\delta_i}) \cap B_{r_i}) \Rightarrow (x(T_{i+1}) \in (\mathcal{N}_{i+1} \cup B_\varepsilon) \cap B_{r_{i+1}}).$$

Proof. Using (2.79), (2.81), and (2.85), one has

$$(2.89) \qquad (x \in \mathcal{N}_i \cap B_{r_i} \text{ and } t \in [0, T_{i+1} - T_i]) \Rightarrow (\phi(x,t) \in \mathcal{N}_i \cap B_{r_{i+1}}).$$

Note also that $\phi(0, T_{i+1} - T_i) \in \mathcal{N}_i$ and, therefore, for some δ_i small enough but positive,

$$(2.90) \qquad \phi(B_{\delta_i}, T_{i+1} - T_i) \subset \mathcal{N}_i.$$

Moreover, $\phi(0, [0, T_{i+1} - T_i]) \subset \mathcal{N}_i$, and, therefore, $\varepsilon > 0$ being given, there exists η in $(0, \varepsilon/2)$—depending on i—such that

$$(2.91) \qquad \{\phi(x,s); x \in B_\eta, s \in [0, T_{i+1} - T_i]\} \subset \mathcal{N}_i \cup B_{\varepsilon/2}.$$

Let $u^* \in C^\infty(\mathbb{R}^n; \mathbb{R}^m)$ be equal to 0 on $B_{\eta/2}$ and such that $u^*(x) = u_5(x)$ for all x in $\mathbb{R}^n\backslash B_\eta$, $|u^*(x)| \le |x|^{1/(p+1)}$ for all x in $\mathbb{R}^n$; then, using (2.81), (2.90), and (2.91), we have

$$(2.92) \ (\dot{x} = f(x, u^*(x)), x(T_i) \in (\mathcal{N}_i \cup B_{\delta_i}) \cap B_{r_i}) \Rightarrow \ x(T_{i+1}) \in (\mathcal{N}_i \cup B_{\varepsilon/2}) \cap B_{r_{i+1}}.$$

Finally, let $\tilde{u} : \mathbb{R}^n \times [T_i, T_{i+1}] \to \mathbb{R}^m$ be defined by $\tilde{u}(x,t) = \theta(t)u^*(x)$, where $\theta \in C^\infty([T_i, T_{i+1}]; [0,1])$ is equal to 1 on $[\mu, T_i + T_{i+1} - \mu]$ for some μ in $(0, +\infty)$ and vanishes on a neighborhood of $\{T_i, T_{i+1}\}$. Then $u \in C^\infty(\mathbb{R}^n \times [T_i, T_{i+1}]; \mathbb{R}^m)$ satisfies (2.87) and (2.88). Moreover, if μ is small enough, we have (2.88) from (2.92). □

Let us now end the proof of Proposition 2.14. In Lemma 2.16 we take $\varepsilon = \delta_{i+1}$; we get a map $\tilde{u}$ that we denote $\tilde{u}_i$. Let $\bar{u}_4$ be the map from $\mathbb{R}^n \times [0, T)$ into $\mathbb{R}^m$ that is equal to $\tilde{u}_i$ on $\mathbb{R}^n \times [T_i, T_{i+1}]$ for all nonnegative integers i and vanishes on $\mathbb{R}^n \times [0, T_0]$. This map is of class C^∞ and satisfies

$$(2.93) \qquad |\bar{u}_4(x,t)| \le |x|^{1/(p+1)} \quad \forall(x,t) \in \mathbb{R}^n \times [0, T],$$

$$(2.94) \qquad \partial^\alpha \bar{u}_4 = 0 \text{ on } \mathbb{R}^n \times \{0\} \quad \forall\alpha \in \mathbb{N}^{n+1}.$$

Let $\varepsilon_4 > 0$ be such that

$$(2.95) \qquad (\dot{x} = f(x,0), x(0) \in B_{\varepsilon_4}) \Rightarrow (x(T_0) \in B_{\delta_0} \cap B_{r_0}).$$

Then, using (2.88) and (2.95), we have

$$(2.96) \qquad (\dot{x} = f(x, \bar{u}_4(x,t)), x(0) \in B_{\varepsilon_4}) \Rightarrow (x(T) \in \mathcal{N}_\infty \cup \{0\}).$$

Finally, let $\bar{\theta} \in C^\infty((\mathbb{R}^n \times [0,T])\backslash\{(0,T)\}; [0,1])$ be such that $\bar{\theta}(x,t) = 1$ if $t \leq T - \bar{\mu}(x)$, $\bar{\theta}(x,t) = 0$ if $t \geq T - (\bar{\mu}(x)/2)$, where $\bar{\mu} \in C^0(\mathbb{R}^n\backslash\{0\}; (0,T))$, and let $u_4(x,t) = \bar{\theta}(x,t)\bar{u}_4(x,t) + (1 - \theta(x,t))u_0(x,t)$. Then u_4 is in $C^*(\mathbb{R}^n \times [0,T]; \mathbb{R}^m)$ and satisfies (2.68) and (2.69). Moreover, if $\bar{\mu}(x) \to 0$ as $x \to 0$, (2.12) holds and, if $\bar{\mu}$ is in a small enough neighborhood of 0 in $C^0(\mathbb{R}^n\backslash\{0\}; \mathbb{R})$ for the Whitney C^0-topology (see, e.g., [GG, Chap. II, §3] for a definition), (2.11) follows from (2.96). This ends the proof of Proposition 2.14 and also the proof of Theorem 1.4.

3. Particular control systems.

In this section we show that, for some special types of control systems, the assumption $n \geq 4$ can be removed and the proof can be simplified. Let us first remark that the assumption $n \geq 4$ is used only in Step 2 of §2. Let us also note that, in Steps 2 and 3 of §2, Γ is assumed to be without boundary, but straightforward modifications of the proofs given in these steps show that Propositions 2.7 and 2.10 remain valid if Γ has a boundary, except that for Proposition 2.10 $\mathcal{N}$ is now a neighborhood of $\Gamma\backslash\partial\Gamma$, where $\partial\Gamma$ is the boundary of Γ.

3.1. Systems without drift.

In this section we prove Proposition 1.6. Hence we now have $f(x,u) = \sum_{i=1}^m u_i f_i(x)$. By (1.9) we may assume, without loss of generality, $f_m(x) = e_n$ for all x in $\mathbb{R}^n$ and $f_i(x) \in \mathbb{R}^{n-1} \times \{0\}$ for all (x,i) in $\mathbb{R}^n \times [1, m-1]$. Therefore, our system can be written $\dot{y} = \bar{g}(y,z,v)$, $\dot{z} = w$, where the state is $x = (y,z) \in \mathbb{R}^{n-1} \times \mathbb{R}$ and the control $u = (v,w) \in \mathbb{R}^{m-1} \times \mathbb{R}$. Note that $\bar{g}(y,z,-v) = -\bar{g}(y,z,v)$ for all (y,z,v) in $\mathbb{R}^{n-1} \times \mathbb{R} \times \mathbb{R}^{m-1}$. So Proposition 1.6 is a corollary of the following proposition.

PROPOSITION 3.1. *Let $g \in C^\infty(\mathbb{R}^{n-1} \times \mathbb{R} \times \mathbb{R}^{m-1} \times \mathbb{R}; \mathbb{R}^{n-1})$ and let $\varphi \in C^\infty(\mathbb{R}^{m-1}; \mathbb{R}^{m-1})$ be such that*

$$(3.1) \qquad g(y,z,\varphi(v),-w) = -g(y,z,v,w) \quad \forall (y,z,v,w) \in \mathbb{R}^{n-1} \times \mathbb{R} \times \mathbb{R}^{m-1} \times \mathbb{R},$$

$$(3.2) \qquad g(0,z,0,w) = 0 \quad \forall (z,w) \in \mathbb{R} \times \mathbb{R},$$

$$(3.3) \qquad \varphi(0) = 0.$$

Let Σ' be the control system $\Sigma' : \dot{y} = g(y,z,v,w)$, $\dot{z} = w$, where the state is $x = (y,z) \in \mathbb{R}^{n-1} \times \mathbb{R} \simeq \mathbb{R}^n$ and the control $u = (v,w) \in \mathbb{R}^{m-1} \times \mathbb{R} \simeq \mathbb{R}^m$. Assume that Σ' satisfies the strong jet accessibility rank condition at $(0,0,0,0)$. Then Σ' is locally smoothly stabilizable in small time by means of periodic time-varying feedback law.

Proof. Indeed, let u_0 in $C^\infty(\mathbb{R}^n \times \mathbb{R}; \mathbb{R}^m)$ be such that, with the notations of Step 1 in §2,

$$(3.4) \qquad \partial^\alpha u_0 = 0 \text{ on } (\mathbb{R}^n \times \{0, T/4\}) \cup (\{0\} \times \mathbb{R}^n) \quad \forall \alpha \in \mathbb{N}^{n+1},$$

$$(3.5) \qquad u_0 \circ h \text{ saturates } X \text{ on } \mathbb{R}^n \times (\{0\} \times \mathbb{R}^n) \times (0, T/4)$$

with $X(x,a,s,u) = ((g(y,z,v,w),w),0,1) \in \mathbb{R}^n \times \mathbb{R}^n \times \mathbb{R}$. The existence of u_0 follows again from [Co3, Thm. 1.3]. Let $\bar{u}_0 : \mathbb{R} \times [0,T] \to \mathbb{R}^m$ be defined by (with $\bar{u}_0 = (\bar{v}_0, \bar{w}_0) \in \mathbb{R}^{m-1} \times \mathbb{R}, u_0 = (v_0, w_0) \in \mathbb{R}^{m-1} \times \mathbb{R}$)

$$\bar{u}_0(s,t) = u_0(se_n,t) \quad \forall (s,t) \in \mathbb{R} \times [0,T/4], \tag{3.6}$$

$$\bar{w}_0(s,t) = -w_0(se_n,(T/2)-t) \quad \forall (s,t) \in \mathbb{R} \times [T/4,T/2], \tag{3.7}$$

$$\bar{v}_0(s,t) = (\varphi \circ v_0)(se_n,(T/2)-t) \quad \forall (s,t) \in \mathbb{R} \times [T/4,T/2], \tag{3.8}$$

$$\bar{v}_0(s,t) = 0 \quad \forall (s,t) \in \mathbb{R} \times [T/2,T], \tag{3.9}$$

$$\bar{w}_0(s,t) = \frac{\partial}{\partial t}\left(\frac{s^3}{s^2+\beta(t)}\right) = \frac{-\beta'(t)s^3}{(s^2+\beta(t))^2} \quad \forall (s,t) \in \mathbb{R} \times (T/2,T), \tag{3.10}$$

$$\bar{w}_0(s,T) = 0 \quad \forall s \in \mathbb{R}, \tag{3.11}$$

with

$$\beta(t) = \exp\left(\frac{1}{T-t} - \frac{1}{t-(T/2)}\right) \quad \forall t \in (T/2,T). \tag{3.12}$$

For $\bar{\varepsilon}_0 > 0$ let $\bar{x}_0 : [0,\bar{\varepsilon}_0] \times [0,T] \to \mathbb{R}^n$ be defined by $\partial\bar{x}_0/\partial t(s,t) = f(\bar{x}_0(s,t),\ \bar{u}_0(s,t))$, $\bar{x}_0(s,0) = se_n$. This map is defined on all $[0,\bar{\varepsilon}_0] \times [0,T]$ if $\bar{\varepsilon}_0$ is small enough. Note that by (3.1), (3.7), and (3.8) $\bar{x}_0(s,T/2-t) = \bar{x}_0(s,t)$ for all (s,t) in $[0,\bar{\varepsilon}_0] \times [T/4,T/2]$. In particular, we have $\bar{x}_0(s,T/2) = se_n$ for all s in $[0,\bar{\varepsilon}_0]$ and, therefore, by (3.10), (3.11), and (3.12), $\bar{x}_0(s,t) = s^3e_n/(s^2+\beta(t))$ for all (s,t) in $[0,\bar{\varepsilon}_0] \times (T/2,T)$ and $\bar{x}_0(s,T) = 0$ for all s in $[0,\bar{\varepsilon}_0]$, which, in particular, implies that, with $\bar{x}_0 = (\bar{y}_0, \bar{z}_0) \in \mathbb{R}^{n-1} \times \mathbb{R}$, $\partial\bar{z}_0/\partial s > 0$ on $[0,\bar{\varepsilon}_0] \times [T/2,T)$. Let us remark that, if u_0 is in a small enough neighborhood of 0 in $C^\infty(\mathbb{R}^n \times \mathbb{R}; \mathbb{R}^m)$, we have $\partial\bar{z}_0/\partial s > 0$ on $[0,\bar{\varepsilon}_0] \times [0,T/2]$. Recalling that Σ satisfies the strong jet accessibility rank condition at (0, 0) we get from (3.4), (3.5), and (3.6) that $(\bar{x}_0(s,\cdot),\bar{u}_0(s,\cdot))$ is supple on $(0,T/4)$ for all s in $(0,\bar{\varepsilon}_0]$ if $\bar{\varepsilon}_0$ is still small enough. Then, using Appendix A—with $T_1 = 0$, $T_2 = T/32$, $T_3 = T/16$, $T_4 = T/8$, and $\Lambda = (0,\bar{\varepsilon}_0)$—and [Co3, Thm. 1.3] in a similar way as in Step 1 of §2, we get the existence of u_1 in $C^0(\mathbb{R} \times [0,T]; \mathbb{R}^m)$ of class C^∞ on $(\mathbb{R}\backslash\{0\}) \times [0,T]$, vanishing on $\{0\} \times [0,T]$ and on $\mathbb{R} \times \{T\}$, equal to $\bar{u}_0$ on $(\mathbb{R} \times [T/8,T])\backslash((0,\bar{\varepsilon}_0) \times (T/8,T))$, and such that

$$u_1/s^2 \in L^\infty(([-1,1]\backslash\{0\}) \times [0,\tau]) \quad \forall \tau \in [0,T), \tag{3.13}$$

$$x_1(s,0) = se_n, \tag{3.14}$$

$$\frac{\partial y_1}{\partial s} > 0 \quad \text{on} \quad (0,\bar{\varepsilon}_0] \times [0,T), \tag{3.15}$$

$$(x_1(s,\cdot),u_1(s,\cdot)) \text{ is supple on } (0,T) \quad \forall s \in (0,\varepsilon_1), \tag{3.16}$$

for some ε_1 in $(0,\bar{\varepsilon}_0)$, and where $x_1 = (y_1,z_1) : (0,\varepsilon_0] \times [T/8,T] \to \mathbb{R}^n$ is defined by $\partial x_1/\partial t(s,t) = f(x_1(s,t),u_1(s,t))$, $x_1(s,T) = 0$. Note that by (3.15) $se_n \in \Gamma = \{se_n; s \in (0,\varepsilon_1]\} \to x_1(s,t) \in \mathbb{R}^n \backslash \{0\}$ is an embedding for all t in $[0,T)$. Clearly, there exists $u_2 \in C^*(\mathbb{R}^n \times [0,T];\mathbb{R}^m)$ satisfying (2.15) with u_2 instead of u_1, vanishing on $\mathbb{R}^n \times \{T\}$, and such that $u_2(se_n,t) = u_1(s,t)$ for all (s,t) in $(0,\varepsilon_1] \times [0,T]$. Then, (2.4) holds with $a \in \Gamma$ instead of $|a| \leq \varepsilon_1$ and (2.5) holds with $(0,T)$ instead of $[0,T]$ and with $a \in \Gamma$ instead of $0 < |a| \leq \varepsilon_1$. Note also that, by (3.15), (2.6) holds. All these properties are sufficient to perform Step 3 in §2 if, with the notations of Proposition 2.10, $\mathcal{N}$ is now a neighborhood of $\{se_n; s \in (0,\varepsilon_1)\}$ in $\mathbb{R}^n \backslash \{0\}$. Moreover, using (3.2), one easily sees that the proof of Step 4 in §2 holds for this Γ. In fact, with (3.2), this proof can be simplified slightly. Indeed, let $\gamma \in C^0((0,+\infty);(0,+\infty))$ and let $\tilde{u} \in C^0(\mathbb{R}^n;\{0\} \times [0,+\infty)) \cap C^\infty(\mathbb{R}^n \backslash \{0\};\mathbb{R}^m)$ be such that $\tilde{u}(x) = (0,|x|^{1/p})$ if $|y|^2 \geq \gamma(z)$, $\tilde{u}(0,z) = 0$ if $z \geq 0$. Then one deduces from (3.2) that if $\dot{x} = f(x,\tilde{u}(x))$, $x(0) = 0$, then $x(t) = 0$ for all positive t. Moreover, one easily checks that, if γ is in a small enough neighborhood of 0 in $C^0((0,+\infty);\mathbb{R})$ for the Whitney C^0-topology, there exists $\tilde{\varepsilon}$ in $(0,+\infty), T_0$ in $(0,T)$ and $\mathcal{N}' \subset \mathbb{R}^n \backslash \{0\}$, whose closure is included $\mathcal{N} \cup \{0\}$, such that

$$(\dot{x} = f(x,\tilde{u}(x)), |x(T_0)| \leq \tilde{\varepsilon}) \Rightarrow x(T) \in \mathcal{N}'. \tag{3.17}$$

Then it suffices for one to take u_4 in $C^*(\mathbb{R}^n \times [0,T];\mathbb{R}^m)$ satisfying (2.69), vanishing on $\mathbb{R}^n \times [0,T_0]$, and such that

$$|u_4(x,t)| \leq |u_0(x,t)| + |\tilde{u}(x)| \quad \forall (x,t) \in \mathbb{R}^n \times [0,T], \tag{3.18}$$

$$u_4(x,t) = \tilde{u}(x) \quad \forall (x,t) \in (\mathbb{R}^n \times [T_0,T]) \backslash V, \tag{3.19}$$

where V is a small neighborhood of $(\mathbb{R}^n \backslash \{0\}) \times \{T_0,T\}$ in $(\mathbb{R}^n \backslash \{0\}) \times \mathbb{R}$. □

3.2. Dynamical time-varying stabilization. In this section we prove Proposition 1.6. Let us first note that this proposition is equivalent to the following proposition.

PROPOSITION 1.6*. *Assume that* (1.8) *and* (1.9) *hold for* Σ. *Then the control system*

$$\dot{x} = f(x,u), \quad \dot{y} = v \in \mathbb{R}, \tag{3.20}$$

where the state is $(x,y) \in \mathbb{R}^n \times \mathbb{R}$ *and the control* $(u,v) \in \mathbb{R}^m \times \mathbb{R}$, *is locally smoothly stabilizable in small time by means of periodic time-varying feedback law.*

Let us remark that this last proposition can be, roughly speaking, rephrased in the following manner: If (1.8) and (1.9) hold then Σ can be locally smoothly stabilizable in small time by means of dynamical time-varying feedback law, which increases the dimension of the state space by only 1. Let us also remark that in (3.20) $y \in \mathbb{R}$. If one replaces $\dot{y} = v \in \mathbb{R}$ by $\dot{y} = v \in \mathbb{R}^n$, a shorter proof is given in [Co2]—with less regularity on the feedback law but without assuming (1.9). Let us now prove Proposition 1.6*. Let u_1 be as in Proposition 3.2. By [Co3, Thm. 1.3] there exists u_0 in $C^0(\mathbb{R}^n \times [0,2T];\mathbb{R}^m) \cap C^\infty((\mathbb{R}^n \backslash \{0\}) \times [0,2T];\mathbb{R}^m)$ such that

$$u_0(a,t) = u_1(a,t-T) \quad \forall (a,t) \in \mathbb{R}^n \times [T,2T], \tag{3.21}$$

$$u_0/|a|^2 \in L^\infty(B_1' \times [0,\tau]) \quad \forall \tau \in [0,2T), \tag{3.22}$$

$$u_0 \circ h \text{ saturates } X_0 \text{ on } \mathbb{R}^n \times (\mathbb{R}^n \backslash \{0\}) \times (0,T), \tag{3.23}$$

where in (3.23) we have used the notations of Step 1 in §2. Let, for ε_2 in $(0,+\infty)$, $u_2 : [0,\varepsilon_2] \times [0,2T] \to \mathbb{R}^m$ be defined by

$$u_2(s,t) = u_0(se_n,t) \quad \forall (s,t) \in [0,\varepsilon_2] \times [0,T], \tag{3.24}$$

$$u_2(s,t) = u_1(x_0(s,T), t-T) \quad \forall (s,t) \in [0,\varepsilon_2] \times [T,2T], \tag{3.25}$$

where $x_0 : [0,\varepsilon_2] \times [0,T] \to \mathbb{R}^n$ is defined by $\partial x_0/\partial t(s,t) = f(x_0(s,t), u_0(s,t))$, $x_0(s,0) = 0$. For ε_2 small enough u_2 is well defined. Let $x_2 : [0,\varepsilon_2] \times [0,2T] \to \mathbb{R}^n$ be defined by $\partial x_2/\partial t(s,t) = f(x_2(s,t), u_2(s,t))$, $x_2(s,0) = 0$. Note that, by (1.9) and (3.23), we have, if ε_2 is small enough,

$$(x_2(s,\cdot), u_2(s,\cdot)) \text{ is supple on } (0,T) \quad \forall s \in (0,\varepsilon_2]. \tag{3.26}$$

Let us also remark that, from (2.15), (3.22), (3.24), and (3.25), we have (still for ε_2 small enough)

$$u_2/s^2 \in L^\infty((0,\varepsilon_2] \times [0,\tau]) \quad \forall \tau \in [0,2T). \tag{3.27}$$

Proceeding again as in Step 1 in §2 one can prove that, if we perturb u_2 slightly and in a suitable way, u_3 can be constructed in $C^0([0,\varepsilon_2]\times[0,2T];\mathbb{R}^m) \cap C^\infty((0,\varepsilon_2]\times[0,2T];\mathbb{R}^m)$ such that

$$u_3/s^2 \in L^\infty((0,\varepsilon_2] \times [0,2\tau]) \quad \forall \tau \in [0,2T), \tag{3.28}$$

$$x_3(s,2T) = 0 \quad \forall s \in [0,\varepsilon_2], \tag{3.29}$$

$$(x_3(s,\cdot), u_3(s,\cdot)) \text{ is supple on } (0,2T) \quad \forall s \in (0,\varepsilon_2), \tag{3.30}$$

where $x_3 : [0,\varepsilon_2] \times [0,2T]$ is defined by $\partial x_3/\partial t(s,t) = f(x_3(s,t), u_3(s,t))$, $x_3(s,0) = 0$. Let $v_3 : [0,\varepsilon_2] \times [0,2T] \to \mathbb{R}$ be defined by $v_3 = 0$ on $[0,\varepsilon_2] \times \{0,2T\}$ and

$$v_3(s,t) = \frac{\partial}{\partial t}\left(\frac{s^3}{s^2+\beta(t)}\right) = -\beta'(t)\frac{s^3}{(s^2+\beta(t))^2} \quad \forall (s,t) \in [0,\varepsilon_2] \times (0,2T) \tag{3.31}$$

with $\beta(t) = \exp(2(t-T)/t(2T-t))$ for all t in $(0,2T)$. Let $y_3 : [0,\varepsilon_2] \times [0,2T] \to \mathbb{R}$ be defined by $\partial y_3/\partial t(s,t) = v_3(s,t)$, $y_3(s,0) = s$. By (3.31) we have $y_3(s,t) = s^3/(s^2+\beta(t))$ for all (s,t) in $[0,\varepsilon_2] \times (0,2T)$, $y_3(s,2T) = 0$ for all s in $[0,\varepsilon_2]$. Finally, one just notes that $(|u_3| + |v_3|)/s^2 \in L^\infty((0,\varepsilon_2] \times [0,\tau])$ for all $\tau \in [0,2T)$ and that, for system (3.20),

$$((x_3(s,\cdot), y_3(s,\cdot)),\ (u_3(s,\cdot), v_3(s,\cdot))) \text{ is supple on } (0,2T) \quad \forall s \in (0,\varepsilon_2), \tag{3.32}$$

(3.33) $s \in (0,\varepsilon_2) \to (x_3(s,t), y_3(s,t)) \in \mathbb{R}^n \times (\mathbb{R}\backslash\{0\})$ is an embedding $\forall t \in [0,T)$.

The conclusion then follows as in the above section (note that, with obvious changes of notations, (3.2) is satisfied).

Appendix A. Let Λ be an open subset of $\mathbb{R}^p$ (or a manifold). Let T_1, T_2, T_3, and T_4 be four real numbers such that $T_1 < T_2 < T_3 < T_4$. Let $x \in C^\infty(\Lambda \times [T_1,T_4]; \mathbb{R}^n)$ and $u \in C^\infty(\Lambda \times [T_1,T_4]; \mathbb{R}^m)$ be such that

$$(A.1) \qquad \frac{\partial x}{\partial t} = f(x(\lambda,t), u(\lambda,t)) \quad \forall(\lambda,t) \in \Lambda \times [T_1,T_4],$$

$$(A.2) \qquad (x(\lambda,\cdot), u(\lambda,\cdot)) \text{ is supple on } (T_2,T_3) \quad \forall \lambda \in \Lambda.$$

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

LEMMA A.1. *There exists δ in $C^0(\Lambda; (0,+\infty))$ and a map H in $C^\infty(\omega \times [T_1,T_4]; \mathbb{R}^m)$ with $\omega = \{(\lambda,b) \in \Lambda \times \mathbb{R}^n; |b| < \delta(s)\}$ such that*

$$(A.3) \qquad \textit{Support}\,(H(\lambda,b,\cdot) - u(\lambda,\cdot)) \subset (T_2,T_3) \quad \forall(\lambda,b) \in \omega,$$

$$(A.4) \qquad (\dot{y} = f(y, H(\lambda,b,t)),\ y(T_1) = x(\lambda,T_1)) \Rightarrow (y(\lambda,T_4) = x(\lambda,T_4) + b) \quad \forall(\lambda,b) \in \omega.$$

Proof. Let $P : \Lambda \times \mathbb{C}^\infty([T_1,T_4]; \mathbb{R}^m) \to \mathbb{R}^m$ be defined by $P(\lambda,v) = y(T_4)$ with $\dot{y} = f(y,v(t))$, $y(T_1) = x(\lambda,T_1)$. The map P is defined and of class C^∞ on an open neighborhood of $\{(\lambda,u(\lambda,\cdot)); \lambda \in \Lambda\}$. Moreover, it follows from $(A.2)$ that for any i in $[1,n]$ and any λ in Λ there exists w^i_λ in $C^\infty([T_1,T_4]; \mathbb{R}^m)$, the support of which is included in (T_2,T_3), and, if $(e_1,e_2,\dots,e_n)$ is the usual basis of $\mathbb{R}^n$, $\partial P/\partial v(\lambda,u)\cdot w^i_\lambda = e_i$. For $\bar{\lambda} \in \Lambda$, *let* $\mathcal{O}(\bar{\lambda})$ be an open neighborhood of $\bar{\lambda}$ in Λ whose closure is compact and such that

$$(A.5) \qquad \left|\frac{\partial P}{\partial v}(\lambda,u(\lambda,\cdot))\cdot w^i_{\bar{\lambda}} - e_i\right| \leq (1/2n)^{1/2} \quad \forall(i,\lambda) \in [1,m] \times \mathcal{O}(\bar{\lambda}).$$

The open sets $(\mathcal{O}(\bar{\lambda}); \bar{\lambda} \in \Lambda)$ cover Λ; let $(\varphi_j \in C^\infty(\Lambda; [0,1]); j \in \mathbb{N})$ be a partition of unity associated to this covering:

$$(A.6) \qquad \sum_{j=0}^{+\infty} \varphi_j(\lambda) = 1 \quad \forall \lambda \in \Lambda,$$

$$(A.7) \qquad \forall j \in \mathbb{N} \ \ \exists \lambda_j \in \Lambda \text{ such that Support } (\varphi_j) \subset \mathcal{O}(\lambda_j),$$

$$(A.8) \qquad \forall K \text{ compact } \subset \Lambda \ \ \exists p \text{ such that } \sum_{i=0}^{p} \varphi_j(\lambda) = 1 \quad \forall \lambda \in K.$$

Now, let $\widetilde{P} : \Lambda \times \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}^m$ be defined by—with $c = (c_1,c_2,\dots,c_n) \in \mathbb{R}^n$—$\widetilde{P}(\lambda,c) = P(\lambda, u(\lambda,\cdot) + \sum_{i=1}^n \sum_{j=0}^{+\infty} c_i\varphi_j(\lambda)w^i_{\lambda_j})$. Then $\widetilde{P}$ is defined and of class C^∞ on an open neighborhood of $\Lambda \times \{0\}$ in $\Lambda \times \mathbb{R}^n$. Moreover, it follows easily from $(A.5)$

and $(A.6)$ that the vectors $\partial\widetilde{P}/\partial c_i(\lambda, 0)$, $i \in [1,n]$, span all of $\mathbb{R}^n$. Then Lemma A.1 follows from the usual implicit function theorem. □

Appendix B. In this appendix we prove Lemma 3.8. We first prove that Ω_i is a residual subset of Ω if n is at least 3. Let e_n be the last vector of the usual basis of $\mathbb{R}^n$ and let a_n be the last component of $a \in \mathbb{R}^n$. Since Γ can be covered by many countably small compact curves with boundary and since $(0,T)$ is the countable union of the set $[T/k, (k-1)T/k], k \in \mathbb{N}\backslash\{0,1\}$, it suffices to show that for any integer $k \in \mathbb{N}\backslash\{0,1\}$

$$(B.1) \qquad \Omega_i^k := \left\{ v \in \Omega; \frac{\partial x}{\partial a_n}(se_n, t; u_v) \neq 0 \ \forall (s,t) \in \left[\frac{\varepsilon_1}{3}, \frac{2\varepsilon_1}{3}\right] \times \left[\frac{T}{k}, (k-1)\frac{T}{k}\right]\right\}$$

is a residual subset of Γ—perform a suitable diffeomorphism of $\mathbb{R}^n$. Since Ω_i^k is open, it suffices to check that Ω_i^k is dense, which will be proved if one can show that for any (v^*, s^*, t^*) in $\Omega \times (0, \varepsilon_1) \times (0,T)$ there exists an open neighborhood Ω_0^* of v^* and ε in $(0, \text{Min } (s^*, \varepsilon_1 - s^*, t^*, T - t^*))$ such that the open set

$$(B.2)$$
$$\widetilde{\Omega}_0^* = \{v \in \Omega_0^*; \partial x/\partial a_n\,(se_n, t; u_v) \neq 0 \ \forall (s,t) \in [s^* - \varepsilon, s^* + \varepsilon] \times [t^* - \varepsilon, t^* + \varepsilon]\}$$

is dense in Ω_0^*. Let Ω^* be an open neighborhood of v^* included in Ω and, for v in Ω^*, let $y \in C^\infty(B'_\varepsilon \times (0,T); \mathbb{R}^n)$ be defined by $\partial y/\partial t = f(y, v(a,t))$, $y(a,T) = 0$. The linearized control system around $(y(a,\cdot), v(a,\cdot))$ is $\dot{z} = A(a,t)z + B(a,t)w$, where $A(a,t) = \partial f/\partial x(y(a,t), v(a,t))$, $B(a,t) = \partial f/\partial u(y(a,t), v(a,t))$. By (2.48)

$$(B.3) \qquad (y(a,\cdot), v(a,\cdot)) \text{ is supple on } (0,T) \quad \forall a \in B'_{\varepsilon_1}.$$

Let Q be a closed ball of $\mathbb{R}^n \times \mathbb{R}$ centered at (s^*e_n, t^*) such that $Q \subset B'_{\varepsilon_1} \times (0,T)$. By the proof of [INS, Thm. 6.4]—or [G, Chap. 2, §3.8 (B)] and, for more details, [Col, pp. 304–305]; see also [FLMR, Prop. 1.7] for a related result—we know from (B.3) that

$$(B.4) \qquad \frac{\partial z}{\partial t} - A(a,t)z - B(a,t)w = r$$

is algebraically solvable; i.e., there exist an integer q, $(q+1)$ maps $(\mu_i)_{0 \leq i \leq q}$ in $C^\infty(Q; \mathcal{L}(\mathbb{R}^n; \mathbb{R}^n))$, and $(q+1)$ maps $(\nu_i)_{0 \leq i \leq q}$ in $C^\infty(Q; \mathcal{L}(\mathbb{R}^n; \mathbb{R}^m))$ such that, for all r in $C^\infty(Q; \mathbb{R}^n)$, z and w defined by

$$(B.5) \qquad z(a,t) = \sum_{i=0}^{q} \mu_i(a,t) \frac{\partial^i r}{\partial t^i}(a,t), \quad w(a,t) = \sum_{i=0}^{q} \nu_i(a,t) \frac{\partial^i r}{\partial t^i}(a,t)$$

are solutions of (B.4). Note that q, μ_i, and ν_i depend on v but, if Ω^* is a small enough neighborhood of v^*, using the construction of the algebraic inverse given in [Col], we may impose that q is independent of v in Ω^* and that the maps $v \in \Omega^* \to \mu_i \in C^\infty(Q; \mathcal{L}(\mathbb{R}^n; \mathbb{R}^n))$, $v \in \Omega^* \to \nu_i \in C^\infty(Q; \mathcal{L}(\mathbb{R}^n; \mathbb{R}^m))$ are continuous for all i in $[0,q]$. Recall that Q is a *compact* ball; hence the topology on $C^\infty(Q; \mathcal{L}(\mathbb{R}^n; \mathbb{R}^p))$ is the topology in which a net $\{\varphi_\beta\}$ converges to φ if and only if $\{\partial^\alpha \varphi_\beta\}$ converges to $\partial^\alpha \varphi$ uniformly in Q $\forall \alpha \in \mathbb{N}^{n+1}$. Let Q_0 be a closed ball included in the interior of Q and let η in $C^\infty(\mathbb{R}^n \times \mathbb{R}; [0,1])$ be equal to 1 on a neighborhood of Q_0 and such that its support is included in the interior of Q. For a positive integer k and a positive real number M let

$$(B.6)\qquad \Lambda = \left\{ \lambda = (\lambda_0, \lambda_1, \dots, \lambda_k) \in (\mathbb{R}^n)^{k+1} ; \lambda_0 \in B_1, \sum_{i=1}^{k} |\lambda_i|^2 < M \right\}.$$

For λ in Λ we define $\alpha_\lambda \in C^\infty(Q; \mathbb{R}^n)$ by

$$(B.7)\qquad \alpha_\lambda(a,t) = (a_n - s^*)\,\eta(a,t) \sum_{i=0}^{k} (t - t^*)^i \lambda_i$$

and $z_\lambda \in C^\infty(Q; \mathbb{R}^n)$, $w_\lambda \in (Q; \mathbb{R}^m)$ by

$$(B.8)\quad z_\lambda(a,t) = \alpha_\lambda(a,t) + \sum_{i=0}^{q} \mu_i(a,t) \frac{\partial^i r_\lambda}{\partial t^i}(a,t), \quad w_\lambda(a,t) = \sum_{i=0}^{q} \nu_i(a,t) \frac{\partial^i r_\lambda}{\partial t^i}(a,t)$$

with $r_\lambda(a,t) = -\partial\alpha_\lambda/\partial t + A(a,t)\alpha_\lambda(a,t)$. We extend the maps z_λ and w_λ by 0 on $(B'_{\varepsilon_1} \times (0,T)) \backslash Q$ and still denote these extensions by z_λ and w_λ. Then z_λ and w_λ are of class C^∞ on $B'_{\varepsilon_1} \times (0,T)$ and satisfy, on $B'_{\varepsilon_1} \times (0,T)$, $\partial z_\lambda/\partial t(a,t) = A(a,t)z_\lambda(a,t) + B(a,t)w_\lambda(a,t)$. For δ in $(0, +\infty)$ let $v_{\delta,\lambda} = v + \delta w_\lambda \in C^\infty(B'_{\varepsilon_1} \times (0,T); \mathbb{R}^m)$. Let Ω^* be fixed and also let M be fixed. Then, for δ_0 small enough but positive and Ω_0^* a small enough neighborhood of v^*, $v_{\delta,\lambda}$ is in Ω^* for all (δ, v, λ) in $[0,\delta_0] \times \Omega_0^* \times \Lambda$. For δ in $[0,\delta_0]$ and v in Ω_0^* let $G \in C^\infty(B'_{\varepsilon_1} \times (0,T) \times \Lambda; \mathbb{R}^n)$ be defined by $G(a,t,\lambda) = x\left(a,t; u_{v_{\delta,\lambda}}\right)$. Straightforward computations show that, on $Q_0 \times \Lambda$,

$$(B.9)\qquad G(a,t,\lambda) = x(a,t,u_v) + \delta z_\lambda(a,t) + R(a,t,\lambda),$$

where, for some constant C independent of (δ, v) in $[0,\delta_0] \times \Omega_0^*$,

$$(B.10)\qquad \|R\|_{C^2(Q_0\times\Lambda)} \le C\delta^2.$$

Let $G_\lambda(a,t) = G(a,t,b)$ and let $J^1(B'_{\varepsilon_1} \times (0,T); \mathbb{R}^n)$ be the set of 1-jet of mappings from $B'_{\varepsilon_1} \times (0,T)$ to $\mathbb{R}^n$ (see, e.g., [GG, Chap. II, Def. 2.1]) and $\phi : B'_{\varepsilon_1} \times (0,T) \times \Lambda \to J^1(B'_{\varepsilon_1} \times (0,T); \mathbb{R}^n)$ be the map that associates the 1-jet of G_λ at (a,t) with (a,t,λ). Let ε be a positive real number and let W be the submanifold of $J^1(B'_{\varepsilon_1} \times (0,T); \mathbb{R}^n)$ defined by

$$(B.11)\quad W = \left\{ \left(a, t, x(a,t), \frac{\partial x}{\partial a}(a,t) \right) ; (a,t) \in (0, \varepsilon_1 e_n) \times (0,T), \right.$$
$$\left. |a_n - s^*| + |t - t^*| < \varepsilon, \frac{\partial x}{\partial a_n}(a,t) = 0, \right\},$$

where $(0, \varepsilon_1 e_n) = \{\tau\varepsilon_1 e_n ; \tau \in (0,1)\}$. From (B.7) we get

$$(B.12)\quad \frac{\partial z_\lambda}{\partial a_n}(s^* e_n, t^*) = \lambda_0 + \sum_{i=0}^{q} \mu_i(s^*, t^*) \left(\frac{\partial^i}{\partial t^i} \left(-\dot{p}(t) + A(s^* e_n, t^*) p(t) \right) \right) \Bigg|_{t=t^*}$$

with $p(t) = \sum_{i=0}^{k} (t - t^*)^i \lambda_i$. We now choose the mapping v^* for v and denote by A^*, B^*, z^*_λ, ϕ^* the corresponding maps. For λ_0 in B_1 let $d : (0,T) \longrightarrow \mathbb{R}^n$ be the solution of $\dot{d} = A^*(s^* e_n, t)\, d$, $d(t^*) = \lambda_0$. Expanding d near t^* we get

$$(B.13) \qquad d(t) = \lambda_0 + \sum_{i=1}^{k} \lambda_i(\lambda_0)(t - t_i^*)^i + 0\left(|t - t^*|^{k+1}\right).$$

We choose $k > q$ and M such that $\sum_{i=1}^{k} |\lambda_i(\lambda_0)|^2 < M - 1$ for all l_0 in B_1. For $\lambda = (\lambda_0, \lambda_1(\lambda_0), \ldots, \lambda_k(\lambda_0))$ we get that

$$(B.14) \qquad \frac{\partial z^*_\lambda}{\partial a_n}(s^* e_n, t^*) = \lambda_0.$$

Using $(B.9)$, $(B.10)$, and $(B.14)$, we get that, if δ and ε are small enough but positive, ϕ^* is transversal to W; moreover, straightforward computations show that, similarly, if Ω_0^* is a small enough neighborhood of v^*, if δ and ε are small enough but positive, then

$$(B.15) \qquad \phi \text{ is transversal to } W \quad \forall v \in \Omega_0^*.$$

We take such Ω_0^*, ε, and δ. We also require Ω_0^* to be open. From $(B.15)$ and a classical result on transversality (see, e.g., [GG, Chap. II, Cor. 4.7]) we get that, if $v \in \Omega_0^*$, then, for a dense set of λ in Λ,

$$(B.16) \qquad \phi(\cdot, \lambda) \text{ is transversal to } W.$$

But the dimension of $B'_{\varepsilon_1} \times (0,T)$ is $n+1$, the dimension of W is $n^2 + n + 2$, and the dimension of $J^1(B'_\varepsilon \times (0,T); \mathbb{R}^n)$ is $n^2 + 3n + 1$. Hence, if $n \geq 3$, $(B.16)$ implies that the image of $\phi(\cdot, \lambda)$ does not meet W. Since $v_{\delta,\lambda} \longrightarrow v$ as $\lambda \longrightarrow 0$, we get that $\widetilde{\Omega}_0^*$ defined by $(B.2)$ is dense in Ω_0^*, and so Ω_i is a residual subset of Ω.

We now prove that Ω_0 is a residual subset of Ω if n is at least 3. One proceeds as for Ω_0 with the following modifications:

- in $(B.1)$ and $(B.2)$ one replaces $\partial x / \partial a_n$ by x;
- in $(B.7)$ one suppresses $(a_n - s^*)$;
- $J^1(B'_{\varepsilon_1} \times (0,T); \mathbb{R}^n)$ is replaced by $J^0(B'_{\varepsilon_1} \times (0,T); \mathbb{R}^n)$ and ϕ associates the 0-jet of G_λ at (a,t) with (a,t,λ);
- W is now submanifold of $J^0(B'_{\varepsilon_1} \times (0,T); \mathbb{R}^n)$ defined by

$$W = \{(a, t, x(a,t)) ; (a,t) \in (0, \varepsilon_1 e_n) \times (0,T); |a_n - s^*| + |t - t^*| < \varepsilon; x(a,t) = 0\};$$

- in $(B.14)$ one replaces $\partial z^* / \partial a_n$ by z^*;
- the dimension of W is 2 and the dimension of $J^0(B'_{\varepsilon_1} \times (0,T); \mathbb{R}^n)$ is $2n+1$.

We finally prove that Ω_1 is a residual subset of Ω if n is at least 4. One again proceeds as for Ω_i but with the following modifications. One first notices that it suffices to prove that for any (s_1^*, s_2^*) in $(0, \varepsilon_1 e_n)$ with $s_1^* \neq s_2^*$, any v^* in Ω, and any t^* in $(0,T)$ there exists ε in $(0, +\infty)$, an open neighborhood Ω_0^* of v^* such that if $\widetilde{\Omega}_0^*$ is the set of v in Ω_0^* satisfying for all s_1 in $[s_1^* - \varepsilon, s_1^* + \varepsilon]$, all s_2 in $[s_2^* - \varepsilon, s_2^* + \varepsilon]$, and all t in $[t^* - \varepsilon, t^* + \varepsilon]$ $x(s_1 e_n, t; u_v) \neq x(s_2 e_n, t; u_v)$, then

$$(B.17) \qquad \widetilde{\Omega}_0^* \text{ is dense in } \Omega_0^*.$$

Note that $\widetilde{\Omega}_0^*$ is open. Let $\bar{\varepsilon}$ be a positive real number such that the intersection of $[s_1^* - 2\bar{\varepsilon}, s_1^* + 2\bar{\varepsilon}]$ with $[s_2^* - 2\bar{\varepsilon}, s_2^* + 2\bar{\varepsilon}]$ is empty, $[s_1^* - 2\bar{\varepsilon}, s_1^* + 2\bar{\varepsilon}]$ and $[s_2^* - 2\bar{\varepsilon}, s_2^* + 2\bar{\varepsilon}]$ are included in $(0, \varepsilon_1)$, and $[t^* - 2\bar{\varepsilon}, t^* + 2\bar{\varepsilon}]$ is included in $(0,T)$. Let

$$(B.18) \qquad Q = \left\{(a,t) \in B'_{\varepsilon_1} \times (0,T); |a - s_1^* e_n|^2 + |t - t^*|^2 \leq 4\bar{\varepsilon}^2\right\},$$

$$(B.19) \qquad Q_0 = \left\{(a,t) \in B'_{\varepsilon_1} \times (0,T); |a - s_1^* e_n|^2 + |t - t^*|^2 \leq \bar{\varepsilon}^2\right\}.$$

Now we replace $(B.7)$ by $\alpha_\lambda(a,t) = \eta(a,t) \sum_{i=0}^k (t - t^*)^i \lambda_i$ and define G in $C^\infty(B'_{\varepsilon_1} \times B'_{\varepsilon_1} \times (0,T) \times \Lambda; \mathbb{R}^n)$ by $G(a_1, a_2, t, \lambda) = x(a_1, t; u_{v_{\delta,\lambda}}) - x(a_2, t; u_{v_{\delta,\lambda}})$, and ϕ is now the map from $B'_{\varepsilon_1} \times B'_{\varepsilon_1} \times (0,T) \times \Lambda$ to $J^0(B'_{\varepsilon_1} \times B'_{\varepsilon_1} \times (0,T); \mathbb{R}^n)$ that associates (a_1, a_2, t, λ) the 0-jet of $G(\cdot,\cdot,\cdot,\lambda)$ at (a_1, a_2, t). Computations similar to those of the proof of Lemma B.8 show that if M is large enough, Ω_0^* is a small enough open neighborhood of v^*, and δ and ε are small enough but positive, then $(B.15)$ holds again with

$$(B.20) \qquad W = ((s_1^* - \varepsilon) e_n, (s_1^* + \varepsilon) e_n) \times ((s_2^* - \varepsilon) e_n, (s_2^* + \varepsilon) e_n) \times (t^* - \varepsilon, t^* + \varepsilon) \times \{0\}.$$

Since the sum of the dimensions of W and $(B'_{\varepsilon_1} \times B'_{\varepsilon_1} \times (0,T))$ is less than the dimension of $J^0(B_{\varepsilon'_1} \times B'_{\varepsilon_1} \times (0,T); \mathbb{R}^n)$ if (and only if) $n \geq 4$ we again get $(B.17)$ by again using [GG, Chap. II, Cor. 4.7].

Appendix C. The goal of this appendix is to prove Lemma 2.12. To simplify the notations we first perform a change of variables. Let $\phi : \mathbb{R}^n \times (0,1] \to \mathbb{R}^n \times [0,T)$ be defined by $\phi(\widehat{y}, \widehat{t}) = (y(\widehat{y}, \widehat{t}), T(1 - \widehat{t}))$, where, if we denote by y_n the last component of $y \in \mathbb{R}^n$, $\partial y / \partial \widehat{t} = -TA(y_n, T(1 - \widehat{t}))\, y$, $y(\widehat{y}, 1) = \widehat{y}$. The domain of definition of ϕ is an open subset D of $\mathbb{R}^n \times (0,1]$ that contains $\mathbb{R}^n \times \{1\}$. Note that D contains $\{0\} \times \mathbb{R} \times (0,1]$ and $\phi((0,\widehat{s}), \widehat{t}) = ((0,\widehat{s}), T(1 - \widehat{t}))$ for all $(\widehat{s}, \widehat{t})$ in $\mathbb{R} \times (0,1]$. Note also that ϕ is a diffeomorphism of class C^∞ between D and $\phi(D)$. By pulling back the time-varying control system $dy/dt = F(y,t,u)$ with ϕ we get a time-varying control system that we denote by $d\widehat{y}/d\widehat{t} = \widehat{F}(\widehat{y}, \widehat{t}, u)$. One has $\widehat{F}(0, \widehat{s}, \widehat{t}, 0) = 0$ for all $(\widehat{s}, \widehat{t})$ in $\mathbb{R} \times (0,1]$. Let, for $\widehat{s}$ in $\mathbb{R}$ and $\widehat{t}$ in $(0,1]$, $\widehat{A}(\widehat{s}, \widehat{t}) = \partial \widehat{F} / \partial \widehat{y}((0,\widehat{s}), \widehat{t}, 0)$ and $\widehat{B}(\widehat{s}, \widehat{t}) = \partial \widehat{F} / \partial u((0,\widehat{s}), \widehat{t}, 0)$. One easily checks that $\widehat{A} = 0$ and, using (2.64), we have

$$(C.1) \qquad \operatorname{Span}\left\{\frac{\partial^i \widehat{B}}{\partial \widehat{t}^i}(\widehat{s}, \widehat{t}) w; \omega \in \mathbb{R}^m, i \geq 0\right\} = \mathbb{R}^n \quad \forall (\widehat{s}, \widehat{t}) \in \mathbb{R} \times (0,1].$$

Note also that, since D contains a neighborhood of $(\mathbb{R}^n \times \{1\}) \cup (\{0\} \times \mathbb{R} \times (0,1])$ in $\mathbb{R}^n \times (0,1]$, there exists g in $C^\infty(\mathbb{R}^n \times (0,1] \times \mathbb{R}^m; \mathbb{R}^n)$ such that $g = \widehat{F}$ on a neighborhood of $(\mathbb{R}^n \times \{1\}) \cup (\{0\} \times \mathbb{R} \times (0,1])$. Hence Lemma 2.12 is proved if one can show the following proposition.

PROPOSITION C.1. *Let* $g : \mathbb{R}^n \times (0,1] \times \mathbb{R}^m \to \mathbb{R}^n, (x,t,u) \to g(x,t,u)$ *be such that*

$$(C.2) \qquad g \in C^\infty(\mathbb{R}^n \times (0,1] \times \mathbb{R}^m; \mathbb{R}^n),$$

$$(C.3) \qquad g((0,s),t,0) = 0 \quad \forall (s,t) \in \mathbb{R} \times (0,1],$$

$$\frac{\partial g}{\partial x}((0,s),t,0)=0 \quad \forall (s,t)\in \mathbb{R}\times(0,1], \tag{C.4}$$

$$\operatorname{Span}\left\{\frac{\partial^i B}{\partial t^i}(s,t)w; i\geq 0, w\in\mathbb{R}^m\right\}\supset \mathbb{R}^{n-1}\times\{0\} \quad \forall (s,t)\in\mathbb{R}\times(0,1] \tag{C.5}$$

with $B(s,t)=\partial g/\partial u((0,s),t,0)$. *Then, for any* ε *in* $C^0(\mathbb{R}\times(0,1];(0,+\infty))$, *there exist* u *in* $C^\infty(Q\times(0,1];\mathbb{R}^m)$ *with* $Q=\{a=(b,s)\in\mathbb{R}^{n-1}\times\mathbb{R}\simeq\mathbb{R}^n; |b|\leq 1\}$ *and* $x\in C^\infty(Q\times(0,1];\mathbb{R}^n)$ *such that*

$$|u(a,t)|\leq\varepsilon(s,t) \quad \forall(a,t)\in Q\times(0,1], \tag{C.6}$$

$$|x(a,t)-(0,s)|\leq\varepsilon(s,t) \quad \forall(a,t)\in Q\times(0,1], \tag{C.7}$$

$$\frac{\partial x}{\partial t}(a,t)=g(x(a,t),t,u(a,t)) \quad \forall(a,t)\in Q\times(0,1], \tag{C.8}$$

$$u((0,s),t)=0 \quad \forall(s,t)\in Q\times(0,1], \tag{C.9}$$

$$x((0,s),t)=(0,s) \quad \forall(s,t)\in Q\times(0,1], \tag{C.10}$$

$$a\in Q\rightarrow x(a,t)\in\mathbb{R}^n \textit{ is an embedding of } Q \textit{ into } \mathbb{R}^n \quad \forall t\in(0,1]. \tag{C.11}$$

Let us prove Proposition C.1. For a C^∞-submanifold M of $\mathbb{R}^p$, which has a boundary and is of dimension p, we define a topology on $C^\infty(M;\mathbb{R}^q)$ in the same way as we defined our topology on $C^\infty(\mathcal{O};\mathbb{R}^q)$ in Step 1 in §2, where $\mathcal{O}$ is an open subset of $\mathbb{R}^p$: ε and A that appear in this definition are now in $C^0(M;(0,+\infty))$ instead of $C^0(\mathcal{O};(0,+\infty))$. Let us emphasize that this topology on $C^\infty(M;\mathbb{R}^q)$ is quite different from the topology induced on $C^\infty(M;\mathbb{R}^q)$ by the topology on $C^\infty(\overset{\circ}{M};\mathbb{R}^q)$. Our first lemma used to prove Proposition C.1 is the following.

LEMMA C.2. *Assume* (*C*.2) *and* (*C*.3). *Then there exist an open neighborhood* Ω *of* 0 *in* $C^\infty(Q\times(0,1];\mathbb{R}^m)$ *and a continuous map* $X:\Omega\rightarrow C^\infty(Q\times(0,1];\mathbb{R}^m)$ $u\rightarrow X(u)=x$ *such that* (*C*.8) *holds for all* u *in* Ω *and, for all* u *in* Ω *and all* $a=(b,s)$ *in* Q,

$$(u(a,t)=0 \quad \forall t\in(0,1])\Rightarrow(X(u)(a,t)=(0,s) \quad \forall t\in(0,1]). \tag{C.12}$$

Note that if g is in $C^\infty(\mathbb{R}^n\times[0,1]\times\mathbb{R}^m;\mathbb{R}^m)$ then Lemma C.2 follows directly from the classical theorems on existence and smoothness—with respect to time, parameters (a), and initial data—of the solution to the Cauchy problem (see, e.g., [Ha, Chap. V]) applied to $\dot{x}=g(x,t,u(a,t)), x(a,0)=(0,s)$. When g is only in $C^\infty(\mathbb{R}^n\times(0,1]\times\mathbb{R}^m;\mathbb{R}^m)$ one just needs to modify slightly the proof of these theorems. Let us limit ourselves to showing how to modify the usual proof of the existence of a solution to the Cauchy problem in order to get a solution to (*C*.8) and (*C*.10). For η in $C^0(\mathbb{R}\times(0,1];(0,1])$ let E be the space of maps x in $C^0(Q\times(0,1];\mathbb{R}^n)$ such that $|x(a,t)-(0,s)|\leq\eta(s,t)$ for all (a,t) in $Q\times(0,1]$. We define a metric d on E by

$d(x_1,x_2) = \text{Sup}\{\eta(s,t)\,|x_1(a,t)-x_2(a,t)|\,;(a,t)\in Q\times(0,1]\}$. Note that (E,d) is a complete metric space. For u in $C^0(Q\times(0,1];\mathbb{R}^m)$ and x in E let $T_u(x):Q\times(0,1]\to\mathbb{R}^n$ be defined by $T_u(x)(a,t)=(0,s)+\int_0^t g(x(a,\tau),\tau,u(a,\tau))d\tau$. Then straightforward computations show the existence of η and δ in $C^0(\mathbb{R}\times(0,1];(0,1])$ such that, if u is in $\Omega_0 := \{u\in C^\infty(Q\times(0,1];\mathbb{R}^m)\,;|u(a,t)|\le\lambda(s,t)\ \forall(a,t)\in Q\times(0,1]\}$, then $T_u(x)$ is defined on $Q\times(0,1]$ for all x in E and, moreover, for all x_1 and x_2 in E, $T_u(x_1)\in E$, and $d(T_u(x_1),T_u(x_2))\le d(x_1,x_2)/2$. Hence, for any u is in Ω_0, T_u has a unique fix point; this fix point is a solution to $(C.8)$ and $(C.10)$.

We now assume that the assumptions of Proposition C.1 hold. Let ε be in $C^\infty(\mathbb{R}\times(0,1];(0,+\infty))$. We are going to construct u in Ω satisfying $(C.6)$ and $(C.9)$ such that $(C.7)$ and $(C.11)$ hold with $x=X(u)$. (Note that $(C.8)$ holds for $x=X(u)$ and that $(C.10)$ for $x=X(u)$ follows from $(C.9)$ and $(C.12)$.) Let $v\in C^\infty((0,1];[0,1])$ be such that $v(t)=\exp-(1/(t-(1/(i+2))))$ for all integers i and for all t in $(1/(i+2),2/(2i+3)]$. Let Λ be the set of sequences $\lambda=(\lambda_i;i\in\mathbb{N})$ of maps in $C^\infty(\mathbb{R};(0,1])$. For λ in Λ we defined two maps $w_\lambda\in C^\infty(\mathbb{R}\times(0,1];[0,1])$ and $u_\lambda\in C^\infty(Q\times(0,1];\mathbb{R}^m)$ by $w_\lambda(s,t)=\lambda_i^2(s)v(t)$ if $t\in(1/(i+2),1/(i+1)]$, $u_\lambda(a,t)=w_\lambda(s,t)B(s,t)^*(b,0)$ if $t\in(1/(i+2),1/(i+1)]$, where $B(s,t)^*\in\mathcal{L}(\mathbb{R}^n;\mathbb{R}^m)$ is the transpose of $B(s,t)$. Note that $(C.9)$, with $u=u_\lambda$, is satisfied. We will say that a property P holds if $\lambda(\in\Lambda)$ is small enough if there exists a sequence $(\Lambda_i;i\in\mathbb{N})$ of neighborhoods of 0 in $C^\infty(\mathbb{R};\mathbb{R})$ such that, if $\lambda\in\Lambda$ satisfies $\lambda_i\in\Lambda_i$ for all i in $\mathbb{N}$, then P holds. For example, for λ small enough $u_\lambda\in\Omega$, where Ω is as in Lemma C.2, so for λ small enough $x_\lambda := X(u_\lambda)$ is defined. Note also that for any sequence $(\Lambda_i;i\in\mathbb{N})$ of neighborhoods of 0 in $C^\infty(\mathbb{R};\mathbb{R})$ there exists $\lambda\in\Lambda$ such that $\lambda_i\in\Lambda_i$ for all $i\in\mathbb{N}$. Therefore, Proposition C.1 is proved if one can show that, for λ small enough, $(C.6)$, $(C.7)$, and $(C.11)$ with $u=u_\lambda$ and $x=x_\lambda$ are satisfied. For simplicity let us write u,w, and x for u_λ,w_λ, and x_λ. Clearly $(C.6)$ is satisfied if λ is small enough. Moreover, X is continuous and, by $(C.12)$, $X(0)(a,t)=(0,s)\ \forall(a,t)\in Q\times(0,1]$; so, if λ is small enough, $(C.7)$ holds. It only remains to prove that, if λ is small enough, $(C.11)$ holds.

Let $\overline{x}:Q\times[0,1]\to\mathbb{R}^m$ be defined by $\overline{x}(a,t)=(0,s)+\int_0^t B(s,\tau)u(a,\tau)d\tau$. For λ small enough $\overline{x}$ is well defined and belongs to $C^\infty(Q\times[0,1];\mathbb{R}^m)$. Our next lemma is as follows.

LEMMA C.3. *There exists $C_1\in C^0(\mathbb{R}\times(0,1];(0,+\infty))$ such that, if λ is small enough, then, for all (a,t) in $Q\times(0,1]$,*

$$(C.13)\qquad \left|\frac{\partial x}{\partial a}-\frac{\partial\overline{x}}{\partial a}\right|(a,t)\le\int_0^t C_1(s,\tau)\left(w(s,\tau)^2+\left(\frac{\partial w}{\partial s}\right)^2(s,\tau)\right)d\tau.$$

The proof of this lemma follows easily from $(C.12)$, $(C.13)$, $(C.18)$, and usual estimates on solutions of differential equations. We omit it.

Let $M\in C^\infty(\mathbb{R}\times(0,1];\mathcal{L}(\mathbb{R}^{n-1};\mathbb{R}^{n-1}))$ and $N\in C^\infty(\mathbb{R}\times(0,1];\mathcal{L}(\mathbb{R}^{n-1};\mathbb{R}))$ be such that $\overline{x}(a,t)=(M(s,t)b,s+N(s,t)b)$. Note that $\overline{x}(a,t)=\int_0^t B(s,\tau)B^*(s,\tau)(b,0)d\tau+(0,s)$; so M and N exist, and, moreover, $M(s,t)=M^*(s,t)$ for all (s,t) in $\mathbb{R}\times(0,1]$ and $b\cdot(M(s,t)b)\ge0$ for all (b,s,t) in $\mathbb{R}^{n-1}\times\mathbb{R}\times(0,1]$, where $\cdot$ denotes the scalar product in $\mathbb{R}^{n-1}$. Let μ be positive real number. Clearly, for λ small enough,

$$(C.14)\qquad |M(s,t)|+\left|\frac{\partial M}{\partial s}(s,t)\right|+|N(s,t)|+\left|\frac{\partial N}{\partial s}(s,t)\right|\le\mu\quad\forall(s,t)\in\mathbb{R}\times(0,1]$$

and by $(C.13)$ we have, still for λ small enough,

$$(C.15)\qquad \left|\frac{\partial x}{\partial a}(a,t)-\frac{\partial\overline{x}}{\partial a}(a,t)\right|\le\mu\quad\forall(a,t)\in Q\times(0,1].$$

Let us write $x(a,t) = (y(a,t), z(a,t)) \in \mathbb{R}^{n-1} \times \mathbb{R}$. Note that

$$(C.16)\qquad \frac{\partial \overline{x}}{\partial a}(a,t)(b',s') = \left(M(s,t)b' + s'\frac{\partial M}{\partial s}(s,t)b, s' + s'\frac{\partial N}{\partial s}(s,t)b + N(s,t)b'\right).$$

Taking μ small enough and using $(C.14)$, $(C.15)$, and $(C.16)$ we get that, if λ is small enough, $|\partial z/\partial s(a,t) - 1| \leq 1/2$ and, therefore, for any b in $\mathbb{R}^{n-1}$ with $|b| \leq 1$ and any t in $[0,1]$, the map $s \in \mathbb{R} \to z((b,s),t) \in \mathbb{R}$ is a diffeomorphism of $\mathbb{R}$ onto $\mathbb{R}$. Let σ in $C^\infty(Q \times (0,1]; \mathbb{R})$ be such that $z((b, \sigma(a,t)), t) = s$ for all (a,t) in $Q \times (0,1]$. Note that, if λ is small enough, we have for all (a,t) in $Q \times (0,1]$,

$$(C.17)$$
$$\left|\frac{\partial z}{\partial b}((b,\sigma(a,t)),t)\right| \leq 2\left\{\left|\frac{\partial x}{\partial a}((b,\sigma(a,t)),t) - \frac{\partial \overline{x}}{\partial a}((b,\sigma(a,t)),t)\right| + |N(\sigma(a,t),t)|\right\}.$$

Clearly, it suffices to prove that, for λ small enough,

$$(C.18)\qquad b \in B^{n-1} \to y^*(b,s,t) \in \mathbb{R}^{n-1} \text{ is an embedding } \quad \forall (s,t) \in \mathbb{R} \times (0,1],$$

where $B^{n-1} = \{b \in \mathbb{R}^{n-1}; |b| \leq 1\}$, $y^*(b,s,t) = y((b, \sigma((b,s),t)),t)$. Using Lemma C.3, $(C.16)$, and $(C.17)$ one gets the existence of $C_2 \in C^0(\mathbb{R} \times (0,1]; (0,+\infty))$ such that, if λ is small enough and with $\sigma = \sigma((b,s),t) = \sigma(a,t)$,

$$(C.19)\qquad \left|\frac{\partial y^*}{\partial b}(b,s,t) - M(\sigma,t)\right| \leq \int_0^t C_2(\sigma,\tau)\left(w^2(\sigma,\tau) + \left(\frac{\partial w}{\partial s}\right)^2(\sigma,\tau)\right)d\tau$$

for all (a,t) in $Q \times (0,1]$. Let us assume, for the moment, the following lemma.

LEMMA C.4. *For any $K \in C^0(\mathbb{R} \times (0,1]; (0,+\infty))$ then, if λ is small enough, we have for all (b', s, t) in $\mathbb{R}^{n-1} \times \mathbb{R} \times (0,1]$,*

$$(C.20)\qquad b' \cdot (M(s,t)b') \geq |b'|^2 \int_0^t K(s,\tau)\left(w^2(s,\tau) + \left(\frac{\partial w}{\partial s}\right)^2(s,\tau)\right)d\tau.$$

From Lemma C.4 we get that, if λ is small enough,

$$(C.21)\qquad b' \cdot \frac{\partial y^*}{\partial b}(b,s,t)b' > 0 \quad \forall (b',b,s,t) \in (\mathbb{R}^{n-1}\backslash\{0\}) \times B^{n-1} \times \mathbb{R} \times (0,1].$$

Note that $(C.21)$ implies $(C.18)$. Indeed, $(C.21)$ implies that $\partial y^*/\partial b(b,s,t)$ is invertible for all (b,s,t) in $B^{n-1} \times \mathbb{R} \times (0,1]$ and that, for all (b_1, b_2, s, t) in $B^{n-1} \times B^{n-1} \times \mathbb{R} \times (0,1]$ with $b_1 \neq b_2$, the real valued function $\tau \in [0,1] \to y^*(\tau b_2 + (1-\tau)b_1, s, t) \cdot (b_2 - b_1)$ is increasing, which implies that for all (s,t) in $\mathbb{R} \times (0,1]$, $b \in B^{n-1} \to y^*(b,s,t)$ is one-to-one.

Now we need only to prove Lemma C.4. Let us assume, for the moment, the following lemma.

LEMMA C.5. *For any integer i, there exist δ_i in $C^0(\mathbb{R}; (0,+\infty))$ such that, for all (b,s,t) in $\mathbb{R}^{n-1} \times \mathbb{R} \times [1/(i+2), 1/(i+1)]$,*

$$(C.22)\qquad \int_{1/(i+2)}^t |B^*(s,\tau)(b,0)|^2 v(\tau)d\tau \geq \delta_i(s)|b|^2 \int_{1/(i+2)}^t v^2(\tau)d\tau.$$

Let $K_i(s) = \operatorname{Max}\{K(s,t); t \in [1/(i+2), 1/(i+1)]\}$; then, if for all integer i, $\lambda_i^2(s) + 4\left(d\lambda_i/ds\right)^2(s) \leq \delta_i(s)/K_i(s)$ for all s in $\mathbb{R}$, $(C.20)$ follows from $(C.22)$.

Now we need only to prove Lemma C.5. Clearly, it suffices to prove that for any $(\bar{b}, \bar{s}, i) \in (\mathbb{R}^{n-1}\backslash\{0\}) \times \mathbb{R} \times \mathbb{N}$ there exists a neighborhood V of $(\bar{b}, \bar{s}, 1/(i+2))$ in $(\mathbb{R}^{n-1}\backslash\{0\}) \times \mathbb{R} \times [1/(i+2), 1/(i+1)]$ such that

$$(C.23) \qquad \int_{1/(i+2)}^{t} |B^*(s,\tau)(b,0)|^2 v(\tau) d\tau \geq \int_{1/(i+2)}^{t} v^2 d\tau$$

for all (b,s,t) in V. It follows from $(C.5)$ that there exists an integer p such that

$$(C.24)$$
$$\frac{\partial^j}{\partial t^j} \left|B^*(\bar{s},t)(\bar{b},0)\right|^2_{|_{t=1/(i+2)}} = 0 \quad \forall j \in [0,p-1], \ \frac{\partial^p}{\partial t^p} \left|B^*(\bar{s},t)(\bar{b},0)\right|^2_{|_{t=1/(i+2)}} \neq 0.$$

From $(C.24)$ and the Malgrange preparation theorem (see, e.g., [GG, Chap. IV, §2]) it follows that there exist p maps $(\theta_j; 0 \leq j \leq p-1)$ in $C^\infty(\mathbb{R}^{n-1} \times \mathbb{R}; \mathbb{R})$, a compact neighborhood W of $(\bar{b}, \bar{s}, 1/(i+2)$ in $\mathbb{R}^{n-1} \times \mathbb{R} \times \mathbb{R}$, and a positive real number μ_1 such that, with $t_i = 1/(i+2)$, $|B^*(s,t)(b,0)|^2 \geq \mu_1 |(t-t_i)^p + \sum_{j=0}^{p-1} \theta_j(b,s)\,(t-t_i)^j|$ for all (b,s,t) in W and, therefore, since W is compact, there exists a positive real number μ_2 such that, for all (b,s,t) in W, $|B^*(s,t)(b,0)|^2 \geq \mu_1\mu_2((t-t_i)^p + \sum_{j=0}^{p-1} \theta_j(b,s)(t-t_i)^j)^2$. So Lemma C.5 is a consequence of the following lemma.

LEMMA C.6. *Let R be a positive real number and let p be an integer. Then there exists a positive real number t_0 such that*

$$(C.25) \qquad \int_0^t \left(\tau^p + \sum_{j=0}^{p-1} C_j \tau^j\right)^2 \exp(-1/\tau) d\tau \geq \int_0^t \exp(-2/\tau) d\tau$$

for all t in $[0,t_0]$ and sequence of p real numbers $(C_j; j \in [0,p-1])$ with $|C_j| \leq R$ for all j in $[0,p-1]$.

Proof. Let $\mathcal{C}$ be the set of sequences C of real numbers $(C_i; i \in [0,p-1])$ such that $|C_i| \leq M$ for all i in $[0,p-1]$. For C in $\mathcal{C}$ let $P_C(t) = t^p + \sum_{j=0}^{p-1} C_j t^j$. Let us first notice that, for all $C \in \mathcal{C}$,

$$(C.26) \quad \int_0^t P_C(\tau)^2 \exp(-1/\tau) d\tau \geq (\exp(-3/(2t))) \int_{2t/3}^t P_C(\tau)^2 d\tau \quad \forall t \in (0,+\infty).$$

Note that $\int_{2t/3}^t P_C(\tau)^2 d\tau$ is an algebraic function of t and $(C_i; i \in [0,p-1])$ which is positive if $t > 0$. Hence, by the Lojasiewicz inequality (see, e.g., [BCR, Chap. 2, §6.7]), there exist an integer N and a real number δ such that $\int_{2t/3}^t P_C(\tau)^2 d\tau \geq \delta t^N$ for all (C,t) in $\mathcal{C} \times [0,1]$, which, with $(C.26)$, gives Lemma C.6. □

REFERENCES

[Ba] A. BACCIOTTI, *Local Stabilization of Nonlinear Control Systems*, Series on Advances in Mathematics for Applied Sciences, Vol. 8, World Scientific, Singapore, New Jersey, London, Hong Kong, 1992.

[Br] R. W. BROCKETT, *Asymptotic stability and feedback stabilization*, in Differential Geometric Control Theory, R. W. Brockett, R. S. Millman, and H. J. Sussmann eds., Progr. Math. 27, Birkhäuser, Basel, Boston, 1983, pp. 181–191.

[BCR] J. Bochnak, M. Coste, and M.-F. Roy, *Géométrie algébrique réelle*, Ergebnisse der Mathematik und ihrer Grenzgebiete, Folge 3, Bd 12, Springer-Verlag, Berlin, Heidelberg, New York, 1987.

[BS] R. M. Bianchini and G. Stefani, *Sufficient conditions for local controllability*, in Proc. 25th Conference on Decision and Control, Athens, IEEE, 1986, pp. 967–970.

[Ch] W. L. Chow, *Uber systeme von linearen partiellen differentialgleichungen ester ordnung*, Math. Ann., 117 (1940–41), pp. 98–105.

[Co1] J.-M. Coron, *Global asymptotic stabilization for controllable systems without drift*, Math. Control Signals Systems, 5 (1992), pp. 295–312.

[Co2] ———, *Links between local controllability and local continuous stabilization*, IFAC Nonlinear Control Systems Design 1992, M. Fliess ed., Bordeaux, 1992, pp. 165–171.

[Co3] ———, *Linearized control systems and applications to smooth stabilization*, SIAM J. Control Optim., 32 (1994), pp. 358–386.

[Co4] ———, *Sur la stabilisation des systèmes commandables et observables*, Ann. Fac. Sci. Toulouse Math. III, 1994, to appear.

[DMK] W. P. Dayawansa, C. Martin, and G. Knowles, *Asymptotic stabilization of a class of smooth two-dimensional systems*, SIAM J. Control Optim., 28 (1990), pp. 1321–1349.

[F] A. F. Filippov, *Differential Equations with Discontinuous Right Hand Sides,* Mathematics and Its Applications, Kluwer Academic Publishers, Dordrecht, (1988).

[FLMR] M. Fliess, J. Lévine, P. Martin, and P. Rouchon, *Sur les systèmes non linéaires différentiellement plats*, C. R. Acad. Sci. Paris Ser. I Math., 315 (1992), pp. 619–624.

[G] M. Gromov, *Partial Differential Relations*, Ergebnisse der Mathematik und ihrer Grenzgebiete, Folge 3, Bd. 9, Springer-Verlag, Berlin, Heidelberg, 1986.

[GG] M. Golubitsky and V. Guillemin, *Stable Mappings and their Singularities*, in Grad. Texts in Math. 14, Springer, New York, Heidelberg, Berlin, 1973.

[Ha] P. Hartman, *Ordinary Differential Equations,* John Wiley, New York, London, Sydney, 1964.

[He1] H. Hermes, *On the synthesis of a stabilizing feedback control via Lie algebraic methods*, SIAM J. Control Optim., 18 (1980), pp. 352–361.

[He2] ———, *Control systems which generate decomposable Lie algebras*, J. Differential Equations, 44 (1982), pp. 166–187.

[INS] A. Ilchmann, I. Nürnberger, and W. Schmale, *Time-varying polynomial matrix systems*, Internat. J. Control, 40 (1984), pp. 329–362.

[Kai] T. Kailath, *Linear Systems*, Prentice Hall, London, 1980.

[Kaw1] M. Kawski, *Stabilization of nonlinear systems in the plane*, Systems Control Lett., 12 (1989), pp. 169–175.

[Kaw2] ———, *High-order small time local controllability*, in Nonlinear Controllability and Optimal Control, H. J. Sussmann, ed., Monographs and Textbooks in Pure and Applied Mathematics, 113, Marcel Dekker, New York, 1990, pp. 431–467.

[Ku] J. Kurzweil, *On the inversion of Lyapunov's second theorem on stability of motion*, Ann. Math. Soc. Trans. Ser. 2, 24 (1956) pp. 19–77.

[Sa] C. Samson, *Velocity and torque feedback control of a non-holonomic cart*, Internat. Workshop in Adaptative and Nonlinear control: Issues in Robotics, Grenoble, 1990, Lecture Notes in Control and Information Sciences, Vol. 162, Springer-Verlag, Berlin, New York, pp. 125–151.

[So1] E. D. Sontag, *Finite dimensional open-loop control generators for nonlinear systems*, Internat. J. Control, 47 (1988), pp. 537–556.

[So2] ———, *Mathematical Control Theory: Deterministic Finite Dimensional Systems*, Texts in Applied Mathematics 6, Springer-Verlag, New York, Berlin, Heidelberg, London, Paris, Tokyo, Hong Kong, 1990.

[So3] ———, *Universal nonsingular controls*, Systems Control Lett., 19 (1992), pp. 221–224.

[Su1] H. J. Sussmann, *Subanalytic sets and feedback control*, J. Differential Equations, 31 (1979), pp. 31–52.

[Su2] ———, *Single-input observability of continuous-time systems*, Math. Systems Theory, 12 (1979), pp. 371–393.

[Su3] ———, *Lie brackets and local controllability: a sufficient condition for scalar-input systems*, SIAM J. Control Optim., 21 (1983), pp. 686–713.

[Su4] ———, *A general theorem on local controllability*, SIAM J. Control Optim., 25 (1987), pp. 158–194.

[SJ] H. J. SUSSMANN AND V. JURDJEVIC, *Controllability of nonlinear systems*, J. Differential Equations, 12 (1972), pp. 95–116.

[SM] L. M. SILVERMAN AND H. E. MEADOWS, *Controllability and observability in time variable linear systems*, SIAM J. Control, 5 (1967), pp. 64–73.

[SS] E. D. SONTAG AND H. J. SUSSMANN, *Remarks on continuous feedback*, in Proc. Conference on Decision and Control, Vol. 2, Albuquerque, NM, 1980, pp. 916–921.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 33, No. 3, pp. 834–862, May 1995

© 1995 Society for Industrial and Applied Mathematics
009

SAMPLED-DATA AND DISCRETE-TIME H_2 OPTIMAL CONTROL*

H. L. TRENTELMAN† AND A. A. STOORVOGEL‡

Abstract. This paper deals with the sampled-data H_2 optimal control problem. Given a linear time-invariant continuous-time system, the problem of minimizing the H_2 performance over all sampled-data controllers with a fixed sampling period can be reduced to a pure discrete-time H_2 optimal control problem. This discrete-time H_2 problem is always singular. Motivated by this, in this paper we give a treatment of the discrete-time H_2 optimal control problem in its full generality. The results we obtain are then applied to the singular discrete-time H_2 problem arising from the sampled-data H_2 problem. In particular, we give conditions for the existence of optimal sampled data controllers. We also show that the H_2 performance of a continuous-time controller can always be recovered asymptotically by choosing the sampling period sufficiently small. Finally, we show that the optimal sampled-data H_2 performance converges to the continuous-time optimal H_2 performance as the sampling period converges to zero.

Key words. sampled-data, lifting technique, discrete-time, H_2 optimal control, algebraic Riccati equation, small sampling periods

AMS subject classifications. 93C05, 93C35, 93C60

1. Introduction.

Recently, much attention has been paid to H_2 and H_∞ optimal control of linear systems using sampled-data control (see [6], [7], [12], [2], [4] and [5], [11], [10], [1], [3], [17], [21]). For a given a continuous-time plant, a sampled-data controller consists of the cascade connection of an A/D converter, a discrete-time controller, and a D/A converter. The A/D device converts the continuous-time measured plant output into a discrete-time signal, which is used as an input for the discrete-time controller. The discrete-time controller generates a discrete-time output signal, which, in turn, is converted into a continuous-time signal that is used as a control input for the continuous-time plant.

Apart from a control input and a measurement output, the plant under consideration has an exogenous input and an output to be controlled. The quality of a controller is given by the performance of the corresponding closed-loop system. This performance measures the influence of the exogenous input on the output to be controlled. In the present paper, we will take the H_2 performance of the closed-loop system as performance measure.

In contrast to the H_∞ performance of a sampled-data control system, which in analogy with the pure continuous-time context can simply be defined as the norm of the input/output operator between the exogenous inputs and the outputs to be controlled, it is not clear from the outset how one should define the H_2 performance of a sampled-data control system. One definition was proposed in [6]: the H_2 performance of the closed-loop system is the number obtained by applying at each input channel a Dirac distribution and by taking the sum of integral squares of the resulting outputs. Of course, this definition exactly mimics the one that is common in the pure continuous-time context.

* Received by the editors December 28, 1992; accepted for publication (in revised form) November 24, 1993.

† Mathematics Institute, University of Groningen, P.O. Box 800, 9700 AV Groningen, The Netherlands (h.l.trentelman@math.rug.nl).

‡ Faculty of Mathematics and Computing Science, Eindhoven University of Technology, P.O. Box 513, 5600 MB Eindhoven, The Netherlands (wscoas@win.tue.nl). This research was supported by a fellowship of the Royal Netherlands Academy of Sciences and Arts.

In our opinion, a more natural definition was given independently in [12] and [2]. In these references, the crucial observation is that the closed-loop system resulting from a sampled data controller, albeit time-varying, is in fact a periodic system, with period equal to the sampling period. It is then argued that, instead of applying impulsive inputs at time $t = 0$, one should in fact apply these inputs at all time instances between 0 and the sampling period and take the mean of the integral squares of the resulting outputs. This leads to an H_2 performance measure that captures the essential features of a sampled-data closed-loop system more satisfactorily. For a given continuous-time plant, the sampled-data H_2 optimal control problem is then to minimize the H_2 performance of the closed-loop system over all internally stabilizing sampled-data controllers with a fixed sampling period. It is the latter problem that will be studied in this paper.

It was shown in [12] and [2] (see also [4]) that the sampled-data H_2 optimal control problem can be reduced to a pure discrete-time H_2 optimal control problem in the following way. First one defines an auxiliary time-invariant discrete-time system (involving the parameters of the original continuous-time plant and the given sampling period). Next, one expresses the sampled-data H_2 performance in terms of the 'normal' H_2 performance of the closed-loop system obtained by interconnecting the auxiliary discrete-time system and the discrete-time controller defining the sampled-data controller. Thus, the sampled-data H_2 optimal control problem under consideration is completely resolved once the auxiliary discrete-time H_2 problem is. This procedure makes use of the so-called *lifting technique* (see [20], [1], [3])

Now it turns out that the auxiliary discrete-time H_2 problem obtained in this way *is always a singular problem*: the direct feedthrough matrix from the exogenous input to the measurement output is always equal to 0. Apart from this, in the auxiliary discrete-time system the direct feedthrough matrix from the control input to the output to be controlled is in general not injective. (Note that, in general, an H_2 optimal control problem is called *regular* if the direct feedthrough matrix from the control input to the output to be controlled is injective, and the direct feedthrough matrix from the exogenous input to the measurement output is surjective. If the problem is not regular it is called *singular*.) In [12], this difficulty is partly removed by introducing an additional noise on the sampled measured output signal and by assuming the corresponding feedthrough matrix to be surjective.

In the present paper we want to consider the completely general formulation of the sampled-data H_2 problem. As a starting point we will take the auxiliary discrete-time H_2 problem derived in [12] and [2]. As noted, this problem is inherently singular. To our best knowledge, no resolution of the discrete-time singular H_2 optimal is known in the literature. Therefore, a substantial part of this paper is devoted to a study of the completely general discrete-time H_2 problem (no assumptions on the direct feedthrough matrices, no assumptions on the absence of zeros on the unit circle). We will describe a complete resolution to this problem, including a characterization of the optimal performance, and necessary *and* sufficient conditions for the existence of optimal controllers. The expression for the optimal performance is different from the one that might be expected in analogy with the continuous-time case (see [15]). Due to the fact that the role of the imaginary axis is taken over by the unit circle, for the discrete-time H_2 performance to be finite it is no longer required that the closed-loop transfer matrix is *strictly proper*. Intuitively, this enlarges the class of admissible controllers and yields a smaller optimal performance.

We will apply our results on the discrete-time H_2 optimal control problem to the sampled-data H_2 problem by simply applying them to the auxiliary discrete-time system derived in [12] and [2]. Our expression for the optimal sampled-data H_2 performance will be an immediate consequence of these results. We will, however, also be interested in conditions guaranteeing the existence of optimal sampled-data controllers. Our results on the general discrete-time H_2 problem give such conditions in terms of the auxiliary discrete-time system, but we will reformulate these conditions *in terms of the original continuous-time plant.* Preliminary results in that direction were also found in [12].

Obviously, the sampled-data H_2 optimal performance is a function of the sampling period. An important question is: what happens if the sampling period tends to zero. In particular, we will answer the following two questions. First, if we control the original continuous-time plant by a "normal" continuous-time compensator, is it then possible to recover this performance asymptotically by using a sampled-data controller with sufficiently small sampling period? This question was also studied for the H_∞ performance and for the H_2 performance à la Chen and Francis in [6]. A second, related, question that we will answer is: does the optimal sampled-data H_2 performance converge to the optimal continuous-time H_2 performance as the sampling period decreases to zero?

The outline of this paper is as follows. In §2 we will define the sampled-data H_2 optimal control problem and recall the main results of [12] and [2]. We will also introduce some notation and recall the notions of left-invertibility and right-invertibility of linear systems, zeros, and their most important state space interpretations. In §3 we deal with the discrete-time H_2 optimal control problem. In this section we will not yet treat the completely general case but make some assumptions on the absence of zeros on the unit circle. In §4, the results of §3 will be extended to derive a resolution of the general discrete-time H_2 optimal control problem. Then, in §5, we return to the sampled-data context and apply the results of §§3 and 4 to the sampled-data H_2 optimal control problem. In particular, we will derive conditions in terms of the original continuous-time plant that guarantee the existence of optimal controllers for the sampled-data H_2 problem. Finally, in §6 we study the aforementioned questions regarding the behavior of the (optimal) performance as the sampling period tends to zero.

2. Problem formulation. Consider a continuous-time, linear, time-invariant, finite-dimensional plant Σ. Let Σ have inputs d and u and outputs z and y, where d is an exogenous input, u is a control input, z is an output to be controlled, and y is a measured output. We want to control Σ by means of sampled-data feedback control. We take a fixed $\Delta > 0$, called the *sampling period.* From the measured output y we obtain a discrete-time signal $\bar{y} = \{y_k\}$ defined by $y_k := (S_\Delta y)_k$, where S_Δ denotes the sampling operator defined by $(S_\Delta y)_k := y(k\Delta)$. This discrete-time signal is taken as input for a discrete-time, linear, time-invariant, finite-dimensional compensator Γ_{dis}. The latter compensator generates a discrete-time signal $\bar{u} = \{u_k\}$, which, in turn, yields a (piecewise constant) continuous-time input signal u for the plant by defining $u(t) := (H_\Delta \bar{u})(t)$, where H_Δ is the hold operator defined by $(H_\Delta \bar{u})(t) := u_k$ $(t \in [k\Delta, (k+1)\Delta))$. This type of feedback control is depicted in Fig. 1.

If we control the system Σ by means of a sampled-data controller with sampling period Δ, then the resulting closed-loop system will no longer be time-invariant. In [12] and [2] the following definition of H_2 performance in the context of sampled-data control is proposed. First, it is observed that the closed-loop system resulting from a sampled-data controller with sampling period Δ is always a time-varying, Δ-periodic

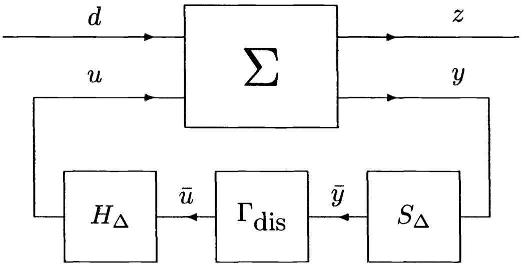


FIG. 1.

system. Then, for Δ-periodic systems the notion of H_2 performance is defined as follows. Suppose we have a finite-dimensional, time-varying, Δ-periodic system $\Sigma_{\rm per}$ described by

$$z(t) = \int_0^t G(t,s)d(s)ds. \tag{2.1}$$

It is argued in [12] and [2] that a natural way to define the H_2 performance of (2.1) is

$$\|\Sigma_{\rm per}\|_2^2 := \frac{1}{\Delta}\int_0^\Delta \mathrm{tr} \int_s^\infty G^{\rm T}(t,s)G(t,s)dt\,ds. \tag{2.2}$$

Next, if Γ is a sampled-data controller with sampling period Δ, the associated performance is defined as $J_{\Sigma,\Delta}(\Gamma) := \|\Sigma \times \Gamma\|_2^2$, the H_2 performance of the (Δ-periodic) closed-loop system $\Sigma \times \Gamma$. The sampled-data H_2 problem is then to minimize, for a fixed sampling period Δ, the performance criterion $J_{\Sigma,\Delta}(\Gamma)$ over all internally stabilizing sampled-data controllers Γ with sampling period Δ. It was shown in [12] and [2] that this problem can be reduced to a discrete-time 'normal' H_2 optimal control problem. To be specific, let the plant Σ be given by the equations

$$\begin{aligned} \dot{x}(t) &= Ax(t) + Bu(t) + Ed(t)\ , \\ y(t) &= C_1x(t)\ , \\ z(t) &= C_2x(t) + D_2u(t)\ , \end{aligned} \tag{2.3}$$

with $x(t) \in \mathbb{R}^n$, $u(t) \in \mathbb{R}^m$, $d(t) \in \mathbb{R}^r$, $y(t) \in \mathbb{R}^p$, and $z(t) \in \mathbb{R}^q$. It will be a standing assumption in this paper that (A, B) is stabilizable and that (C_1, A) is detectable, both with respect to $\mathcal{C}^- := \{s \in \mathcal{C} \mid \Re e\ s < 0\}$. Introduce a finite-dimensional linear time-invariant discrete-time system Σ_Δ:

$$\begin{aligned} x_{k+1} &= A_\Delta x_k + B_\Delta u_k + E_\Delta d_k\ , \\ y_k &= C_1 x_k\ , \\ z_k &= C_{2,\Delta}x_k + D_{2,\Delta}u_k\ , \end{aligned} \tag{2.4}$$

where we define

$$A_\Delta := e^{\Delta A}, \quad B_\Delta := \int_0^\Delta e^{tA}dtB,$$

where E_Δ is any matrix satisfying

$$E_\Delta E_\Delta^{\mathrm{T}} = \int_0^\Delta e^{tA} E E^{\mathrm{T}} e^{tA^{\mathrm{T}}}\, dt, \tag{2.5}$$

and where $C_{2,\Delta}$ and $D_{2,\Delta}$ are matrices satisfying

$$(C_{2,\Delta} \quad D_{2,\Delta})^{\mathrm{T}} (C_{2,\Delta} \quad D_{2,\Delta}) = \int_0^\Delta e^{t\underline{A}^{\mathrm{T}}} (C_2 \quad D_2)^{\mathrm{T}} (C_2 \quad D_2)\, e^{t\underline{A}}\, dt. \tag{2.6}$$

Here we have denoted

$$\underline{A} := \begin{pmatrix} A & B \\ 0 & 0 \end{pmatrix}. \tag{2.7}$$

Let $\boldsymbol{\Delta}$ denote the set of sampling periods for which either (A_Δ, B_Δ) is not stabilizable or (C_1, A_Δ) is not detectable, both with respect to the open unit disc $\{z \in \mathcal{C} \mid |z| < 1\}$. It is well known [13], [8] that if (A, B) is stabilizable and (C_1, A) is detectable, then every bounded subset of $\mathbb{R}^+$ contains only finitely many elements of $\boldsymbol{\Delta}$. We will restrict ourselves to sampling periods that are not in $\boldsymbol{\Delta}$. The plant Σ is controlled using sampled-data controllers $\Gamma := H_\Delta \Gamma_{\mathrm{dis}} S_\Delta$, with Γ_{dis} given by the equations

$$\begin{aligned} w_{k+1} &= K w_k + L y_k\ , \\ u_k &= M w_k + N y_k\ . \end{aligned} \tag{2.8}$$

Let us denote by $J_{\Sigma_\Delta}(\Gamma_{\mathrm{dis}})$ the discrete-time H_2 performance of the closed-loop system $\Sigma_\Delta \times \Gamma_{\mathrm{dis}}$, i.e., the value $\sum_k \mathrm{tr}\ (G_k G_k^{\mathrm{T}})$, where $\{G_k\}$ denotes the pulse response of the closed-loop system. The main result of [12] and [2] is the following:

THEOREM 2.1. *Assume that $\Delta \notin \boldsymbol{\Delta}$. Then there exists a sampled-data controller Γ with sampling period Δ such that the closed-loop system $\Sigma \times \Gamma$ is internally stable. The sampled-data controller $\Gamma = H_\Delta \Gamma_{\mathrm{dis}} S_\Delta$ internally stabilizes Σ if and only if the discrete-time controller Γ_{dis} internally stabilizes Σ_Δ. Furthermore, for every such controller we have*

$$J_{\Sigma,\Delta}(\Gamma) = \frac{1}{\Delta} \int_0^\Delta \int_0^{\Delta - s} \mathrm{tr}\ \left(C_2 e^{tA} E E^{\mathrm{T}} e^{tA^{\mathrm{T}}} C_2^{\mathrm{T}} \right) dt\, ds + \frac{1}{\Delta} J_{\Sigma_\Delta}(\Gamma_{\mathrm{dis}}).$$

We shall use this theorem as a starting point and study in this paper the discrete-time H_2 optimal control problem for the discrete-time system Σ_Δ given by (2.4). This H_2 problem is inherently singular, due to the fact that the direct feedthrough matrix from the disturbance input to the measured output is always equal to zero.

We conclude this section by introducing some notation and recalling some basic concepts. In this paper, any given continuous-time system $\dot{x} = Ax + Bu, y = Cx + Du$ or discrete-time system $x_{k+1} = Ax_k + Bu_k, y_k = Cx_k + Du_k$ will be denoted simply by (A, B, C, D). It will be clear from the context which interpretation we have in mind. For any such system, the *system matrix* is defined as the first-order polynomial matrix

$$P(s) = \begin{pmatrix} sI - A & -B \\ C & D \end{pmatrix}.$$

If the underlying system is discrete-time, we will rather use the indeterminate z instead of s. For a real rational matrix R, its *normal rank*, normrank R, is defined as the rank of R as a matrix with entries in the field of real rational functions. It is well known that normrank $R = \max_\sigma$ rank $R(\sigma)$. A *zero* of the system (A, B, C, D) is any complex number λ with the property that rank $P(\lambda) <$ normrank P. The system (A, B, C, D) is called left-invertible (right-invertible) if its transfer matrix $G(s) = C(sI - A)^{-1}B + D$ is a left-invertible (right-invertible) rational matrix. Assuming that $A \in \mathbb{R}^{n\times n}$, $B \in \mathbb{R}^{m\times n}$, and $C \in \mathbb{R}^{p\times n}$ we have that (A, B, C, D) is left-invertible (right-invertible) if and only if its system matrix has normal rank $n + m$ $(n + p)$.

If $M \in \mathbb{R}^{n\times n}$ and $\mathcal{L}$ is a subspace of $\mathbb{R}^n$, then $\langle M \mid \mathcal{L}\rangle$ will denote the smallest M-invariant subspace containing $\mathcal{L}$. The largest M-invariant subspace contained in $\mathcal{L}$ will be denoted by $\langle \mathcal{L} \mid M\rangle$. In particular, given (A, B, C, D), the reachable subspace is equal to $\langle A \mid \operatorname{im} B\rangle$ and the unobservable subspace is equal to $\langle \ker C \mid A\rangle$.

Given the system (A, B, C, D), we define *the weakly unobservable subspace* $\mathcal{V}$ to be the smallest subspace $\mathcal{L}$ of $\mathbb{R}^n$ with the property that there exists $F \in \mathbb{R}^{m\times n}$ such that $(A + BF)\mathcal{L} \subset \mathcal{L}$ and $(C + DF)\mathcal{L} = 0$ (see [14]). In addition, *the controllability subspace* $\mathcal{R}$ of (A, B, C, D) is defined as follows:

$$\mathcal{R} := \langle A + BF \mid \mathcal{V} \cap B \ker D\rangle,$$

for any F such that $(A + BF)\mathcal{V} \subset \mathcal{V}$ and $(C + DF)\mathcal{V} = 0$ (any such F yields the same $\mathcal{R}$). It was shown in [14] that the system (A, B, C, D) is left-invertible if and only if $\ker B \cap \ker D = 0$ and $\mathcal{V} \cap B \ker D = 0$. Note that $\mathcal{V} \cap B \ker D = 0$ if and only if $\mathcal{R} = 0$.

Finally, the set of zeros of (A, B, C, D) can be shown to be equal to $\sigma(A + BF \mid \mathcal{V}/\mathcal{R})$, for any F such that $(A + BF)\mathcal{V} \subset \mathcal{V}$ and $(C + DF)\mathcal{V} = 0$. Here, $A + BF \mid \mathcal{V}/\mathcal{R}$ is the quotient map of $A + BF \mid \mathcal{V}$ modulo $\mathcal{R}$ (see, e.g., [19]).

3. The discrete-time H_2 problem: No zeros on the unit circle. In this section we shall consider the discrete-time H_2 problem. Consider the finite-dimensional, linear, time-invariant, discrete-time system Σ_{dis} given by the equations

$$\begin{aligned} x_{k+1} &= Ax_k + Bu_k + Ed_k\,, \\ y_k &= C_1x_k \qquad\quad\; + D_1d_k\,, \\ z_k &= C_2x_k + D_2u_k\,. \end{aligned} \tag{3.1}$$

There will be no assumptions on the direct feedthrough matrices D_1 and D_2. In the present section, however, we will have assumptions on the absence of system zeros on the unit circle in the complex plane: it will be assumed that (A, B, C_2, D_2) and (A, E, C_1, D_1) do not have zeros on the unit circle $|z| = 1$. In the next section we will drop these assumptions and treat the completely general case. Of course, it will be a standing assumption that (A, B) is stabilizable and that (C_1, A) is detectable, both with respect to the open unit disc.

We will consider discrete-time controllers Γ_{dis} given by (2.8). For any internally stabilizing controller Γ_{dis}, let $J_{\Sigma_{\text{dis}}}(\Gamma_{\text{dis}})$ be its H_2 performance. Denote by J^* the optimal performance, i.e., the infimum over all internally stabilizing controllers Γ_{dis}.

For a given matrix M, we will denote by M^+ its Moore–Penrose inverse. The solution of the discrete-time H_2 optimal control problem centers around the following two algebraic Riccati equations:

$$P = A^{\mathrm{T}}PA + C_2^{\mathrm{T}}C_2 - (C_2^{\mathrm{T}}D_2 + A^{\mathrm{T}}PB)(D_2^{\mathrm{T}}D_2 + B^{\mathrm{T}}PB)^+(D_2^{\mathrm{T}}C_2 + B^{\mathrm{T}}PA), \tag{3.2}$$

$$Q = AQA^{\mathrm{T}} + EE^{\mathrm{T}} - (AQC_1^{\mathrm{T}} + ED_1^{\mathrm{T}})(D_1D_1^{\mathrm{T}} + C_1QC_1^{\mathrm{T}})^+(D_1E^{\mathrm{T}} + C_1QA^{\mathrm{T}}). \tag{3.3}$$

For any real symmetric matrix P, we shall denote

$$D_P := (D_2^{\mathrm{T}} D_2 + B^{\mathrm{T}} P B)^{\frac{1}{2}}, \tag{3.4}$$

$$C_P := D_P^{+}(D_2^{\mathrm{T}} C_2 + B^{\mathrm{T}} P A). \tag{3.5}$$

Note that, since for any matrix $M \geq 0$ we have $(M^{\frac{1}{2}})^{+} = (M^{+})^{\frac{1}{2}}$, we have $D_P^{+} C_P = (D_2^{\mathrm{T}} D_2 + B^{\mathrm{T}} P B)^{+}(D_2^{\mathrm{T}} C_2 + B^{\mathrm{T}} P A)$. If, in addition, P is a real symmetric solution of (3.2), then $C_P^{\mathrm{T}} C_P = A^{\mathrm{T}} P A - P + C_2^{\mathrm{T}} C_2$. Note also that D_P is symmetric by definition. Finally, since im $(D_2^{\mathrm{T}} C_2 + B^{\mathrm{T}} P A) \subset$ im D_P, we have $D_P C_P = D_2^{\mathrm{T}} C_2 + B^{\mathrm{T}} P A$. (Note that it is a property of the Moore–Penrose inverse that MM^{+} is the orthogonal projection onto im M.)

The following is a corrected and slightly extended version of a theorem from [14]. A proof can be given along the lines of the proof of [14, Thm. 18].

THEOREM 3.1. *Consider the system (A, B, C_2, D_2) together with the algebraic Riccati equation* (3.2). *The following two statements are equivalent* :

(i) *(A, B) is stabilizable and (A, B, C_2, D_2) has no zeros on the unit circle $|z| = 1$,*

(ii) *Equation* (3.2) *has a real symmetric solution P with the following property: there exists a matrix F_1 such that*

$$|\sigma(A - BD_P^{+} C_P + B(I - D_P^{+} D_P) F_1)| < 1. \tag{3.6}$$

Furthermore, if P satisfies this condition, it is the unique real symmetric solution of (3.2) *for which this condition holds. In addition, P is positive semidefinite and is in fact the largest real symmetric solution of* (3.2).

Next we consider the dual algebraic Riccati equation (3.3). For any real symmetric matrix Q, denote

$$D_Q := (D_1 D_1^{\mathrm{T}} + C_1 Q C_1^{\mathrm{T}})^{\frac{1}{2}}, \tag{3.7}$$

$$E_Q := (AQC_1^{\mathrm{T}} + ED_1^{\mathrm{T}}) D_Q^{+}. \tag{3.8}$$

By dualizing the previous theorem, the corresponding result on the Riccati equation (3.3) can be found:

THEOREM 3.2. *Consider the system (A, E, C_1, D_1) together with the algebraic Riccati equation* (3.3). *The following two statements are equivalent* :

(i) *(C_1, A) is detectable and (A, E, C_1, D_1) has no zeros on the unit circle $|z| = 1$.*

(ii) *Equation* (3.3) *has a real symmetric solution Q with the following property: there exists a matrix K_1 such that*

$$|\sigma(A - E_Q D_Q^{+} C_1 + K_1(I - D_Q D_Q^{+}) C_1)| < 1. \tag{3.9}$$

Furthermore, if Q satisfies this condition, it is the unique real symmetric solution of (3.3) *for which this condition holds. In addition, Q is positive semidefinite and is in fact the largest real symmetric solution of* (3.3).

In the remainder of this section we will always denote by P and Q the largest real symmetric solution of (3.2) and (3.3), respectively. Now we will state the main result of this section:

THEOREM 3.3. *Consider the system* (3.1). *Assume that (A, B) is stabilizable and (C_1, A) is detectable. Assume that (A, B, C_2, D_2) and (A, E, C_1, D_1) have no zeros on the unit circle. Then we have the following:*

(i)

$$J^* = \mathrm{tr}\,(E^{\mathrm{T}}PE) + \mathrm{tr}\,(C_PQC_P^{\mathrm{T}}) - \mathrm{tr}\,((D_PN^*D_Q)(D_PN^*D_Q)^{\mathrm{T}}), \tag{3.10}$$

where N^* is defined by

$$N^* := -(D_P^+)^2(D_PC_PQC_1^{\mathrm{T}} + B^{\mathrm{T}}PED_1^{\mathrm{T}})(D_Q^+)^2. \tag{3.11}$$

(ii) *There exists an optimal controller, i.e., an internally stabilizing controller Γ^*_{dis} such that $J_{\Sigma_{\mathrm{dis}}}(\Gamma^*_{\mathrm{dis}}) = J^*$. One such optimal controller is given by the following "construction":*

(a) *Choose a state feedback matrix F such that $|\sigma(A+BF)| < 1$ and $C_P + D_PF = 0$.*

(b) *Choose an output injection matrix G such that $|\sigma(A+GC_1)| < 1$ and $E_Q + GD_Q = 0$.*

(c) *Define $\Gamma^*_{\mathrm{dis}} = (K^*, L^*, M^*, N^*)$ by choosing N^* given by (3.11), and by choosing $K^* := A + BF + GC_1 - BN^*C_1$, $L^* := BN^* - G$, and $M^* := F - N^*C_1$.*

In the remainder of this section we shall prove this theorem. In addition to the system Σ_{dis}, consider the system $\Sigma_{\mathrm{dis},P}$ given by the equations

$$\begin{aligned} x_{k+1} &= Ax_k + Bu_k + Ed_k\,, \\ y_k &= C_1x_k + D_1d_k, \\ z_k &= C_Px_k + D_Pu_k\,, \end{aligned} \tag{3.12}$$

with P the largest real symmetric solution of the algebraic Riccati equation (3.2). The following basic lemma can be proven by a standard completion-of-the-squares argument:

LEMMA 3.4. *For every compensator $\Gamma_{\mathrm{dis}} = (K, L, M, N)$ we have Γ_{dis} internally stabilizes Σ_{dis} if and only if Γ_{dis} internally stabilizes $\Sigma_{\mathrm{dis},P}$. For any such compensator we have*

$$J_{\Sigma_{\mathrm{dis}}}(\Gamma_{\mathrm{dis}}) = \mathrm{tr}\,(E^{\mathrm{T}}PE) + 2\mathrm{tr}\,(D_1^{\mathrm{T}}N^{\mathrm{T}}B^{\mathrm{T}}PE) + J_{\Sigma_{\mathrm{dis},P}}(\Gamma_{\mathrm{dis}}). \tag{3.13}$$

In addition to $\Sigma_{\mathrm{dis},P}$ we consider the system $\Sigma_{\mathrm{dis},P,Q}$ defined by

$$\begin{aligned} x_{k+1} &= Ax_k + Bu_k + E_Qd_k\,, \\ y_k &= C_1x_k + D_Qd_k, \\ z_k &= C_Px_k + D_Pu_k\,, \end{aligned} \tag{3.14}$$

with Q the largest real symmetric solution of the dual algebraic Riccati equation (3.3). It is clear that the H_2 performance of a given compensator Γ_{dis} applied to Σ_{dis} is equal to the H_2 performance of the dual compensator $\Gamma^{\mathrm{T}}_{\mathrm{dis}} := (K^{\mathrm{T}}, M^{\mathrm{T}}, L^{\mathrm{T}}, N^{\mathrm{T}})$ applied to the dual system $\Sigma^{\mathrm{T}}_{\mathrm{dis}}$. By applying Lemma 3.4 to the dual system $\Sigma^{\mathrm{T}}_{\mathrm{dis},P}$ and the dual compensator $\Gamma^{\mathrm{T}}_{\mathrm{dis}}$ we thus arrive at the following theorem:

THEOREM 3.5. *For every compensator $\Gamma_{\mathrm{dis}} = (K, L, M, N)$ we have : Γ_{dis} internally stabilizes Σ_{dis} if and only if Γ_{dis} internally stabilizes $\Sigma_{\mathrm{dis},P,Q}$. For any such compensator we have*

$$\begin{aligned} J_{\Sigma_{\mathrm{dis}}}(\Gamma_{\mathrm{dis}}) = \mathrm{tr}\,(E^{\mathrm{T}}PE) + \mathrm{tr}\,(C_PQC_P^{\mathrm{T}}) + 2\mathrm{tr}\,(D_1^{\mathrm{T}}N^{\mathrm{T}}B^{\mathrm{T}}PE) \\ + 2\mathrm{tr}\,(C_PQC_1^{\mathrm{T}}N^{\mathrm{T}}D_P^{\mathrm{T}}) + J_{\Sigma_{\mathrm{dis},P,Q}}(\Gamma_{\mathrm{dis}}). \end{aligned}$$

Now note that in the above formula the first two terms do not depend on the compensator Γ_{dis}. The remaining three terms do depend on the compensator. Also note that in the closed-loop system $\Sigma_{\text{dis},P,Q} \times \Gamma_{\text{dis}}$ the direct feedthrough matrix from the disturbance input to the output to be controlled is equal to $D_P N D_Q$. As a consequence, $J_{\Sigma_{\text{dis},P,Q}}(\Gamma_{\text{dis}}) \geq \text{tr }((D_P N D_Q)(D_P N D_Q)^{\mathrm{T}})$, with equality if and only if the transfer matrix $G_{P,Q,\Gamma_{\text{dis}}}(z)$ of the closed-loop system $\Sigma_{\text{dis},P,Q} \times \Gamma_{\text{dis}}$ is equal to the constant matrix $D_P N D_Q$. It thus follows immediately from Theorem 3.5 that

LEMMA 3.6. *For every internally stabilizing compensator* $\Gamma_{\text{dis}} = (K, L, M, N)$ *we have*

$$J_{\Sigma_{\text{dis}}}(\Gamma_{\text{dis}}) \geq \text{tr }(E^{\mathrm{T}}PE) + \text{tr }(C_P Q C_P^{\mathrm{T}}) + 2\text{tr }(D_1^{\mathrm{T}} N^{\mathrm{T}} B^{\mathrm{T}} PE)$$

$$+ 2\text{tr }(C_P Q C_1^{\mathrm{T}} N^{\mathrm{T}} D_P^{\mathrm{T}}) + \text{tr }((D_P N D_Q)(D_P N D_Q)^{\mathrm{T}}),$$

with equality if and only if $G_{P,Q,\Gamma_{\text{dis}}}(z) = D_P N D_Q$.

This lemma shows that, in order to minimize $J_{\Sigma_{\text{dis}}}(\Gamma_{\text{dis}})$ over all internally stabilizing compensators, we should do the following:

(i) First minimize the quadratic matrix function

$$\Phi(N) := 2\text{tr }(D_1^{\mathrm{T}} N^{\mathrm{T}} B^{\mathrm{T}} PE) + 2\text{tr }(C_P Q C_1^{\mathrm{T}} N^{\mathrm{T}} D_P^{\mathrm{T}}) \tag{3.15}$$

$$+ \text{tr }((D_P N D_Q)(D_P N D_Q)^{\mathrm{T}}),$$

yielding an optimal N^*.

(ii) Next find a compensator Γ_{dis}^*, described by the quadruple (K^*, L^*, M^*, N^*), that is internally stabilizing and yields $G_{P,Q,\Gamma_{\text{dis}}^*}(z) = D_P N^* D_Q$, i.e., the closed-loop system $\Sigma_{\text{dis},P,Q} \times \Gamma_{\text{dis}}^*$ has the constant transfer matrix $D_P N^* D_Q$.

Indeed, if N^* minimizes $\Phi(N)$ and if $G_{P,Q,\Gamma_{\text{dis}}^*}(z) = D_P N^* D_Q$, then we have

$$J_{\Sigma_{\text{dis}}}(\Gamma_{\text{dis}}^*) = \text{tr }(E^{\mathrm{T}}PE) + \text{tr }(C_P Q C_P^{\mathrm{T}}) + \Phi(N^*),$$

while for any internally stabilizing compensator $\Gamma_{\text{dis}} = (K, L, M, N)$ we have

$$J_{\Sigma_{\text{dis}}}(\Gamma_{\text{dis}}) \geq \text{tr }(E^{\mathrm{T}}PE) + \text{tr }(C_P Q C_P^{\mathrm{T}}) + \Phi(N) \geq \text{tr }(E^{\mathrm{T}}PE) + \text{tr }(C_P Q C_P^{\mathrm{T}}) + \Phi(N^*).$$

This clearly implies that

$$J^* = \text{tr }(E^{\mathrm{T}}PE) + \text{tr }(C_P Q C_P^{\mathrm{T}}) + \Phi(N^*)$$

and that

$$J_{\Sigma_{\text{dis}}}(\Gamma_{\text{dis}}^*) = J^*.$$

We will first study the minimization of $\Phi(N)$.

LEMMA 3.7. *Let* $\Phi(N)$ *be defined by* (3.15). *Define*

$$R^* := D_P^+(D_P C_P Q C_1^{\mathrm{T}} + B^{\mathrm{T}} PED_1^{\mathrm{T}})D_Q^+.$$

Then

$$\Phi^* := \min\{\Phi(N) \mid N \in \mathbb{R}^{m \times p}\} = -\text{tr }(R^* R^{*T}).$$

N minimizes Φ, i.e., $\Phi(N) = \Phi^*$, if and only if N is a solution to the linear equation $D_P N D_Q = -R^*$. One particular solution of this linear equation is given by $N^* = -D_P^+ R^* D_Q^+$. We have $\Phi^* = -\text{tr }((D_P N^* D_Q)(D_P N^* D_Q)^{\mathrm{T}})$.

Proof. Using the facts that

$$\ker D_Q \subset \ker (D_P C_P Q C_1^{\mathrm{T}} + B^{\mathrm{T}} P E D_1^{\mathrm{T}}),$$
$$\mathrm{im}\, D_P \supset \mathrm{im}\, (D_P C_P Q C_1^{\mathrm{T}} + B^{\mathrm{T}} P E D_1^{\mathrm{T}}),$$

it can be shown by straightforward calculation that

$$\Phi(N) = -\mathrm{tr}\, (R^* R^{*T}) + \mathrm{tr}\, ((D_P N D_Q + R^*)(D_P N D_Q + R^*)^{\mathrm{T}}).$$

The equation $D_P N D_Q = -R^*$ has a solution since $\ker D_Q = \ker D_Q^{\mathrm{T}} = \ker D_Q^+ \subset \ker R^*$ and $\mathrm{im}\, D_P = \mathrm{im}\, D_P^{\mathrm{T}} = \mathrm{im}\, D_P^+ \supset \mathrm{im}\, R^*$. Clearly, one particular solution is then given by $N^* = -D_P^+ R^* D_Q^+$. Finally, the expression for Φ^* can be checked in a straightforward manner. □

Next we study the question whether, starting with N^* above, it is possible to find K^*, L^*, M^* such that the resulting compensator $\Gamma^*_{\mathrm{dis}} = (K^*, L^*, M^*, N^*)$ yields a closed-loop system $\Sigma_{\mathrm{dis},P,Q} \times \Gamma^*_{\mathrm{dis}}$ with constant transfer matrix $D_P N^* D_Q$. We will first prove the following lemma:

LEMMA 3.8. *Assume that (A, B) is stabilizable and that (A, B, C_2, D_2) has no zeros on the unit circle. Let P be the largest real symmetric solution of the algebraic Riccati equation* (3.2). *There exists a matrix F such that*

(i) $|\sigma(A + BF)| < 1$,

(ii) $C_P + D_P F = 0$.

Proof. Let F_1 be such that (3.6) holds, and define $F := -D_P^+ C_P + (I - D_P^+ D_P) F_1$. Then (i) is satisfied. To prove (ii), note that $\mathrm{im}\, C_P \subset \mathrm{im}\, D_P^+ = \mathrm{im}\, D_P$. Consequently, $-D_P D_P^+ C_P = -C_P$, which proves (ii). □

We will also need the dual of this lemma, which reads as follows:

LEMMA 3.9. *Assume that (C_1, A) is detectable and that (A, E, C_1, D_1) has no zeros on the unit circle. Let Q be the largest real symmetric solution of the dual algebraic Riccati equation* (3.3). *There exists a matrix G such that*

(i) $|\sigma(A + GC_1)| < 1$,

(ii) $E_Q + G D_Q = 0$.

Now we show that by suitable choice of compensator Γ_{dis}, the transfer matrix of $\Sigma_{\mathrm{dis},P,Q} \times \Gamma_{\mathrm{dis}}$ can be made equal to any constant matrix product $M_1 M_2$, as long as $\mathrm{im}\, D_P \subset \mathrm{im}\, M_1$ and $\ker D_Q \subset \ker M_2$.

LEMMA 3.10. *Consider the system* (3.1). *Assume that (A, B) is stabilizable and (C_1, A) is detectable. Assume that (A, B, C_2, D_2) and (A, E, C_1, D_1) have no zeros on the unit circle. Let P and Q be the largest real symmetric solution of the algebraic Riccati equation* (3.2) *and* (3.3), *respectively. Then for any pair of matrices M_1, M_2 such that the product $M_1 M_2$ is defined and such that* $\mathrm{im}\, D_P \subset \mathrm{im}\, M_1$ *and* $\ker D_Q \subset \ker M_2$ *there exists an internally stabilizing compensator Γ_{dis} such that the transfer matrix of $\Sigma_{\mathrm{dis},P,Q} \times \Gamma_{\mathrm{dis}}$ is equal to the constant $M_1 M_2$.*

Specifically, for given M_1 and M_2 let F_2 be a solution of $M_1 = D_P F_2$ and G_2 be a solution of $M_2 = -G_2 D_Q$ and take F such that the conditions in Lemma 3.8 *are satisfied and G such that the conditions of Lemma* 3.9 *are satisfied. Then the compensator $\Gamma_{\mathrm{dis}} := (K, L, M, N)$ with $K := A + BF + GC_1 + BF_2 G_2 C_1$, $L := -BF_2 G_2 - G$, $M := F + F_2 G_2 C_1$, and $N := -F_2 G_2$ satisfies the requirements.*

Proof. The equations of the compensator are given by (2.8). Using the specifications of K, L, M, and N given above, we find that the error $e_k := w_k - x_k$ satisfies $e_{k+1} = (A + GC_1) e_k$. Thus, if $w_0 = 0$ and $x_0 = 0$, we have $x_k = w_k$ for all k. In particular, this implies that $u_k = F x_k + F_2 M_2 w_k$. The output of the closed-loop system is then equal to $z_k = C_P x_k + D_P u_k = M_1 M_2 w_k$. This implies that the closed-loop transfer matrix is equal to the constant matrix $M_1 M_2$. Finally, the spectrum of the closed-loop system matrix A_e is easily shown to be equal to $\sigma(A+BF) \cup \sigma(A+GC_1)$. This implies that the closed-loop system is internally stable. □

Clearly, if in this lemma we take $M_1 = D_P$ and $M_2 = N^* D_Q$, we arrive at an internally stabilizing compensator Γ_{dis} such that the closed-loop transfer matrix is equal to the constant matrix $D_P N^* D_Q$. In the formulas for the compensator as given in the lemma, we should then take $F_2 = I$ and $G_2 = -N^*$. The result of Theorem 3.3 follows immediately by combining the above lemmas.

Remark 3.11. For later use we note that Lemma 3.8 also provides a resolution of the discrete-time *linear quadratic problem* for the case that (A, B, C_2, D_2) has no zeros on the unit circle (see also [14]). Given $x_{k+1} = Ax_k + Bu_k$, the problem is to minimize the cost-functional $J(x_0, u) := \sum_k \|(C_2 x_k + D_2 u_k\|^2$ over all inputs $u = \{u_k\}$ such that $x_k \to 0$. It was pointed out in [14] that for each such input u we have the completion-of-the-squares formula $J(x_0, u) = x_0^{\mathrm{T}} P x_0 + J_P(x_0, u)$, with $J_P(x_0, u) := \sum_k \|C_P x_k + D_P u_k\|^2$. Thus, if we take F satisfying (i) and (ii) of Lemma 3.8, then the input $u_k = F x_k$ leads to the optimal cost $J^*(x_0) = x_0^{\mathrm{T}} P x_0$. Note that we could also formulate the linear quadratic problem as a minimization over all internally stabilizing feedback laws: minimize the cost-functional $J(x_0, F) := \sum_k \|(C_P + D_P F) x_k\|^2$ over all $F \in \mathbb{R}^{m \times n}$ such that $|\sigma(A + BF)| < 1$. By the above argument, any F satisfying (i) and (ii) of Lemma 3.8 is then optimal and the optimal cost is again given by $x_0^{\mathrm{T}} P x_0$.

Remark 3.12. An interesting question is under what conditions the Moore–Penrose inverse $(D_2^{\mathrm{T}} D_2 + B^{\mathrm{T}} P B)^+$ reduces to the inverse $(D_2^{\mathrm{T}} D_2 + B^{\mathrm{T}} P B)^{-1}$, equivalently, under what conditions $D_2^{\mathrm{T}} D_2 + B^{\mathrm{T}} P B$ is positive definite. Using the ideas from [14] it can be shown that if P is a positive semidefinite solution of the algebraic Riccati equation (3.2), then $D_2^{\mathrm{T}} D_2 + B^{\mathrm{T}} P B > 0$ if and only if (A, B, C_2, D_2) is a left-invertible system. Of course, dually, if Q is a positive semidefinite solution of the algebraic Riccati equation (3.3), then $D_1 D_1^{\mathrm{T}} + C_1 Q C_1^{\mathrm{T}} > 0$ if and only if the system (A, E, C_1, D_1) is right-invertible. In view of this, it is perhaps more natural to call the discrete-time H_2 problem regular if (A, B, C_2, D_2) is a left-invertible system and (A, E, C_1, D_1) is a right-invertible system.

4. The discrete-time H_2 problem: The general case. In this section we will extend the results of the previous section and treat the discrete-time H_2 problem in its full generality. This means that we will drop the assumption on the absence of zeros on the unit circle that was made in the previous section. First we will prove that also without the assumption that (A, B, C_2, D_2) has no zeros on the unit circle, the Riccati equation (3.2) has a largest real symmetric solution. We will prove that this solution can be obtained as the limit of solutions of algebraic Riccati equations associated with suitable perturbations of the system (A, B, C_2, D_2).

THEOREM 4.1. *If (A, B) is stabilizable, then the Riccati equation (3.2) has a largest real symmetric solution, say P. P is positive semidefinite. We have $P = \lim_{\varepsilon \downarrow 0} P_\varepsilon$, where for $\varepsilon > 0$ P_ε is the largest real symmetric solution of the algebraic Riccati equation*

$$\begin{aligned} &A^{\mathrm{T}} P_\varepsilon A - P_\varepsilon + C_2^{\mathrm{T}} C_2 + \varepsilon^2 I \\ &\quad -(A^{\mathrm{T}} P_\varepsilon B + C_2^{\mathrm{T}} D_2)(D_2^{\mathrm{T}} D_2 + B^{\mathrm{T}} P_\varepsilon B)^+ (B^{\mathrm{T}} P_\varepsilon A + D_2^{\mathrm{T}} C_2) = 0. \end{aligned} \tag{4.1}$$

Remark 4.2. Note that (4.1) is the Riccati equation associated with the perturbed system $(A, B, \binom{C_2}{\varepsilon I}, \binom{D_2}{0})$. (Here, I denotes the $n\times n$ identity matrix, and 0 denotes the $n \times m$ zero matrix). For $\varepsilon > 0$, the perturbed system has no zeros. Consequently, the existence of P_ε follows from Theorem 3.1.

The idea of the proof of Theorem 4.1 is to show first that the P_ε indeed converge to some matrix P and next to show that P satisfies (3.2). The difficulty is that in the general case we are considering, the term $D_2^{\mathrm{T}}D_2 + B^{\mathrm{T}}PB$ need not be invertible, so that we cannot conclude that $(D_2^{\mathrm{T}}D_2+B^{\mathrm{T}}P_\varepsilon B)^+$ converges to $(D_2^{\mathrm{T}}D_2+B^{\mathrm{T}}PB)^+$. We will show, however, that we can get around this difficulty by considering the so-called linear matrix inequality. Our proof is split up in three lemmas. In the following, let $J(x_0, u)$ be the cost-functional of the linear quadratic problem, and let $J^*(x_0)$ be the optimal cost (see Remark 3.11).

LEMMA 4.3. *Let P_ε be the largest real symmetric solution of* (4.1). *There exists a real positive semidefinite matrix P such that $P_\varepsilon \downarrow P$ ($\varepsilon \downarrow 0$). For all $x_0 \in \mathbb{R}^n$ we have $J^*(x_0) = x_0^{\mathrm{T}}Px_0$.*

Proof. Let $J_\varepsilon(x_0, u) := \sum_k \|C_P x_k + D_P u_k\|^2 + \varepsilon^2\|x_k\|^2$, and let $J_\varepsilon^*(x_0)$ be the infimum of $J_\varepsilon(x_0, u)$ over all u such that $x_k \to 0$. According to Remark 3.11 we have $J_\varepsilon^*(x_0) = x_0^{\mathrm{T}}P_\varepsilon x_0$. From this interpretation it follows that P_ε is monotonically nonincreasing as $\varepsilon \downarrow 0$. Being bounded from below by 0, this yields the existence of a limit P. Obviously, for all $\varepsilon > 0$ we have $J^*(x_0) \le J_\varepsilon^*(x_0) = x_0^{\mathrm{T}}P_\varepsilon x_0$, so $J^*(x_0) \le x_0^{\mathrm{T}}Px_0$. Conversely, for all $\varepsilon > 0$ and for all u we have $J_\varepsilon(x_0, u) \ge x_0^{\mathrm{T}}P_\varepsilon x_0$. Taking the limit on both sides this yields $J(x_0, u) \ge x_0^{\mathrm{T}}Px_0$ for all u. Taking the infimum over u then yields the converse inequality. □

LEMMA 4.4. *P is the largest real symmetric solution of the linear matrix inequality*

$$M(P) := \begin{pmatrix} A^{\mathrm{T}}PA - P + C_2^{\mathrm{T}}C_2 & C_2^{\mathrm{T}}D_2 + A^{\mathrm{T}}PB \\ D_2^{\mathrm{T}}C_2 + B^{\mathrm{T}}PA & D_2^{\mathrm{T}}D_2 + B^{\mathrm{T}}PB \end{pmatrix} \ge 0.$$

Proof. Denote the left-hand side of (4.1) by $R_\varepsilon(P_\varepsilon)$. Also consider the linear matrix inequality associated with the perturbed system:

$$M_\varepsilon(P_\varepsilon) := \begin{pmatrix} A^{\mathrm{T}}P_\varepsilon A - P_\varepsilon + C_2^{\mathrm{T}}C_2 + \varepsilon^2 I & C_2^{\mathrm{T}}D_2 + A^{\mathrm{T}}P_\varepsilon B \\ D_2^{\mathrm{T}}C_2 + B^{\mathrm{T}}P_\varepsilon A & D_2^{\mathrm{T}}D_2 + B^{\mathrm{T}}P_\varepsilon B \end{pmatrix} \ge 0.$$

We have $M_\varepsilon(P_\varepsilon) \ge 0$ if and only if $R_\varepsilon(P_\varepsilon) \ge 0$. This follows from the fact that the latter is equal to the Schur complement of $D_2^{\mathrm{T}}D_2 + B^{\mathrm{T}}P_\varepsilon B$ in $M_\varepsilon(P_\varepsilon)$. The Schur complement is defined here with matrix inverse replaced by Moore–Penrose inverse. This can be done because of the fact that

$$\ker(D_2^{\mathrm{T}}D_2 + B^{\mathrm{T}}P_\varepsilon B) \subset \ker(C_2^{\mathrm{T}}D_2 + A^{\mathrm{T}}P_\varepsilon B).$$

Since $R_\varepsilon(P_\varepsilon) = 0$, we indeed have $M_\varepsilon(P_\varepsilon) \ge 0$. Taking the limit $\varepsilon \downarrow 0$ then yields $M(P) \ge 0$. To show that P is the largest real symmetric solution, let P_1 be any real symmetric solution of the linear matrix inequality. Using a standard completion-of-the-squares argument then yields $J(x_0, u) \ge x_0^{\mathrm{T}}P_1x_0$ for any x_0 and any u such that $x_k \to 0$. Taking the infimum over all such u yields $x_0^{\mathrm{T}}Px_0 = J^*(x_0) \ge x_0^{\mathrm{T}}P_1x_0$. □

Now we will show that P in fact satisfies the algebraic Riccati equation (3.2). Denote

$$R(P) := A^{\mathrm{T}}PA - P + C_2^{\mathrm{T}}C_2 - (C_2^{\mathrm{T}}D_2 + A^{\mathrm{T}}PB)(D_2^{\mathrm{T}}D_2 + B^{\mathrm{T}}PB)^+(D_2^{\mathrm{T}}C_2 + B^{\mathrm{T}}P).$$

Again, by the fact that $\ker(D_2^{\mathrm{T}}D_2 + B^{\mathrm{T}}PB) \subset \ker(C_2^{\mathrm{T}}D_2 + A^{\mathrm{T}}PB)$, $R(P)$ is equal to the Schur complement of $D_2^{\mathrm{T}}D_2 + B^{\mathrm{T}}PB$ in $M(P)$. In particular this implies that

$$\operatorname{rank} M(P) = \operatorname{rank}\,(D_2^{\mathrm{T}}D_2 + B^{\mathrm{T}}PB) + \operatorname{rank} R(P).$$

In order to prove that $R(P) = 0$ we should therefore prove that P has the property expressed in the following lemma:

LEMMA 4.5. $\operatorname{rank} M(P) = \operatorname{rank}\,(D_2^{\mathrm{T}}D_2 + B^{\mathrm{T}}PB)$.

Proof. Let $\tilde{C}$ and $\tilde{D}$ be matrices such that

$$M(P) = \begin{pmatrix} \tilde{C} & \tilde{C} \end{pmatrix}^{\mathrm{T}} \begin{pmatrix} \tilde{C} & \tilde{D} \end{pmatrix}.$$

Again using a standard completion-of-the-squares argument, for any initial state x_0 and for any input sequence u such that $x_k \to 0$ we have

$$J(x_0, u) = x_0^{\mathrm{T}}Px_0 + \sum_k \|\tilde{C}x_k + \tilde{D}u_k\|^2 \geq x_0^{\mathrm{T}}x_0 + \|\tilde{C}Px_0 + \tilde{D}u_0\|^2 \tag{4.2}$$

From Lemma 4.3 we have that $J^*(x_0) = x_0^{\mathrm{T}}Px_0$. In particular this implies that the infimum of $\|\tilde{C}x_0 + \tilde{D}u_0\|^2$ over all $u_0 \in \mathbb{R}^m$ is equal to 0. Consequently, for all x_0 there exists $u_0 \in \mathbb{R}^m$ such that $\tilde{C}x_0 + \tilde{D}u_0 = 0$. This implies $\operatorname{im} \tilde{C} \subset \operatorname{im} \tilde{D}$ so

$$\operatorname{rank} M(P) = \operatorname{rank} \begin{pmatrix} \tilde{C} & \tilde{D} \end{pmatrix} = \operatorname{rank} \tilde{D} = \operatorname{rank}\,(D_2^{\mathrm{T}}D_2 + B^{\mathrm{T}}PB). \qquad \square$$

Clearly, the proof of Theorem 4.1 follows by combining these three lemmas. The fact that P is the largest real symmetric solution of the algebraic Riccati equation follows by noting that any real symmetric solution is also a solution of the linear matrix inequality and by applying Lemma 4.4.

Remark 4.6. For later use, note that by combining the above results with Remark 3.11 we obtain that also for the general case the optimal cost $J^*(x_0)$ of the discrete-time linear quadratic problem associated with the system (A, B, C_2, D_2) is given by $J^*(x_0) = x_0^{\mathrm{T}}Px_0$, with P the largest real symmetric solution of the Riccati equation (3.2). We will also need the dual result of Theorem 4.1, which is stated below:

THEOREM 4.7. *If (C_1, A) is detectable, then the Riccati equation* (3.3) *has a largest real symmetric solution, say Q. Q is positive semidefinite. We have $Q = \lim_{\varepsilon\downarrow 0} Q_\varepsilon$, where for $\varepsilon > 0$ Q_ε is the largest real symmetric solution of the algebraic Riccati equation*

$$\begin{aligned} &AQ_\varepsilon A^{\mathrm{T}} - Q_\varepsilon + EE^{\mathrm{T}} + \varepsilon^2 I \\ &\quad -(AQ_\varepsilon C_1^{\mathrm{T}} + ED_1^{\mathrm{T}})(D_1D_1^{\mathrm{T}} + C_1Q_\varepsilon C_1^{\mathrm{T}})^+(C_1Q_\varepsilon A^{\mathrm{T}} + D_1E^{\mathrm{T}}) = 0. \end{aligned} \tag{4.3}$$

Now we are in a position to state the main results of this section. It turns out that also for the discrete-time H_2 problem in its full generality, so without any assumptions on the zeros, the optimal performance J^* is given by (3.10), with P and Q the largest real symmetric solutions of the respective Riccati equations. However, in general no optimal controller will exist. We will, however, derive necessary and sufficient conditions for the existence of an optimal controller. Our first main result deals with the optimal performance.

THEOREM 4.8. *Consider the system* (3.1). *Assume that (A, B) is stabilizable and (C_1, A) is detectable. Then the optimal performance J^* is given by* (3.10), *where P and Q are the largest real symmetric solutions of* (3.2) *and* (3.3), *respectively.*

Proof. In addition to the system (3.1), consider its perturbation $\Sigma^{\varepsilon}_{\mathrm{dis}}$:

$$\begin{aligned} x_{k+1} &= Ax_k + Bu_k + \begin{pmatrix} E & \varepsilon I \end{pmatrix} v_k\,, \\ y_k &= C_1 x_k + \begin{pmatrix} D_1 & 0 \end{pmatrix} v_k\,, \\ z_k &= \begin{pmatrix} C_2 \\ \varepsilon I \end{pmatrix} x_k + \begin{pmatrix} D_2 \\ 0 \end{pmatrix} u_k\,. \end{aligned} \tag{4.4}$$

Let $J_{\Sigma^{\varepsilon}_{\mathrm{dis}}}(\Gamma_{\mathrm{dis}})$ denote the H_2 performance, and let J^*_{ε} denote the optimal H_2 performance. Since, for $\varepsilon > 0$, neither $(A, B, \left(\begin{smallmatrix} C_2 \\ \varepsilon I \end{smallmatrix}\right), \left(\begin{smallmatrix} D_2 \\ 0 \end{smallmatrix}\right))$ nor $(A, (E\ \ \varepsilon I), C_1, (D_1\ \ 0))$ have zeros; we can apply Theorem 3.3 to obtain

$$\begin{aligned} J^*_{\varepsilon} = \operatorname{tr}\,((EE^{\mathrm{T}} + \varepsilon^2 I)P_{\varepsilon}) + \operatorname{tr}\,(A^{\mathrm{T}}P_{\varepsilon}A - P_{\varepsilon} + C_2^{\mathrm{T}}C_2 + \varepsilon^2 I)Q_{\varepsilon}) \\ -\operatorname{tr}\,((D_{P_{\varepsilon}}N^*_{\varepsilon}D_{Q_{\varepsilon}})(D_{P_{\varepsilon}}N^*_{\varepsilon}D_{Q_{\varepsilon}})^{\mathrm{T}}), \end{aligned}$$

where P_{ε} and Q_{ε} are the largest real symmetric solutions of (4.1) and (4.3), respectively, and where $D_{P_{\varepsilon}}$, N^*_{ε}, and $D_{Q_{\varepsilon}}$ are defined by (3.4), (3.11), and (3.7), with P and Q replaced by P_{ε} and Q_{ε}. From Lemma 3.7, recall that

$$-\operatorname{tr}\,((D_{P_{\varepsilon}}N^*_{\varepsilon}D_{Q_{\varepsilon}})(D_{P_{\varepsilon}}N^*_{\varepsilon}D_{Q_{\varepsilon}})^{\mathrm{T}}) = \Phi_{\varepsilon}(N^*_{\varepsilon}) = \min_{N} \Phi_{\varepsilon}(N),$$

with

$$\begin{aligned} \Phi_{\varepsilon}(N) := {} & 2\operatorname{tr}\,\left(\begin{pmatrix} D_1 \\ 0 \end{pmatrix}^{\mathrm{T}} N^{\mathrm{T}}B^{\mathrm{T}}P_{\varepsilon}(E\ \ \varepsilon I)\right) + 2\operatorname{tr}\,(C_{P_{\varepsilon}}Q_{\varepsilon}C_1^{\mathrm{T}}N^{\mathrm{T}}D_{P_{\varepsilon}}) \\ & + \operatorname{tr}\,((D_{P_{\varepsilon}}ND_{Q_{\varepsilon}})(D_{P_{\varepsilon}}ND_{Q_{\varepsilon}})^{\mathrm{T}}) \\ = {} & 2\operatorname{tr}\,(D_1^{\mathrm{T}}N^{\mathrm{T}}B^{\mathrm{T}}P_{\varepsilon}E) + 2\operatorname{tr}\,(Q_{\varepsilon}C_1^{\mathrm{T}}N^{\mathrm{T}}(D_2^{\mathrm{T}}C_2 + B^{\mathrm{T}}P_{\varepsilon}A)) \\ & + \operatorname{tr}\,((D_{P_{\varepsilon}}ND_{Q_{\varepsilon}})(D_{P_{\varepsilon}}ND_{Q_{\varepsilon}})^{\mathrm{T}}). \end{aligned}$$

Since $P_{\varepsilon} \to P$ and $Q_{\varepsilon} \to Q$, we see that for every N we have $\Phi_{\varepsilon}(N) \to \Phi(N)$ ($\varepsilon \downarrow 0$), where $\Phi(N)$ is defined by (3.15). Since of course for all $\varepsilon > 0$ we have $J^* \leq J^*_{\varepsilon}$ we see that for all $\varepsilon > 0$, for all N we have

$$J^* \leq \operatorname{tr}\,((EE^{\mathrm{T}} + \varepsilon^2 I)P_{\varepsilon}) + \operatorname{tr}\,((A^{\mathrm{T}}P_{\varepsilon}A - P_{\varepsilon} + C_2^{\mathrm{T}}C_2 + \varepsilon^2 I)Q_{\varepsilon}) + \Phi_{\varepsilon}(N).$$

Now, letting $\varepsilon \downarrow 0$ on the left in this inequality, we find that for all N

$$J^* \leq \operatorname{tr}\,(EE^{\mathrm{T}}P) + \operatorname{tr}\,((A^{\mathrm{T}}PA - P + C_2^{\mathrm{T}}C_2)Q) + \Phi(N).$$

Finally, taking the minimum over all N, this yields

$$J^* \leq \operatorname{tr}\,(EE^{\mathrm{T}}P) + \operatorname{tr}\,(C_P^{\mathrm{T}}C_PQ) - \operatorname{tr}\,((D_PN^*D_Q)(D_PN^*D_Q)^{\mathrm{T}}).$$

To prove the converse inequality note that by using the fact that P and Q satisfy (3.2) and (3.3) we can apply a repeated completion-of-the-squares argument as in §3 to obtain that for any internally stabilizing compensator Γ_{dis} we have

$$J_{\Sigma_{\mathrm{dis}}}(\Gamma_{\mathrm{dis}}) \geq \operatorname{tr}\,(E^{\mathrm{T}}PE) + \operatorname{tr}\,(C_PQC_P^{\mathrm{T}}) + \Phi(N^*). \tag{4.5}$$

Taking the infimum over all such Γ_{dis} yields the desired inequality. □

Next we will study the question: Under what conditions does there exist an optimal controller? Again, let P and Q be the largest real symmetric solutions of the respective Riccati equations. Define a system $\Sigma_{\mathrm{dis},P,Q}$ by (3.14). Again, for any internally stabilizing compensator $\Gamma_{\mathrm{dis}} = (K, L, M, N)$ we have the inequality (4.5).

As noted in §3, we have equality *if* $N = N^*$ and Γ_{dis} has the property that the closed loop system $\Sigma_{\text{dis},P,Q} \times \Gamma_{\text{dis}}$ has the constant transfer matrix $D_P N^* D_Q$. Of course, the latter statement only gives a sufficient condition for a compensator to be optimal. In the following theorem we will give necessary *and* sufficient conditions for optimality. Let R^* be as defined in Lemma 3.7.

THEOREM 4.9. *A controller* Γ_{dis} *is optimal if and only if* $\Sigma_{\text{dis},P,Q} \times \Gamma_{\text{dis}}$ *is internally stable and has constant transfer matrix* R^*.

Proof. If $\Gamma_{\text{dis}} = (K, L, M, N)$ is optimal, then we have

$$J_{\Sigma_{\text{dis}}}(\Gamma_{\text{dis}}) = \text{tr } (E^{\mathrm{T}} P E) + \text{tr } (C_P^{\mathrm{T}} Q C_P) + \Phi^*.$$

By Lemma 3.6 we also have

$$J_{\Sigma_{\text{dis}}}(\Gamma_{\text{dis}}) \geq \text{tr } (E^{\mathrm{T}} P E) + \text{tr } (C_P^{\mathrm{T}} Q C_P) + \Phi(N).$$

This clearly yields $\Phi(N) = \Phi^*$, i.e., N minimizes the function Φ. Again by Lemma 3.6 this implies that $\Sigma_{\text{dis},P,Q} \times \Gamma_{\text{dis}}$ has the constant transfer matrix $D_P N D_Q$. However, since N minimizes Φ, by Lemma 3.7 we have $D_P N D_Q = -R^*$. The converse statement is also an immediate consequence of Lemma 3.6. □

Our aim is to reformulate these conditions in terms of the original system Σ_{dis}. For any given matrix $N \in \mathbb{R}^{m \times p}$, consider the system $\Sigma^N_{\text{dis},P,Q}$ that is obtained by applying to $\Sigma_{\text{dis},P,Q}$ the static output feedback $u = Ny + v$. This system $\Sigma^N_{\text{dis},P,Q}$ is described by

$$\begin{array}{llll} x_{k+1} = & (A + BNC_1)x_k + & Bv_k + (BND_Q + E_Q)d_k\ , \\ y_k \;\;\; = & C_1 x_k & + \qquad\qquad D_Q d_k\ , \\ z_k \;\;\; & = (C_P + D_P N C_1)x_k + D_P v_k\ , \end{array} \tag{4.6}$$

Also, for a given compensator $\Gamma_{\text{dis}} = (K, L, M, N)$, let $\Gamma^0_{\text{dis}} := (K, L, M, 0)$ be the compensator with direct feedthrough matrix N replaced by 0. It is clear that the closed-loop system $\Sigma_{\text{dis},P,Q} \times \Gamma_{\text{dis}}$ has constant transfer matrix $D_P N D_Q$ if and only if $\Sigma^N_{\text{dis},P,Q} \times \Gamma^0_{\text{dis}}$ has transfer matrix equal to 0. Consequently, an internally stabilizing compensator $\Gamma_{\text{dis}} = (K, L, M, N)$ is optimal if and only if $D_P N D_Q = -R^*$ and $\Sigma^N_{\text{dis},P,Q} \times \Gamma^0_{\text{dis}}$ has transfer matrix 0. In other words, in order to find necessary and sufficient conditions for the existence of an optimal controller, we should study the problem of disturbance decoupling with internal stability. This problem has been studied extensively in [16]. One of the main results of [16] gives necessary and sufficient conditions for the existence of an internally stabilizing strictly proper compensator Γ^0_{dis} for the system Σ_{dis} given by (3.1). We will briefly recall this result here. Given Σ_{dis}, let $\mathcal{V}_g$ denote the largest subspace of $\mathbb{R}^n$ for which there exists $F \in \mathbb{R}^{m \times n}$ such that $(A + BF)\mathcal{V}_g \subset \mathcal{V}_g$, $|\sigma(A + BF \mid \mathcal{V}_g)| < 1$, and $(C_2 + D_2 F)\mathcal{V}_g = 0$. Dually, let $\mathcal{S}_g$ be the smallest subspace of $\mathbb{R}^n$ for which there exists a matrix $G \in \mathbb{R}^{n \times p}$ such that $(A + GC_1)\mathcal{S}_g \subset \mathcal{S}_g$, $|\sigma(A + GC_1 \mid \mathbb{R}^n/\mathcal{S}_g)| < 1$, and im $(E + GD_1) \subset \mathcal{S}_g$. It was shown in [16, Thm. 2.4] that there exists an internally stabilizing compensator $\Gamma^0_{\text{dis}} = (K, L, M, 0)$ such that $\Sigma_{\text{dis}} \times \Gamma^0_{\text{dis}}$ has transfer matrix 0 if and only if the following conditions hold: (i) (A, B) is stabilizable and (C_1, A) is detectable, (ii) the following four subspace inclusions hold: im $E \subset \mathcal{V}_g$, $\mathcal{S}_g \subset \ker C_2$, $\mathcal{S}_g \subset \mathcal{V}_g$, and $A\mathcal{S}_g \subset \mathcal{V}_g$.

Here, we want to apply this result to the system $\Sigma^N_{\text{dis},P,Q}$, with N any solution of $D_P N D_Q = -R^*$. In the following, we will omit some of the details. Using the fact that im $(C_P + D_P N C_1) \subset$ im D_P, it can be shown that the subspace $\mathcal{V}_g$ associated with $\Sigma^N_{\text{dis},P,Q}$ is given by

$$\mathcal{V}_g = \mathcal{X}_g(A - BD_P^+ C_P) + \langle A - BD_P^+ C_P \mid B \ker D_P \rangle, \tag{4.7}$$

where for a given matrix M, $\mathcal{X}_g(M)$ is the sum of the generalized eigenspaces of M associated with its eigenvalues in $|z| < 1$, and where $\langle M \mid \mathcal{L} \rangle$ is the smallest M-invariant subspace contained in $\mathcal{L}$. It can also be shown, using the fact that $\ker D_Q \subset \ker(BND_Q + E_Q)$, that

$$\mathcal{S}_g = \mathcal{X}_b(A - E_Q D_Q^+ C_1) \cap \langle C_1^{-1} \text{im } D_Q \mid A - E_Q D_Q^+ C_1 \rangle, \tag{4.8}$$

where $\mathcal{X}_b(M)$ is the sum of the generalized eigenspaces of M associated with its eigenvalues in $|z| \geq 1$ and where $\langle \mathcal{L} \mid M \rangle$ is the largest M-invariant subspace containing $\mathcal{L}$. Using the fact that, from (4.7), $B \ker D_P \subset \mathcal{V}_g$, it can be shown that im $(BND_Q + E_Q) \subset \mathcal{V}_g$ if and only if

$$\text{im } (E_Q - BD_P^+ R^*) \subset \mathcal{V}_g. \tag{4.9}$$

Using the fact that, by (4.8), $\mathcal{S}_g \subset C_1^{-1}$im D_Q, it can be shown that $\mathcal{S}_g \subset \ker(C_P + D_P N C_1)$ if and only if

$$\mathcal{S}_g \subset \ker(C_P - R^* D_Q^+ C_1). \tag{4.10}$$

Finally, it can be shown that $(A + BNC_1)\mathcal{S}_g \subset \mathcal{V}_g$ if and only if

$$(A - BD_P^+ R^* D_Q^+ C_1)\mathcal{S}_g \subset \mathcal{V}_g. \tag{4.11}$$

Collecting the above facts, we then obtain the following necessary and sufficient conditions for the existence of an optimal controller for the discrete-time H_2 optimal control problem associated with the system Σ_{dis}:

THEOREM 4.10. *Consider the system* (3.1). *Assume that* (A, B) *is stabilizable and* (C_1, A) *is detectable. Let P and Q be the largest real symmetric solution of* (3.2) *and* (3.3), *respectively. Let $\mathcal{V}_g$ and $\mathcal{S}_g$ be given by* (4.7) *and* (4.8). *Then we have: there exists an optimal controller, i.e., an internally stabilizing controller* $\Gamma^*_{\text{dis}} = (K^*, L^*, M^*, N^*)$ *such that* $J_{\Sigma_{\text{dis}}}(\Gamma^*_{\text{dis}}) = J^*$, *if and only if the four subspace inclusions* $\mathcal{S}_g \subset \mathcal{V}_g$, (4.9), (4.10), *and* (4.11) *are satisfied.*

5. The sampled-data H_2 problem.

Now we return to the sampled-data H_2 problem. Consider the continuous-time system Σ given by (2.3), and let $\Delta \notin \mathbf{\Delta}$ be a given sampling period. Let the discrete-time system Σ_Δ be given by (2.4). According to Theorem 2.1, the optimal sampled-data H_2 performance $J^*_{\Sigma,\Delta}$ is equal to

$$J^*_{\Sigma,\Delta} = \frac{1}{\Delta}\int_0^\Delta \int_0^{\Delta - s} \text{tr}\left(C_2 e^{tA} E E^{\mathrm{T}} e^{tA^{\mathrm{T}}} C_2^{\mathrm{T}}\right) dt\, ds + \frac{1}{\Delta} J^*_{\Sigma_\Delta}, \tag{5.1}$$

where $J^*_{\Sigma_\Delta}$ is the optimal discrete-time H_2 performance associated with Σ_Δ. According to Theorem 4.8, the optimal performance $J^*_{\Sigma_\Delta}$ can be found in terms of two algebraic Riccati equations associated with Σ_Δ. According to Theorem 4.10, an optimal compensator $\Gamma_{\text{dis},\Delta}$ exists if and only if four subspace inclusions involving subspaces associated with the system Σ_Δ are satisfied. According to Theorem 3.3, if the systems $(A_\Delta, B_\Delta, C_{2,\Delta}, D_{2,\Delta})$ and $(A_\Delta, E_\Delta, C_1, 0)$ have no zeros on the unit circle, then an optimal compensator $\Gamma_{\text{dis},\Delta}$ exists and can be calculated using the "construction" in the statement of Theorem 3.3. The sampled-data controller $\Gamma := H_\Delta \Gamma_{\text{dis},\Delta} S_\Delta$ is then optimal for the sampled-data H_2 problem under consideration.

In this section we study the following question: what are conditions *in terms of the original system* Σ that guarantee that there exists an optimal compensator for the sampled-data H_2 problem? Instead of being completely general, we will study the following question: what are necessary and sufficient conditions in terms of the original system Σ such that $(A_\Delta, B_\Delta, C_{2,\Delta}, D_{2,\Delta})$ and $(A_\Delta, E_\Delta, C_1, 0)$ have no zeros on the unit circle? In the following, let $\mathcal{R}$ be the controllability subspace of the system (A, B, C_2, D_2) (see §2). The main results of this section are the following:

THEOREM 5.1. *Consider the system* Σ. *Let* $\Delta > 0$.

(i) *Let* λ *be a zero of* $(A_\Delta, B_\Delta, C_{2,\Delta}, D_{2,\Delta})$, $\lambda \neq 1$. *Then there exists a unobservable eigenvalue* μ *of* (C_2, A) *such that* $\lambda = e^{\mu\Delta}$.

(ii) *If* (A, B, C_2, D_2) *is left-invertible, then also the converse of* (i) *holds*: *if* μ *is an unobservable eigenvalue of* (C_2, A), *then* $e^{\mu\Delta}$ *is a zero of* $(A_\Delta, B_\Delta, C_{2,\Delta}, D_{2,\Delta})$.

(iii) 1 *is a zero of* $(A_\Delta, B_\Delta, C_{2,\Delta}, D_{2,\Delta})$ *if and only if at least one of the following two conditions hold*:

(a) *0 is a zero of* (A, B, C_2, D_2),

(b)

$$\mathcal{R} \not\subset \langle \ker C_2 \mid A \rangle. \tag{5.2}$$

(iv) *If* (A, B, C_2, D_2) *is left-invertible, then* 1 *is a zero of* $(A_\Delta, B_\Delta, C_{2,\Delta}, D_{2,\Delta})$ *if and only if* 0 *is a zero of* (A, B, C_2, D_2).

COROLLARY 5.2. *Consider the system* Σ. *Let* $\Delta > 0$.

(i) *If* (C_2, A) *has no unobservable eigenvalues on the imaginary axis,* 0 *is not a zero of* (A, B, C_2, D_2), *and* $\mathcal{R} \subset \langle \ker C_2 \mid A \rangle$, *then* $(A_\Delta, B_\Delta, C_{2,\Delta}, D_{2,\Delta})$ *has no zeros on the unit circle.*

(ii) *If* (A, B, C_2, D_2) *is left-invertible, then* $(A_\Delta, B_\Delta, C_{2,\Delta}, D_{2,\Delta})$ *has no zeros on the unit circle if and only if* (C_2, A) *has no unobservable eigenvalues on the imaginary axis and* 0 *is not a zero of* (A, B, C_2, D_2).

THEOREM 5.3. *Consider the system* Σ. *Let* $\Delta > 0$.

(i) *Let* λ *be a zero of* $(A_\Delta, E_\Delta, C_1, 0)$. *Then there exists an uncontrollable eigenvalue* μ *of* (A, E) *such that* $\lambda = e^{\mu\Delta}$.

(ii) *If* $(A, E, C_1, 0)$ *is right-invertible, then also the converse of* (i) *holds*; *i.e., if* μ *is an uncontrollable eigenvalue of* (A, E), *then* $e^{\mu\Delta}$ *is a zero of* $(A_\Delta, E_\Delta, C_1, 0)$.

COROLLARY 5.4. *Consider the system* Σ. *Let* $\Delta > 0$. *If* (A, E) *has no uncontrollable eigenvalues on the imaginary axis, then* $(A_\Delta, E_\Delta, C_1, 0)$ *has no zeros on the unit circle. If, in addition,* $(A, E, C_1, 0)$ *is right-invertible, then also the converse holds*: $(A_\Delta, E_\Delta, C_1, 0)$ *has no zeros on the unit circle if and only if* (A, E) *has no uncontrollable eigenvalues on the imaginary axis.*

Note that the conditions on Σ obtained in these theorems are independent of the sampling period. In the remainder of this section we shall prove these results.

In order to study the zeros of (A, B, C_2, D_2) and $(A_\Delta, B_\Delta, C_{2,\Delta}, D_{2,\Delta})$, consider the system matrices of these systems. Let

$$P_\Delta(z) := \begin{pmatrix} zI - A_\Delta & -B_\Delta \\ C_{2,\Delta} & D_{2,\Delta} \end{pmatrix}, \quad P(s) := \begin{pmatrix} sI - A & -B \\ C_2 & D_2 \end{pmatrix}.$$

Recall that λ is a zero of $(A_\Delta, B_\Delta, C_{2,\Delta}, D_{2,\Delta})$ if and only if the rank of the complex matrix $P_\Delta(\lambda)$ is less than the normal rank of P_Δ (see §2). In order to find out in which points λ this happens, we will study for $\lambda \in \mathcal{C}$ the subspace

$$\mathcal{V}_\lambda := \ker P_\Delta(\lambda) \subset \mathcal{C}^{n+m}.$$

Clearly, for all λ we have $\dim \mathcal{V}_\lambda = n + m - \text{rank } P_\Delta(\lambda)$. Consequently, for all but finitely many λ we have $\dim \mathcal{V}_\lambda = d$, where

$$d := n + m - \text{normrank } P_\Delta.$$

Hence, λ is a zero of $(A_\Delta, B_\Delta, C_{2,\Delta}, D_{2,\Delta})$ if and only if $\dim \mathcal{V}_\lambda > d$. In the following lemma we will calculate for each λ the subspace $\mathcal{V}_\lambda$, its dimension $\dim \mathcal{V}_\lambda$, and the number d. Denote the unobservable subspace $\langle \ker C_2 \mid A \rangle$ by $\mathcal{N}$. Define a subspace $\mathcal{W}$ as follows:

$$\mathcal{W} := B^{-1}\mathcal{N} \cap \ker D_2. \tag{5.3}$$

LEMMA 5.5. *For every $\lambda \in \mathcal{C}$, $\lambda \neq 1$ we have*

$$\mathcal{V}_\lambda = (\mathcal{N} \times \mathcal{W}) \cap \ker \begin{pmatrix} \lambda I - A_\Delta & B_\Delta \end{pmatrix}, \tag{5.4}$$

$$\dim \mathcal{V}_\lambda = \dim \mathcal{N} + \dim \mathcal{W} - \dim((\lambda I - A_\Delta)\mathcal{N} + B_\Delta \mathcal{W}). \tag{5.5}$$

For all but finitely many λ we have $\dim \mathcal{V}_\lambda = d = \dim \mathcal{W}$, *equivalently,* normrank $P_\Delta = n + m - \dim \mathcal{W}$. *In addition we have*

$$\mathcal{V}_1 = \ker \begin{pmatrix} -A & -B \\ C_2 & D_2 \end{pmatrix}. \tag{5.6}$$

Proof. We will first prove (5.4). We know $\binom{x_0}{u_0} \in \mathcal{V}_\lambda$ if and only if

$$A_\Delta x_0 + B_\Delta u_0 = \lambda x_0, \tag{5.7}$$
$$C_{2,\Delta} x_0 + D_{2,\Delta} u_0 = 0. \tag{5.8}$$

Consider the differential equation $\dot{x}(t) = Ax(t) + Bu_0$, $x(0) = x_0$; and define $z(t) := C_2 x(t) + Du_0$. Clearly, $x(\Delta) = A_\Delta x_0 + B_\Delta u_0$, so (5.7) is equivalent to $x(\Delta) = \lambda x_0$. In turn, this is equivalent to

$$(\lambda - 1)x_0 = \int_0^\Delta e^{At}(Ax_0 + Bu_0)dt. \tag{5.9}$$

Using the definition (2.6) of $C_{2,\Delta}$ and $D_{2,\Delta}$, we see that (5.8) is equivalent to

$$(C_2 \;\; D_2)e^{\underline{A}t}\begin{pmatrix} x_0 \\ u_0 \end{pmatrix} = 0 \text{ for all } t \in [0, \Delta],$$

which, in turn, is equivalent to $z(t) = 0$ for all $t \in [0, \Delta]$. Obviously,

$$z(t) = C_2 e^{At} x_0 + \left[C_2 \int_0^t e^{As} B ds + D_2 \right] u_0.$$

Since $z(t) = 0$ for all $t \in [0, \Delta]$ is satisfied if and only if $z(0) = 0$ and $\dot{z}(t) = 0$ for all $t \in [0, \Delta]$, we find that (5.8) is equivalent to

$$C_2 x_0 + D_2 u_0 = 0 \quad \text{and} \quad C_2 e^{At}(Ax_0 + Bu_0) = 0, \qquad t \in [0, \Delta].$$

In other words (5.8) is satisfied if and only if

$$(5.10) \qquad C_2x_0 + D_2u_0 = 0 \quad \text{and} \quad Ax_0 + Bu_0 \in \mathcal{N}.$$

Now assume that $\lambda \neq 1$. Then (5.9) and (5.10) imply that $x_0 \in \mathcal{N} \subset \ker C_2$, so $u_0 \in \ker D_2$. Also it follows that $Ax_0 \in \mathcal{N}$, so $Bu_0 \in \mathcal{N}$ and, in fact, $u_0 \in \mathcal{W}$. We conclude that, for $\lambda \neq 1$, $\mathcal{V}_\lambda \subset (\mathcal{N} \times \mathcal{W}) \cap \ker \begin{pmatrix} \lambda I - A_\Delta & B_\Delta \end{pmatrix}$. To prove the converse inclusion, note that $u_0 \in \mathcal{W}$ implies that $D_2u_0 = 0$ and $Bu_0 \in \mathcal{N}$. If, in addition, $x_0 \in \mathcal{N}$, then we have $C_2x_0 + D_2u_0 = 0$ and $Ax_0 + Bu_0 \in \mathcal{N}$. By the above this is equivalent to (5.8). This completes the proof of (5.4).

To prove (5.5), note that, in general, if $\mathcal{L}$ is a subspace of some finite-dimensional linear space $\mathcal{X}$ and if T is a linear map acting on $\mathcal{X}$, then we have $\dim(\mathcal{L} \cap \ker T) = \dim \mathcal{L} - \dim T\mathcal{L}$. Applying this to the situation at hand, we find that for any $\lambda \neq 1$ we have

$$\dim \mathcal{V}_\lambda = \dim(\mathcal{N} \times \mathcal{W}) - \dim(\lambda I - A_\Delta \quad B_\Delta)(\mathcal{N} \times \mathcal{W}),$$

which immediately yields (5.5).

Next, we will prove the statement on the dimension of $\mathcal{V}_\lambda$. First note that since $\mathcal{N}$ is A-invariant, it is also e^{At}-invariant, for any t. In particular, this implies that $\mathcal{N}$ is A_Δ-invariant and invariant under $\int_0^\Delta e^{At}dt$. Now assume that $\lambda \notin \sigma(A_\Delta)$. Then we have $(\lambda I - A_\Delta)\mathcal{N} = \mathcal{N}$. Also, since $B\mathcal{W} \subset \mathcal{N}$, we have $B_\Delta \mathcal{W} \subset \mathcal{N}$. This implies that $(\lambda I - A_\Delta)\mathcal{N} + B_\Delta\mathcal{W} = \mathcal{N}$. If, in addition, we assume that $\lambda \neq 1$, then (5.5) yields $\dim \mathcal{V}_\lambda = \dim \mathcal{W}$.

Finally, to prove (5.6), recall that (5.7) is equivalent to (5.9). Note that for all $\Delta > 0$, $\int_0^\Delta e^{At}dt$ is a nonsingular matrix (this can be shown using the Jordan form of A). Thus, for the case that $\lambda = 1$ (5.9) is equivalent to $Ax_0 + Bu_0 = 0$. Together with the fact that (5.8) is equivalent to (5.10), this proves (5.6). □

By applying this lemma, we are now able to prove the statements (i) and (ii) of Theorem 5.1:

Proof of Theorem 5.1 (i) *and* (ii). (i) Assume that $\lambda \neq 1$ is a zero of $(A_\Delta, B_\Delta, C_{2,\Delta}, D_{2,\Delta})$. Then we must have $\dim \mathcal{V}_\lambda > \dim \mathcal{W}$. Using (5.5) this implies

$$(5.11) \qquad \dim \mathcal{N} > \dim((\lambda I - A_\Delta)\mathcal{N} + B_\Delta \mathcal{W}).$$

As noted in the proof of Lemma 5.5, $\mathcal{N}$ is A_Δ-invariant and $B_\Delta\mathcal{W} \subset \mathcal{N}$. Consequently,

$$(\lambda I - A_\Delta)\mathcal{N} + B_\Delta\mathcal{W} \subset \mathcal{N}.$$

Together with the inequality (5.11), this implies that $(\lambda I - A_\Delta)\mathcal{N}$ is a *strict* subspace of $\mathcal{N}$. This implies that the map $(\lambda I - A_\Delta)$ restricted to $\mathcal{N}$ is singular. Thus, $\ker(\lambda I - A_\Delta) \cap \mathcal{N} \neq 0$. Clearly, this intersection is A-invariant, so the restriction of A to this intersection has an eigenvalue, say μ, with corresponding eigenvector p. This eigenvector satisfies $A_\Delta p = \lambda p$. Also, since $Ap = \mu p$, we have $A_\Delta p = e^\mu p$, so $\lambda = e^\mu$. Finally, $p \in \mathcal{N} \subset \ker C_2$, so μ is an unobservable eigenvalue of (C_2, A).

(ii) We claim that if (A, B, C_2, D_2) is left-invertible, then $\dim \mathcal{W} = 0$. Indeed, left-invertibility is equivalent to the conditions $\binom{B}{D_2}$ is injective and $\mathcal{V} \cap B \ker D_2 = 0$, where $\mathcal{V}$ denotes the weakly unobservable subspace associated with (A, B, C_2, D_2) (see §2). Assume that $u_0 \in \mathcal{W}$. Then we have $D_2u_0 = 0$ and $Bu_0 \in \mathcal{N}$. Since $\mathcal{N} \subset \mathcal{V}$, this yields $Bu_0 = 0$. Combining this with $D_2u_0 = 0$ then leads to $u_0 = 0$. This proves our claim. Now let μ be a unobservable eigenvalue of (C_2, A). There exists $x_0 \neq 0$ such that $Ax_0 = \mu x_0$ and $C_2 x_0 = 0$. This yields $A_\Delta x_0 = \lambda x_0$, with $\lambda := e^{\mu\Delta}$. From the definition of $C_{2,\Delta}$ it is also easily seen that $C_{2,\Delta}x_0 = 0$. Consequently, $\binom{x_0}{0} \in \mathcal{V}_\lambda$, so $\dim \mathcal{V}_\lambda > 0 = \dim \mathcal{W}$. This implies that λ is a zero of $(A_\Delta, B_\Delta, C_{2,\Delta}, D_{2,\Delta})$. □

In order to prove statements (iii) and (iv) of Theorem 5.1, we need the following lemma.

LEMMA 5.6. *Let $\Delta > 0$. Then we have*

$$\text{normrank } P_\Delta \geq \text{normrank } P, \tag{5.12}$$

with equality if and only if $\mathcal{R} \subset \mathcal{N}$.

Proof. For each $\lambda \notin \sigma(A)$ define a subspace $\mathcal{L}_\lambda$ by

$$\mathcal{L}_\lambda := \left\{ \begin{pmatrix} x_0 \\ u_0 \end{pmatrix} \mid u_0 \in \mathcal{W}, x_0 = (\lambda I - A)^{-1} B u_0 \right\}.$$

Clearly, $\mathcal{L}_\lambda \subset \ker P(\lambda)$ and $\dim \mathcal{L}_\lambda = \dim \mathcal{W}$. Consequently, for each $\lambda \notin \sigma(A)$ we have $\dim \mathcal{W} \leq \dim \ker P(\lambda)$. This implies normrank $P \leq n + m - \dim \mathcal{W}$. The inequality (5.12) then follows from Lemma 5.5.

Of course, normrank $P_\Delta =$ normrank P if and only if $\dim \ker P(\lambda) = \dim \mathcal{W}$ for all but finitely many λ, which, in turn, is equivalent to $\ker P(\lambda) = \mathcal{L}_\lambda$ for all but finitely many λ, $\lambda \notin \sigma(A)$. We will prove that the latter statement is equivalent to $\mathcal{R} \subset \mathcal{N}$.

Let $k := \dim \mathcal{R}$, and let $\lambda_1, \ldots, \lambda_k$ be distinct complex numbers, $\lambda_i \notin \sigma(A)$, such that $\ker P(\lambda_i) = \mathcal{L}_{\lambda_i}$. There exists $F \in \mathbb{R}^{m \times n}$ such that $(A + BF)\mathcal{R} \subset \mathcal{R}$, $(C_2 + D_2 F)\mathcal{R} = 0$, and $\sigma(A + BF \mid \mathcal{R}) = \{\lambda_1, \ldots, \lambda_k\}$. Let $x_1, \ldots, x_k \in \mathcal{R}$ be corresponding eigenvectors of $A + BF \mid \mathcal{R}$. Then $\{x_1, \ldots, x_k\}$ is a basis of $\mathcal{R}$. We will prove that $x_i \in \mathcal{N}$. Indeed, define $u_i := -Fx_i$. Then $\binom{x_i}{u_i} \in \ker P(\lambda_i) = \mathcal{L}_{\lambda_i}$. Since $u_i \in \mathcal{W}$, we have $Bu_i \in \mathcal{N}$, so $x_i = (\lambda_i I - A)^{-1} B u_i \in \mathcal{N}$ by A-invariance of $\mathcal{N}$. We conclude that $x_i \in \mathcal{N}$, so $\mathcal{R} \subset \mathcal{N}$.

Conversely, assume that $\mathcal{R} \subset \mathcal{N}$. It suffices to show that $\ker P(\lambda) \subset \mathcal{L}_\lambda$ for all but finitely many λ. Let λ be arbitrary, $\lambda \notin \sigma(A)$, and λ not a zero of (A, B, C_2, D_2). Let $\binom{x_0}{u_0} \in \ker P(\lambda)$. We will prove that $x_0 \in \mathcal{R}$, so $x_0 \in \mathcal{N}$. Assume that $x_0 \neq 0$. Let $F \in \mathbb{R}^{m \times n}$ be such that $Fx_0 = u_0$. Then we have $(A + BF)x_0 = \lambda x_0$ and $(C_2 + D_2 F)x_0 = 0$. This implies $x_0 \in \mathcal{V}$, the weakly unobservable subspace associated with the system (A, B, C_2, D_2). (Indeed, the one-dimensional subspace $\mathcal{L}$ spanned by the vector x_0 has the property that $(A + BF)\mathcal{L} \subset \mathcal{L}$ and $(C_2 + D_2 F)\mathcal{L} = 0$ and so must be contained in $\mathcal{V}$, the largest subspace for which such F exists.) By extending the linear map F to the whole subspace $\mathcal{V}$, we obtain that $(A + BF)\mathcal{V} \subset \mathcal{V}$ and $(C_2 + D_2 F)\mathcal{V} = 0$, so $\lambda \in \sigma(A + BF \mid \mathcal{V})$. We have assumed that λ is not a zero. This implies $\lambda \notin \sigma(A + BF \mid \mathcal{V}/\ R)$ (the latter spectrum is equal to the set of zeros of (A, B, C_2, D_2); see [19]). But then we must have $x_0 \in \mathcal{R}$. This implies that $x_0 \in \mathcal{N}$. Now $(\lambda I - A)x_0 - Bu_0 = 0$, so $Bu_0 \in \mathcal{N}$. This implies that $u_0 \in \mathcal{W}$. For $\lambda \notin \sigma(A)$ this then yields $x_0 \in \mathcal{L}_\lambda$. This completes the proof of the lemma. □

Proof of Theorem 5.1 (iii) *and* (iv). (iii) We will prove that 1 is *not* a zero of the system $(A_\Delta, B_\Delta, C_{2,\Delta}, D_{2,\Delta})$ if and only if 0 is not a zero of (A, B, C_2, D_2) ánd normrank $P =$ normrank P_Δ. Clearly, 1 is not a zero of $(A_\Delta, B_\Delta, C_{2,\Delta}, D_{2,\Delta})$ if and only if $\dim \mathcal{V}_1 = n+m-$normrank P_Δ. By (5.6) we have $\dim \mathcal{V}_1 = n+m-$rank $P(0) \geq n + m -$ normrank P, with strict inequality if and only if 0 is a zero of (A, B, C_2, D_2). Combining these facts proves our claim. The proof of (iii) is then completed by applying Lemma 5.6.

(iv) If (A, B, C_2, D_2) is left-invertible, then $\mathcal{R} = 0$. In that case condition (5.2) is never satisfied. □

In order to study the zeros of $(A_\Delta, E_\Delta, C_1, 0)$, consider the system matrix of this system. Let

$$Q_\Delta(z) := \begin{pmatrix} zI - A_\Delta & -E_\Delta \\ C_1 & 0 \end{pmatrix}.$$

As before, λ is a zero of $(A_\Delta, E_\Delta, C_1, 0)$ if and only if the rank of the complex matrix $Q_\Delta(\lambda)$ is less than the normal rank of Q_Δ (see §2). In order to find out in which points λ this happens, we will study for $\lambda \in \mathcal{C}$ the subspace

$$\mathcal{W}_\lambda := (\text{im}\, Q_\Delta(\lambda))^\perp \subset \mathcal{C}^{n+p}.$$

For all λ we have $\dim \mathcal{W}_\lambda = n + p - \text{rank}\ Q_\Delta(\lambda)$. Consequently, for all but finitely many λ we have $\dim \mathcal{W}_\lambda = d_1$, where

$$d_1 := n + p - \text{normrank}\ Q_\Delta.$$

Hence, λ is a zero of $(A_\Delta, E_\Delta, C_1, 0)$ if and only if $\dim \mathcal{W}_\lambda > d_1$. The following lemma calculates for each λ the subspace $\mathcal{W}_\lambda$, its dimension $\dim \mathcal{W}_\lambda$, and the number d_1. Let $\mathcal{M} := \langle A \mid \text{im}\, E\rangle$, the reachable subspace of (A, E).

LEMMA 5.7. *Let* $\Delta > 0$. *Then we have*

$$\mathcal{W}_\lambda = \left(M^\perp \times (C_1^{\mathrm{T}})^{-1} M^\perp\right) \cap \ker \begin{pmatrix} \lambda I - A_\Delta^{\mathrm{T}} & C_1^{\mathrm{T}} \end{pmatrix},$$

$$\begin{aligned} \dim \mathcal{W}_\lambda = {} & \dim M^\perp + \dim (C_1^{\mathrm{T}})^{-1} M^\perp \\ & - \dim((\lambda I - A_\Delta^{\mathrm{T}}) M^\perp + C_1^{\mathrm{T}} (C_1^{\mathrm{T}})^{-1} M^\perp). \end{aligned} \tag{5.13}$$

For all but finitely many λ *we have* $\dim \mathcal{W}_\lambda = d_1 = \dim (C_1^{\mathrm{T}})^{-1} M^\perp$, *equivalently,*

$$\text{normrank}\ Q_\Delta = n + p - \dim (C_1^{\mathrm{T}})^{-1} M^\perp.$$

Proof. By definition, $\binom{x_0}{y_0} \in \mathcal{W}_\lambda$ if and only if

$$(\lambda I - A_\Delta^{\mathrm{T}}) x_0 + C_1^{\mathrm{T}} y_0 = 0 \quad \text{and} \quad x_0^{\mathrm{T}} E_\Delta = 0. \tag{5.14}$$

Since, by definition, im $E_\Delta = \mathcal{M}$, we see that it suffices to show that (5.14) implies $y_0 \in (C_1^{\mathrm{T}})^{-1} M^\perp$. From the fact that $\mathcal{M}^\perp$ is A^{T}-invariant it follows that $\mathcal{M}^\perp$ is A_Δ^{T}-invariant, so $C_1^{\mathrm{T}} y_0 \in \mathcal{M}^\perp$. The statement (5.13) on the dimension of $\mathcal{W}_\lambda$ follows in the same way as the corresponding statement in the previous lemma.

Now let λ be any complex number such that $\lambda \notin \sigma(A_\Delta^{\mathrm{T}})$. Since $\mathcal{M}^\perp$ is A_Δ^{T}-invariant, we then have $(\lambda I - A_\Delta^{\mathrm{T}}) M^\perp = M^\perp$. Also we have $C_1^{\mathrm{T}} (C_1^{\mathrm{T}})^{-1} M^\perp \subset M^\perp$ (no equality!). Thus, for such λ we have $\dim \mathcal{W}_\lambda = \dim (C_1^{\mathrm{T}})^{-1} M^\perp$. □

We are now ready to prove Theorem 5.3.

Proof of Theorem 5.3. Let λ be a zero of $(A_\Delta, E_\Delta, C_1, 0)$. Then we have $\dim \mathcal{W}_\lambda > \dim (C_1^{\mathrm{T}})^{-1} M^\perp$. Consequently, by (5.13), $\dim M^\perp > \dim((\lambda I - A_\Delta^{\mathrm{T}}) M^\perp + C_1^{\mathrm{T}} (C_1^{\mathrm{T}})^{-1} M^\perp)$. In particular, this implies that $(\lambda I - A_\Delta^{\mathrm{T}}) \mathcal{M}^\perp$ is a *strict* subspace of $\mathcal{M}^\perp$, so $\ker(\lambda I - A_\Delta^{\mathrm{T}}) \cap \mathcal{M}^\perp \neq 0$. This subspace is A^{T}-invariant, so there exist μ and

$x_0 \in \mathcal{M}^\perp$, $x_0 \neq 0$, such that $A^T x_0 = \mu x_0$, $A_\Delta^T x_0 = \lambda x_0$, and $x_0 \in \mathcal{M}^\perp$. Obviously, this implies $\lambda = e^\mu \Delta$, and μ is an uncontrollable eigenvalue of (A, E).

Assume that $(A, E, C_1, 0)$ is right-invertible. Let

$$Q(s) := \begin{pmatrix} sI - A & -E \\ C_1 & 0 \end{pmatrix}$$

be the system matrix. We have normrank $Q = n+p$. We claim that also normrank $Q_\Delta = n+p$. Indeed, assume that $y_0 \neq 0$ is an element of $(C_1^T)^{-1} M^\perp$. For $\lambda \notin \sigma(A^T)$, define $x_0 := -(\lambda I - A^T)^{-1} C_1^T y_0$. Then $x_0 \in \mathcal{M}^\perp$ and we have $(x_0^T \;\; y_0^T) Q(\lambda) = (0 \;\; 0)$. Thus, for all but finitely many λ we have rank $Q(\lambda) < n+p$, which is a contradiction. Hence we must have $(C_1^T)^{-1} M^\perp = 0$.

It follows that λ is a zero if and only if $\mathcal{W}_\lambda \neq 0$. Assume that μ is an uncontrollable eigenvalue of (A, E). Then there exists $x_0 \neq 0$, $x_0 \in \mathcal{M}^\perp$, such that $x_0^T A^T = \mu x_0$. Define $\lambda := e^{\mu\Delta}$. Then we have $x_0^T E_\Delta^T = 0$ and $x_0^T(\lambda I - A_\Delta) = 0$. It follows that $\binom{x_0}{0} \in \mathcal{W}_\lambda$, so λ is a zero of $(A_\Delta, E_\Delta, C_1, 0)$. □

6. Performance recovery and convergence of optimal performance. In this section we study the connection between the 'ordinary' continuous-time H_2 problem and the sampled-data H_2 problem. In particular, we are interested in the following questions:

- Suppose that we control the system Σ by means of an internally stabilizing continuous-time compensator Γ_{con}, yielding continuous-time H_2 performance $J_\Sigma(\Gamma_{\text{con}})$. Is it possible to recover this performance asymptotically by using a sampled-data controller with sufficiently small sampling period? More precisely, is it true that for all $\epsilon > 0$ there exists $\Delta > 0$ and an internally stabilizing sampled-data controller Γ with sampling-period Δ such that $|J_\Sigma(\Gamma_{\text{con}}) - J_{\Sigma,\Delta}(\Gamma)| < \epsilon$?
- Does the optimal sampled-data H_2 performance converge to the optimal continuous-time H_2 performance as the sampling period Δ decreases to zero? More precisely, suppose that $J^*_{\Sigma,\text{con}}$ is the optimal continuous-time H_2 performance associated with the system Σ and, as before, denote the optimal sampled-data H_2 performance by $J^*_{\Sigma,\Delta}$. Is it true that $\lim_{\Delta\downarrow 0} J^*_{\Sigma,\Delta} = J^*_{\Sigma,\text{con}}$?

The first question above was studied before in [6, Thm. 4] using a different definition of H_2 performance and for the H_∞ performance criterion [6, Thm. 5]. In this section we will show that both questions have an affirmative answer.

Let Σ be given by (2.2). If the system Σ is controlled by a continuous-time compensator Γ_{con} given by the equations

$$\begin{aligned} \dot{w}(t) &= \bar{K}w(t) + \bar{L}y(t) , \\ u(t) &= \bar{M}w(t) + \bar{N}y(t) , \end{aligned} \tag{6.1}$$

with $w(t) \in \mathbb{R}^\ell$, then the associated closed-loop system $\Sigma \times \Gamma_{\text{con}}$ is given by

$$\begin{aligned} \dot{x}_e(t) &= A_e x_e(t) + E_e y(t) , \\ z(t) &= C_e x_e(t) , \end{aligned}$$

with

$$A_e = \begin{pmatrix} A + B\bar{N}C_1 & B\bar{M} \\ \bar{L}C_1 & \bar{K} \end{pmatrix}, \; E_e := \begin{pmatrix} E \\ 0 \end{pmatrix}, \; C_e := \begin{pmatrix} C_2 + D_2\bar{N}C_1 & D_2\bar{M} \end{pmatrix}.$$

If Γ_{con} is internally stabilizing, i.e., $\sigma(A_e) \subset \mathcal{C}^-$, then the H_2 performance of the closed-loop system $\Sigma \times \Gamma_{\text{con}}$ is equal to

$$J_\Sigma(\Gamma_{\text{con}}) = \text{tr}\ (E_e P_e E_e^{\text{T}}),$$

where P_e is the unique solution of the Lyapunov equation

$$A_e^{\text{T}} P_e + P_e A_e + C_e^{\text{T}} C_e = 0. \tag{6.2}$$

On the other hand, if the system Σ is controlled by the sampled-data controller $\Gamma = H_\Delta \Gamma_{\text{dis}} S_\Delta$, with Γ_{dis} given by (2.8), then the discrete-time closed-loop system $\Sigma_\Delta \times \Gamma_{\text{dis}}$ is given by the equations

$$\begin{aligned} x_{e,k+1} &= A_{e,\Delta} x_{e,k}\ E_{e,\Delta} y_k\ , \\ z_k &= C_{e,\Delta} x_{e,k}\ , \end{aligned}$$

with

$$A_{e,\Delta} = \begin{pmatrix} A_\Delta + B_\Delta N C_1 & B_\Delta M \\ L C_1 & K \end{pmatrix},\ E_{e,\Delta} := \begin{pmatrix} E_\Delta \\ 0 \end{pmatrix},$$

$$C_{e,\Delta} := \begin{pmatrix} C_{2,\Delta} + D_{2,\Delta} N C_1 & D_{2,\Delta} M \end{pmatrix}.$$

If Γ is internally stabilizing, equivalently $|\sigma(A_{e,\Delta})| < 1$, then the H_2 performance of the closed-loop system $\Sigma \times \Gamma$ is given by

$$J_{\Sigma,\Delta}(\Gamma) = \frac{1}{\Delta}\int_0^\Delta \int_0^{\Delta - s} \text{tr}\ \left(C_2 e^{tA} E E^{\text{T}} e^{tA^{\text{T}}} C_2^{\text{T}}\right) dt\, ds + \frac{1}{\Delta}\text{tr}\ (E_{e,\Delta} P_{e,\Delta} E_{e,\Delta}^{\text{T}}), \tag{6.3}$$

where $P_{e,\Delta}$ is the unique solution of the Lyapunov equation

$$A_{e,\Delta}^{\text{T}} P_{e,\Delta} A_{e,\Delta} - P_{e,\Delta} + C_{e,\Delta}^{\text{T}} C_{e,\Delta} = 0. \tag{6.4}$$

The following theorem shows that our first question above indeed has an affirmative answer:

THEOREM 6.1. *Let Γ_{con} be an internally stabilizing continuous-time compensator. For any $\Delta > 0$ define a discrete-time controller Γ_{dis} by $\Gamma_{\text{dis}} := S_\Delta \Gamma_{\text{con}} H_\Delta$, and let $\Gamma_\Delta := H_\Delta \Gamma_{\text{dis}} S_\Delta$ be the corresponding sampled-data controller with sampling period Δ. Then we have that there exists $\Delta_1 > 0$ such that for all $\Delta \notin \boldsymbol{\Delta}$ with $0 < \Delta < \Delta_1$, Γ_Δ is internally stabilizing. Furthermore,*

$$J_{\Sigma,\Delta}(\Gamma_\Delta) \to J_\Sigma(\Gamma_{\text{con}}) \qquad (\Delta \downarrow 0).$$

Proof. It is easily verified that $\Gamma_{\text{dis}} := S_\Delta \Gamma_{\text{con}} H_\Delta$ is described by the equations

$$\begin{aligned} w_{k+1} &= K_\Delta w_k + L_\Delta y_k\ , \\ u_k &= M w_k + N y_k\ , \end{aligned}$$

with $K_\Delta := e^{\bar{K}\Delta}$ and $L_\Delta := \int_0^\Delta e^{\bar{K}t} dt \bar{L}$. Thus we have

$$A_{e,\Delta} = \begin{pmatrix} A_\Delta + B_\Delta N C_1 & B_\Delta M \\ L_\Delta C_1 & K_\Delta \end{pmatrix}.$$

Note that $A_{e,\Delta} \to I$, the $(n+\ell)\times(n+\ell)$ identity matrix, and that $\frac{1}{\Delta}(A_{e,\Delta} - I) \to A_e$ $(\Delta \downarrow 0)$. Now we will first show that for Δ sufficiently small we have $|\sigma(A_{e,\Delta})| < 1$. Since A_e is stable, there exists $Q > 0$ such that $A_e^{\mathrm{T}}Q + QA_e < 0$. Now note that

$$\frac{1}{\Delta}(A_{e,\Delta}^{\mathrm{T}}QA_{e,\Delta} - Q) = \frac{1}{\Delta}(A_{e,\Delta}^{\mathrm{T}} - I)QA_{e,\Delta} + Q\frac{1}{\Delta}(A_{e,\Delta} - I).$$

Since the right-hand term converges to $A_e^{\mathrm{T}}Q + QA_e < 0$, for Δ sufficiently small we have $A_{e,\Delta}^{\mathrm{T}}QA_{e,\Delta} - Q < 0$. This implies that for Δ sufficiently small $A_{e,\Delta}$ is stable.

Next we show the convergence of the H_2 performance. For Δ sufficiently small we have $|\sigma(A_{e,\Delta})| < 1$, so the H_2 performance is given by (6.3), with $P_{e,\Delta}$ given by the Lyapunov equation (6.4). We shall prove that $P_{e,\Delta} \to P_e$, the unique solution of (6.2). For any Δ sufficiently small define a linear map $m_\Delta : \mathbb{R}^{n\times n} \to \mathbb{R}^{n\times n}$ by

$$m_\Delta(X) := \frac{1}{\Delta}A_{e,\Delta}^{\mathrm{T}}XA_{e,\Delta} - \frac{1}{\Delta}X.$$

Also define a linear map $m : \mathbb{R}^{n\times n} \to \mathbb{R}^{n\times n}$ by

$$m(X) := A_e^{\mathrm{T}}X + XA_e.$$

Note that m and m_Δ are all bijections. We can rewrite m_Δ as

$$m_\Delta(X) = \frac{1}{\Delta}(A_{e,\Delta}^{\mathrm{T}} - I)XA_{e,\Delta} + X\frac{1}{\Delta}(A_{e,\Delta} - I).$$

Recall that $A_{e,\Delta} \to I$ and $\frac{1}{\Delta}(A_{e,\Delta} - I) \to A_e$. Thus we see that $m_\Delta \to m$ $(\Delta \downarrow 0)$. Consequently, also $m_\Delta^{-1} \to m^{-1}$ $(\Delta \downarrow 0)$. Obviously, $P_{e,\Delta} = m_\Delta^{-1}(-\frac{1}{\Delta}C_{e,\Delta}^{\mathrm{T}}C_{e,\Delta})$. In addition, it follows from (2.6) that $\frac{1}{\Delta}C_{e,\Delta}^{\mathrm{T}}C_{e,\Delta} \to C_e^{\mathrm{T}}C_e$. This implies that $P_{e,\Delta} \to m^{-1}(C_e^{\mathrm{T}}C_e)$, which, in turn, is equal to P_e. By (2.5) we see that $\frac{1}{\Delta}E_{e,\Delta}E_{e,\Delta}^{\mathrm{T}} \to E_eE_e^{\mathrm{T}}$. Combining these facts we find that

$$\frac{1}{\Delta}\mathrm{tr}\;(E_{e,\Delta}E_{e,\Delta}^{\mathrm{T}}P_{e,\Delta}) \to \mathrm{tr}\;(E_eE_e^{\mathrm{T}}P_e).$$

Finally, it is immediate that

$$\frac{1}{\Delta}\int_0^\Delta\int_0^{\Delta-s} \mathrm{tr}\;\left(C_1e^{tA}EE^{\mathrm{T}}e^{tA^{\mathrm{T}}}C_1^{\mathrm{T}}\right)dt\,ds \to 0, \qquad \Delta \downarrow 0,$$

which completes the proof of the theorem. □

Now we turn to the second question posed above. In order to be able to answer this question, it is useful to consider this question first for the *linear quadratic problem.*

For this, consider the system $\dot{x}(t) = Ax(t) + Bu(t)$, $z(t) = C_2x(t) + D_2u(t)$. Assume that (A, B) is stabilizable. For a given static state feedback control law $u = Fx$ and initial state x_0, the output function is denoted by z_{F,x_0}. The linear quadratic problem is to minimize for each x_0 the cost-functional $J(x_0, F) := \int_0^\infty \|z_{F,x_0}(t)\|^2dt$ over all $F \in \mathbb{R}^{m\times n}$ such that $\sigma(A + BF) \subset C^-$. It is well known (see [9], [18]) that for each x_0 the optimal cost

$$J^*(x_0) := \inf\{J(x_0, F) \mid F \text{ s.t. } \sigma(A + BF) \subset C^-\} = x_0^{\mathrm{T}}Px_0,$$

where P is the largest real symmetric solution of the linear matrix inequality

$$(6.5)\qquad \begin{pmatrix} A^{\mathrm{T}}P + PA^{\mathrm{T}} + C_2^{\mathrm{T}}C_2 & PB + C_2^{\mathrm{T}}D_2 \\ B^{\mathrm{T}}P + D_2^{\mathrm{T}}C_2 & D_2^{\mathrm{T}}D_2 \end{pmatrix} \geq 0.$$

We want to compare this "normal" linear quadratic problem with its sampled-data version.

In the following, take a fixed sampling period $\Delta > 0$. The sampled-data version of the linear quadratic problem is to do the minimization over all stabilizing sampled-data static state feedback laws. More precisely, for a given $F \in \mathbb{R}^{m\times n}$ define the sampled-data state feedback control law $u = \mathcal{F}_\Delta x$ by $u(t) := Fx(k\Delta)$ ($t \in [k\Delta, (k+1)\Delta)$, $k = 0, 1, 2, \ldots$, or with a slight abuse of notation: $\mathcal{F}_\Delta = H_\Delta F S_\Delta$. For a given $\mathcal{F}_\Delta$ and initial state x_0, denote the output by $z_{\mathcal{F}_\Delta, x_0}$. Define the sampled-data cost functional in the obvious way, and denote it by $J(x_0, \mathcal{F}_\Delta)$. The control law $\mathcal{F}_\Delta$ is called internally stabilizing if for each initial state the controlled state trajectory $x(t)$ converges to 0 as $t \to \infty$. The sampled-data linear quadratic problem is to minimize for each x_0 $J(x_0, \mathcal{F}_\Delta)$ over all internally stabilizing control laws $\mathcal{F}_\Delta$. Let

$$J^*_\Delta(x_0) := \inf\{J(x_0, \mathcal{F}_\Delta) \mid \mathcal{F}_\Delta \text{ is internally stabilizing}\}$$

be the optimal cost. If no internally stabilizing $\mathcal{F}_\Delta$ exists, we define $J^*_\Delta(x_0) := \infty$ for all x_0. We will briefly explain here how the sampled-data linear quadratic can be resolved. First, note that for any $\mathcal{F}_\Delta = H_\Delta F S_\Delta$ we have

$$J(x_0, \mathcal{F}_\Delta) = \sum_{k=0}^{\infty} \int_{k\Delta}^{(k+1)\Delta} \|z_{\mathcal{F}_\Delta, x_0}(t)\|^2 dt.$$

Secondly, note that for all $t \in [k\Delta, (k+1)\Delta)$ we have $\dot{x}(t) = Ax(t) + Bu(t)$, $z_{\mathcal{F}_\Delta, x_0}(t) = C_2 x(t) + D_2 u(t)$, with $u(t) = Fx(k\Delta)$. Hence, on the interval $[k\Delta, (k+1)\Delta)$, x and u satisfy

$$\begin{pmatrix} \dot{x} \\ \dot{u} \end{pmatrix} = \begin{pmatrix} A & B \\ 0 & 0 \end{pmatrix} \begin{pmatrix} x \\ u \end{pmatrix},$$

with $u(k\Delta) = Fx(k\Delta)$. Consequently,

$$\begin{pmatrix} x(t) \\ u(t) \end{pmatrix} = e^{\underline{A}(t-k\Delta)} \begin{pmatrix} x(k\Delta) \\ Fx(k\Delta) \end{pmatrix}$$

for $t \in [k\Delta, (k+1)\Delta)$, with $\underline{A}$ defined by (2.7). Using this, it follows immediately from (2.6) that for $t \in [k\Delta, (k+1)\Delta)$ we have $\|z_{\mathcal{F}_\Delta, x_0}(t)\|^2 = \|C_{2,\Delta} x(k\Delta) + D_{2,\Delta} F x(k\Delta)\|^2$. Obviously, $x(k\Delta))$ evoluates according to $x((k+1)\Delta) = A_\Delta x(k\Delta) + B_\Delta F x(k\Delta)$. Hence we see that if $\mathcal{F}_\Delta = H_\Delta F S_\Delta$, then $J(x_0, \mathcal{F}_\Delta) = \sum_{k=0}^{\infty} \|(C_{2,\Delta} + D_{2,\Delta} F) x_k\|^2$, with $x_{k+1} = (A_\Delta + B_\Delta F) x_k$. It is also easily seen that $\mathcal{F}_\Delta$ is internally stabilizing if and only if $|\sigma(A_\Delta + B_\Delta F)| < 1$. Hence, $J^*_\Delta(x_0) < \infty$ for all x_0 if and only if (A_Δ, B_Δ) is stabilizable.

Consequently, we can make the following conclusion: the sampled-data linear quadratic problem under consideration is equivalent to the "normal" discrete-time linear quadratic problem of minimizing, for the system $x_{k+1} = A_\Delta x_k + B_\Delta u_k$, the cost functional $J_{\text{dis}}(x_0, F) := \sum_{k=0}^{\infty} \|(C_{2,\Delta} x_k + D_{2,\Delta} u_k\|^2$ over all $F \in \mathbb{R}^{m\times n}$ such that $|\sigma(A_\Delta + B_\Delta F)| < 1$. The latter problem was discussed in §3, remark (3.11) and §4, remark (4.6). By applying these results to the situation under consideration we can find a characterization of the optimal cost $J^*_\Delta(x_0)$ of the sampled-data linear quadratic problem:

LEMMA 6.2. *Let $\Delta > 0$ be such that (A_Δ, B_Δ) is stabilizable. Then for each x_0 we have*

$$J^*_\Delta(x_0) = x_0^{\mathrm{T}} P_\Delta x_0,$$

where P_Δ is the largest real symmetric solution of the algebraic Riccati equation

$$A_\Delta^{\mathrm{T}} P_\Delta A_\Delta - P_\Delta + C_{2,\Delta}^{\mathrm{T}} C_{2,\Delta} \tag{6.6}$$

$$-(C_{2,\Delta}^{\mathrm{T}} D_{2,\Delta} + A_\Delta^{\mathrm{T}} P_\Delta B_\Delta)(D_{2,\Delta}^{\mathrm{T}} D_{2,\Delta} + B_\Delta^{\mathrm{T}} P_\Delta B_\Delta)^{+}(D_{2,\Delta}^{\mathrm{T}} C_{2,\Delta} + B_\Delta^{\mathrm{T}} P_\Delta A_\Delta) = 0.$$

Now we will show that as $\Delta \downarrow 0$ the largest real symmetric solution P_Δ of (6.6) converges to P, the largest real symmetric solution of (6.5). We will prove this by proving that for each x_0 we have $J_\Delta^*(x_0) \to J^*(x_0)$. Note that if (A, B) is stabilizable, then for $\Delta > 0$ sufficienly small we have that (A_Δ, B_Δ) is stabilizable.

LEMMA 6.3. *Assume that (A, B) is stabilizable. Then there exists $\Delta_1 > 0$ such that for all $0 < \Delta < \Delta_1$, for all x_0 we have $J_\Delta^*(x_0) < \infty$. For all x_0 we have $\lim_{\Delta\downarrow 0} J_\Delta^*(x_0) = J^*(x_0)$. Also, for all $0 < \Delta < \Delta_1$, P_Δ exists and we have $\lim_{\Delta\downarrow 0} P_\Delta = P$.*

Proof. First of all note that for each sampling period Δ we have $J_\Delta^*(x_0) \geq J^*(x_0)$ for all x_0. This can be shown using that, in fact, for each x_0,

$$J^*(x_0) = \inf\left\{\int_0^\infty \|C_2 x(t) + D_2 u(t)\|^2 dt \mid u \text{ is such that } \lim_{t\to\infty} x(t) = 0\right\}.$$

Hence, by taking u to be generated by the internally stabilizing sampled-data control law $\mathcal{F}_\Delta$, it follows that $J(x_0, \mathcal{F}_\Delta) \geq J^*(x_0)$.

Now, let $\delta > 0$. Let F be such that $\sigma(A+BF) \subset \mathcal{C}^-$ and $J(x_0, F) < J^*(x_0) + \frac{\delta}{2}$. Clearly, $J(x_0, F) = x_0^{\mathrm{T}} L x_0$, where L is the unique solution of the Lyapunov equation

$$(A+BF)^{\mathrm{T}} L + L(A+BF) + (C_2 + D_2 F)^{\mathrm{T}}(C_2 + D_2 F) = 0.$$

Now consider the sampled-data control law $\mathcal{F}_\Delta = H_\Delta F S_\Delta$. By previous arguments, $J(x_0, \mathcal{F}_\Delta) = x_0^{\mathrm{T}} L_\Delta x_0$, where L_Δ is the unique solution of the Lyapunov equation

$$(A_\Delta + B_\Delta F)^{\mathrm{T}} L_\Delta (A_\Delta + B_\Delta F) - L_\Delta + (C_{2,\Delta} + D_{2,\Delta} F)^{\mathrm{T}}(C_{2,\Delta} + D_{2,\Delta} F) = 0.$$

Note that $A_\Delta + B_\Delta F \to I$, $\frac{1}{\Delta}(A_\Delta + B_\Delta F - I) \to A$, and $\frac{1}{\Delta}(C_{2,\Delta} + D_{2,\Delta} F)^{\mathrm{T}}(C_{2,\Delta} + D_{2,\Delta} F) \to (C_2 + D_2 F)^{\mathrm{T}}(C_2 + D_2 F)$ as $\Delta \downarrow 0$. Using a completely similar argument as in the proof of Theorem 6.1 we derive from this that $L_\Delta \to L$, which implies $J(x_0, \mathcal{F}_\Delta) \to J(x_0, F)$. Of course, we also have $J^*(x_0) \leq J_\Delta^*(x_0) \leq J(x_0, \mathcal{F}_\Delta)$. Combining this with $J(x_0, F) < J^*(x_0) + \frac{\delta}{2}$, we find that for δ sufficiently small we have $J^*(x_0) \leq J_\Delta^*(x_0) \leq J^*(x_0) + \delta$. Since δ was arbitrary, this proves the claim. The second statement in the formulation of the theorem is then immediate. □

Let $J_{\Sigma,\mathrm{con}}^*$ be the optimal continuous-time H_2 performance, i.e., the infimum of $J_\Sigma(\Gamma_{\mathrm{con}})$ over all internally stabilizing continuous-time compensators (6.1). It was shown in [15] that if (A, B) is stabilizable and (C_1, A) is detectable, then

$$J_{\Sigma,\mathrm{con}}^* = \mathrm{tr}\,(EE^{\mathrm{T}}P) + \mathrm{tr}\,((A^{\mathrm{T}}P + PA + C_2^{\mathrm{T}}C_2)Q), \tag{6.7}$$

where P is the largest real symmetric solution of the linear matrix inequality (6.5) and Q is the largest real symmetric solution of the dual linear matrix inequality

$$\begin{pmatrix} AQ + QA^{\mathrm{T}} + EE^{\mathrm{T}} & C_1^{\mathrm{T}}Q \\ QC_1 & 0 \end{pmatrix} \geq 0. \tag{6.8}$$

Let $J^*_{\Sigma,\Delta}$ be the optimal sampled-data H_2 performance. If $\Delta \in \boldsymbol{\Delta}$, then we define $J^*_{\Sigma,\Delta} := +\infty$. Our next theorem gives an affirmative answer to the second question posed in the introduction to this section.

THEOREM 6.4. *Let (A, B) be stabilizable and (C_1, A) be detectable. Then there exists Δ_1 such that for all $0 < \Delta < \Delta_1$, $J^*_{\Sigma,\Delta} < \infty$. We have $\lim_{\Delta\downarrow 0} J^*_{\Sigma,\Delta} = J^*_{\Sigma,\mathrm{con}}$.*

In the remainder of this section we will prove this theorem. First, recall the expression (5.1) for $J^*_{\Sigma,\Delta}$. Denote the first term in (5.1) by $I(\Delta)$. Then, under the conditions that (A, B) is stabilizable and (C_1, A) is detectable, we know that for $\Delta \notin \boldsymbol{\Delta}$

$$(6.9)\quad J^*_{\Sigma,\Delta} = I(\Delta) + \frac{1}{\Delta}\mathrm{tr}\,(E_\Delta E_\Delta^{\mathrm{T}} P_\Delta) + \frac{1}{\Delta}\mathrm{tr}\,((A_\Delta^{\mathrm{T}} P_\Delta A_\Delta - P_\Delta + C_{2,\Delta}^{\mathrm{T}} C_{2,\Delta})Q_\Delta)$$

$$-\frac{1}{\Delta}\mathrm{tr}\,((D_{P_\Delta} N_\Delta^* D_{Q_\Delta})(D_{P_\Delta} N_\Delta^* D_{Q_\Delta})^{\mathrm{T}}),$$

where P_Δ is the largest real symmetric solution of (6.6), Q_Δ is the largest real symmetric solution of the dual Riccati equation

$$(6.10)\qquad A_\Delta Q_\Delta A_\Delta^{\mathrm{T}} - Q_\Delta + E_\Delta E_\Delta^{\mathrm{T}} + A_\Delta Q_\Delta C_1^{\mathrm{T}} (C_1 Q_\Delta C_1^{\mathrm{T}})^+ C_1 Q_\Delta A_\Delta = 0,$$

and

$$N_\Delta^* = -D_{P_\Delta}(D_{P_\Delta}^+)^2 D_{P_\Delta} C_{P_\Delta} Q_\Delta C_1^{\mathrm{T}} (D_{Q_\Delta}^+)^2 D_{Q_\Delta}.$$

Here, C_{P_Δ}, D_{P_Δ}, and D_{Q_Δ} are defined by (3.5), (3.4), and (3.7), respectively, with $P = P_\Delta$ and $Q = Q_\Delta$. We will prove that $J^*_{\Sigma,\Delta} \to J^*_{\Sigma,\mathrm{con}}$ by analyzing the asymptotic behavior of the four terms appearing in (6.9) separately:

- It is immediate that the first term, $I(\Delta)$, converges to 0 as $\Delta \downarrow 0$.
- From (2.5) it follows that $\frac{1}{\Delta} E_\Delta E_\Delta^{\mathrm{T}} \to EE^{\mathrm{T}}$. Since also $P_\Delta \to P$, we conclude that the second term, $\frac{1}{\Delta}\mathrm{tr}\,(E_\Delta E_\Delta^{\mathrm{T}} P_\Delta)$, converges to $\mathrm{tr}\,(EE^{\mathrm{T}}P)$.
- To prove convergence of the third term, first note that $Q_\Delta \to Q$. This follows immediately by dualizing Lemma 6.3. Next, as before, rewrite

$$\frac{1}{\Delta}\mathrm{tr}\,(A_\Delta^{\mathrm{T}} P_\Delta A_\Delta - P_\Delta + C_{2,\Delta}^{\mathrm{T}} C_{2,\Delta})Q_\Delta)$$

$$(6.11)\qquad = \frac{1}{\Delta}(A_\Delta^{\mathrm{T}} - I)P_\Delta A_\Delta + P_\Delta \frac{1}{\Delta}(A_\Delta - I) + \frac{1}{\Delta} C_{2,\Delta}^{\mathrm{T}} C_{2,\Delta}.$$

  Since $\frac{1}{\Delta}(A_\Delta - I) \to A$, $A_\Delta \to I$, and $\frac{1}{\Delta} C_{2,\Delta}^{\mathrm{T}} C_{2,\Delta} \to C_2^{\mathrm{T}} C_2$, we conclude that the third term in (6.9) converges to $\mathrm{tr}\,(A^{\mathrm{T}}P + PA + C_2^{\mathrm{T}} C_2)$.
- In order to complete the proof of Theorem 6.4, we should hence prove that the fourth term in (6.9) converges to 0 as $\Delta \downarrow 0$. This is done in the following lemma:

LEMMA 6.5. $\frac{1}{\Delta}\mathrm{tr}\,((D_{P_\Delta} N_\Delta^* D_{Q_\Delta})(D_{P_\Delta} N_\Delta^* D_{Q_\Delta})^{\mathrm{T}}) \to 0$ *as* $\Delta \downarrow 0$.

Proof. Rewrite the fourth term in (6.9) as $\frac{1}{\Delta}\|D_{P_\Delta} N_\Delta^* D_{Q_\Delta}\|^2$, where for any matrix M, $\|M\|$ denotes the Frobenius norm $\mathrm{tr}\,(MM^{\mathrm{T}})$. Note that if M is a given matrix, then M^+M and MM^+ are orthogonal projectors, so consequently $\|MM^+\| = \|MM^+\| = \mathrm{rank}\,(M)$. In particular, this implies that if M is $n \times n$ matrix, then $\|MM^+\| = \|MM^+\| \le n$. Now make the following estimates:

$$\frac{1}{\Delta}\|D_{P_\Delta} N_\Delta^* D_{Q_\Delta}\|^2$$

$$
\begin{aligned}
&\leq \frac{1}{\Delta}\|(D_{P_\Delta}D_{P_\Delta}^{+})(D_{P_\Delta}^{+}D_{P_\Delta})C_{P_\Delta}Q_\Delta C_1^{\mathrm{T}}D_{Q_\Delta}^{+}(D_{Q_\Delta}^{+}D_{Q_\Delta})\|^2\\
&\leq \frac{m^4p^2}{\Delta}\|C_{P_\Delta}Q_\Delta C_1^{\mathrm{T}}D_{Q_\Delta}^{+}\|^2\\
&\leq \frac{m^4p^2}{\Delta}\|C_{P_\Delta}\|^2\|Q_\Delta C_1^{\mathrm{T}}D_{Q_\Delta}^{+}\|^2.
\end{aligned}
$$

As noted before, $C_{P_\Delta}^{\mathrm{T}}C_{P_\Delta} = A_\Delta^{\mathrm{T}}P_\Delta A_\Delta - P_\Delta + C_{2,\Delta}^{\mathrm{T}}C_{2,\Delta}$, so $\frac{1}{\Delta}\|C_{P_\Delta}\|^2 \to \mathrm{tr}\,(A^{\mathrm{T}}P + PA + C_2^{\mathrm{T}}C_2)$. On the other hand, by noting that Q_Δ satisfies the Riccati equation (6.10), where $A_\Delta = e^{A\Delta}$ is invertible, we see that

$$
\begin{aligned}
&\|Q_\Delta C_1^{\mathrm{T}}D_{Q_\Delta}^{+}\|^2\\
&\quad = \mathrm{tr}\,(Q_\Delta C_1^{\mathrm{T}}(C_1Q_\Delta C_1)^{+}C_1Q_\Delta)\\
&\quad = \mathrm{tr}\,(Q_\Delta - A_\Delta^{-1}Q_\Delta A_\Delta^{-T} + A_\Delta^{-1}E_\Delta E_\Delta^{\mathrm{T}}A_\Delta^{-\mathrm{T}}).
\end{aligned}
$$

Since $Q_\Delta \to Q$, $A_\Delta^{-1} \to I$ and $E_\Delta E_\Delta^{\mathrm{T}} \to 0$, the latter converges to zero as $\Delta \downarrow 0$. □

REFERENCES

[1] B. Bamieh and J. Pearson, *A general framework for linear periodic systems with application to H_∞ sampled-data control*, IEEE Trans. Autom. Control, 37 (1992), pp. 418–435.

[2] ———, *The H_2 problem for sampled data systems*, Systems and Control Letters, 19 (1992), pp. 1–12.

[3] B. Bamieh, J. Pearson, B. Francis, and A. Tannenbaum, *A lifting technique for linear periodic systems with applications to sampled-data control*, Systems Control Lett., 17 (1991), pp. 79–88.

[4] T. Chen, *A simple derivation of the H_2-optimal sampled-data controllers*, Preprint, University of Calgary, June 1992.

[5] T. Chen and B. Francis, *On the L_2 induced norm of a sampled-data system*, Systems Control Lett., 15 (1990), pp. 211–219.

[6] ———, *H_2-optimal sampled data control*, IEEE Trans. Autom. Control, AC-36 (1991), pp. 387–397.

[7] ———, *Linear time-varying H_2 optimal control of sampled-data systems*, Automatica, 27 (1991), pp. 963–974.

[8] B. Francis and T. Georgiou, *Stability theory for linear time-invariant plants with periodic digital controllers*, IEEE Trans. Autom. Control, 33 (1988), pp. 820–832.

[9] M. Hautus and L. Silverman, *System structure and singular control*, Linear Algebra Appl., 50 (1983), pp. 369–402.

[10] P. Kabamba and S. Hara, *On computing the induced norm of a sampled data system*, in Proc. American Control Conference, San Diego, CA, 1990, pp. 319–320.

[11] ———, *Worst case analysis and design of sampled data control systems*, Preprint, 1990.

[12] P. Khargonekar and N. Sivashankar, *H_2 optimal control for sampled-data systems*, Systems Control Lett., 17 (1992), pp. 425–436.

[13] Y. H. R. Kalman and K. Narendra, *Controllability of linear dynamical systems*, in Contributions to Differential Equations, vol. 1, Interscience, New York, 1963.

[14] L. Silverman, *Discrete Riccati equations*, in Control and Dynamic Systems, Advances in Theory and Applications, C. Leondes, ed., Academic Press, New York, 1976, pp. 313–386.

[15] A. Stoorvogel, *The singular H_2 control problem*, Automatica, 28 (1992), pp. 627–631.

[16] A. Stoorvogel and J. van der Woude, *The disturbance decoupling problem with measurement feedback and stability for systems with direct feedthrough matrices*, Systems Control Lett., 17 (1991), pp. 217–226.

[17] H. Toivonen, *Sampled-data control of continuous-time systems with an H_∞ optimality criterion*, Automatica, 28 (1992), pp. 45–54.

[18] J. Willems, A. Kitapci, and L. Silverman, *Singular optimal control, a geometric approach*, SIAM J. Control Optim., 24 (1986), pp. 323–337.

[19] W. M. Wonham, *Linear Multivariable Control: a Geometric Approach*, Springer-Verlag, New York, 1979.

[20] Y. YAMAMOTO, *New approach to sampled-data systems: a function space method*, in Proc. 29th CDC, 1990, pp. 1882–1887.

[21] ———, *On the state space and frequency domain characterization of H_∞-norm of sampled-data systems*, Preprint, 1991.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 33, No. 3, pp. 863–881, May 1995

© 1995 Society for Industrial and Applied Mathematics
010

A GEOMETRIC APPROACH TO THE MINIMUM SENSITIVITY DESIGN PROBLEM [*]

ERIK I. VERRIEST[†] AND W. STEVEN GRAY[‡]

Abstract. In this paper a Riemannian geometric framework is given for the minimum sensitivity design problem and its solution using a natural optimization criterion. The theory is then applied to the case of linear systems to generate a class of minimum sensitivity realizations related to the so-called balanced realizations. In particular, conditions are given which are applicable in the cases of fixed point and floating point implementations.

Key words. sensitivity, realization theory, Riemannian geometry, balanced realizations

AMS subject classifications. 93B35, 53B21

1. Introduction.

In this paper we consider the problem of selecting parameterizations for a given mathematical model which minimize the sensitivity of the model's behavior to perturbations in the parameter values. We call this problem the *minimum sensitivity design problem.* It often arises in applied mathematics and engineering, and has received a great deal of attention in a variety of applications. For example, given a digital linear filter (model structure) with some fixed input-output relationship (model's behavior), a common question is how to find an internal representation (parameterization) which minimizes the sensitivity of the input-output relationship to quantization (perturbations) of the filter coefficients (parameter values). This problem differs from the parameter identification problem in that one assumes at least one parameterization is already available. The specific context introduced for solving this problem is Riemannian geometry. The primary motivations for a geometric approach are the generality of the synthesis techniques and the success of such an approach in the area of parametric system identification [13]–[15].

Once developed, the primary focus is on applying the technique to the synthesis of minimum sensitivity state space models for dynamical systems. Only linear systems are considered in this paper. The performance index derived in the geometric context is related to (but distinct from) those found via frequency domain techniques in [21], [24], [25], [26], [30]. The main difference is that the performance index introduced here gives a bound on the sensitivity of the impulse response in terms of a mixed ℓ_∞/ℓ_2 norm rather than on the sensitivity of the frequency response using a mixed L_1/L_2 norm with respect to the realization coefficients. Both approaches compliment each other and in certain ways are closely related. A strength of the geometric method, however, is that it generalizes easily to other classes of systems where frequency domain interpretations are not as tractable [9], [11], [12], [27]. Practical applications of this research include the synthesis of state space realizations with superior coefficient quantization properties and the synthesis of analog networks with minimum component sensitivity.

[*] Received by the editors October 2, 1991; accepted for publication (in revised form) December 2, 1993. This research was supported in part by U.S. Air Force contracts FO8635-84-C-0273 and AFOSR-87-0308.

[†] School of Electrical and Computer Engineering, Georgia Institute of Technology, Atlanta, Georgia 30332-0250; and Georgia Tech Lorraine, Technôpole Metz 2000, F-57070 Metz, France (erik.verriest@ee.gatech.edu) and (verriest@gtl.georgiatech-metz.fr).

[‡] Department of Electrical and Computer Engineering, Drexel University, Philadelphia, Pennsylvania 19104-2884 (steven_gray@ece.drexel.edu).

The paper is composed of two main sections. The goal of the first section is to motivate and describe a geometric framework for solving the minimum sensitivity design problem that leads to a general optimality equation. It is then demonstrated in the second section how this theory can be used in the synthesis of minimum sensitivity state space realizations of discrete-time linear time-invariant systems.

2. The minimum sensitivity design problem. The goal of this section is to describe the minimum sensitivity design problem in a differential geometric setting. The main intuitive idea behind the method is as follows. Given the set of all admissible realizations for a given mathematical model, we first partition this set into equivalence classes, where two realizations are said to be equivalent if they map to the same model behavior. Since the applications which follow will deal specifically with input-output systems, the model behavior to which we refer is the (output) response to some fixed input. For linear systems, the behavior we are usually interested in is the impulse response. Generally, each equivalence class will contain more than one realization. Therefore, the idea is to define a sensitivity measure which reflects how perturbations of realization coefficients of a given realization perturb the model behavior. Within a fixed equivalence class, we then designate those realizations which have minimal sensitivity as the ones solving the minimum sensitivity design problem.

2.1. Abstract realization spaces. We consider a realization θ to be an ordered set of real numbers denoted by $\theta = (\theta^1, \theta^2, \ldots, \theta^n)$. It will be assumed that the system under consideration is completely specified by such a vector θ. Since these parameters are real attributes of the system, it is natural to assume that what can be known exactly is their membership to certain intervals. Therefore, the ambient space in which a realization exists is the cartesian product space $\Re^n$, and the usual topology (generated by cartesian products of open intervals) is the natural topology it should be endowed with. It is not natural to assume any other structure at this point. In particular, since neither the sum of two realizations nor the scalar multiple of a realization may have any physical significance, a *global* vector space structure is not naturally associated with it. Thus, the collection of all realizations will be considered as an affine space, denoted by $\mathcal{A}^n$, with a *local* vector space structure isomorphic to the usual vector space $\mathcal{V}^n$ on $\Re^n$. Therefore, there is no assumed relation between the vectors belonging to the vector spaces attached at different points. We will be interested in an open subset of $\mathcal{A}^n$ referred to as either the set of *admissible realizations* or the *realization space*, and denoted by Θ. Typically, but not exclusively, open sets arise as complements of inverse images by continuous maps.

Now we introduce perturbations in the parameters. If these perturbations must be quantified in any way one needs a metric structure, at least locally at each point (i.e., in the local vector space attached at the nominal parametrization) in order to discuss perturbations of fixed norm, and also to allow for the notion of orthogonality. The latter also allows one to talk about the *directions* of perturbations. If Θ is paracompact, then a Riemannian structure g can be put on Θ (or, more exactly, its tangent bundle). This means that for each $\theta \in \Theta$, it follows that $g(\theta)$ is a symmetric, positive definite, nondegenerate bilinear form defined on $T_\theta\Theta \times T_\theta\Theta$ with a local representation

$$g_{ij}(\theta) = g(\theta)\left(\frac{\partial}{\partial\theta^i}, \frac{\partial}{\partial\theta^j}\right), \tag{1}$$

where $\partial/\partial\theta^i$ $i = 1, 2, \ldots, n$ are the canonical basis vectors for the tangent space $T_\theta\Theta$ and each g_{ij} is a smooth function of θ. Then $g : \Theta \mapsto T_2^0\Theta$ corresponds to the

Riemannian metric on Θ, where $T_2^0\Theta$ denotes the bundle of covariant two-tensors defined on Θ. We stress again that we are not using global vector space structures, but only local ones to describe the perturbations with respect to a given nominal point. The following definition summarizes what we mean by a realization space.

DEFINITION 2.1. *A* realization space *in $\mathcal{A}^n$ is a smooth Riemannian manifold (Θ, g), where Θ is an open subset of the affine space $\mathcal{A}^n$ and g is a nondegenerate Riemannian metric on $T\Theta \times T\Theta$.*

In what follows, "g" will be understood if we speak about the realization space Θ.

Now let $T_\theta^*\Theta$ denote the dual space of the tangent space at each $\theta \in \Theta$, and let

$$\flat : T_\theta\Theta \mapsto T_\theta^*\Theta : v \mapsto g(\theta)(v, \cdot), \tag{2}$$

$$\sharp : T_\theta^*\Theta \mapsto T_\theta\Theta : g(\theta)(v, \cdot) \mapsto v \tag{3}$$

denote the natural isomorphism that the metric tensor g induces between the two spaces. Denoting the canonical basis for $T_\theta^*\Theta$ by $\{d\theta^1, d\theta^2, \ldots, d\theta^n\}$, the metric tensor can be represented in the form

$$g = g_{ij} d\theta^i \otimes d\theta^j, \tag{4}$$

where $\otimes$ denotes the usual tensor product and the Einstein summation convention is assumed. Furthermore, if f is a smooth real-valued function defined on Θ, then $df(\theta) \in T_\theta^*\Theta$, and hence can be written as the linear combination

$$df(\theta) = \left.\frac{\partial f}{\partial \theta^i}\right|_\theta d\theta^i. \tag{5}$$

The gradient of f is defined as the vector field given by

$$\nabla f : \Theta \mapsto T_\theta\Theta : \theta \mapsto (df(\theta))^\sharp = g^{ij}(\theta) \left.\frac{\partial f}{\partial \theta^i}\right|_\theta \frac{\partial}{\partial \theta^j}, \tag{6}$$

where $[g^{ij}]$ is the matrix inverse of the metric tensor $[g_{ij}]$.

Within a given realization space (Θ, g), we wish to form equivalence classes which reflect some commonality between realizations based on their associated model behavior. To this end, we adapt a notion from quantum mechanics described in the following definition (see, for example, [1]).

DEFINITION 2.2. *Let f be a smooth real-valued function defined on a realization space (Θ, g) such that the gradient ∇f is nonzero everywhere on Θ. Then, f will be called an* observable *over Θ, and the scalar $f(\theta)$, where $\theta \in \Theta$, will be referred to as the value of the observable at θ.*

The significance of an observable is that two realizations are considered equivalent relative to f if they yield the same observable value, i.e., $\theta \sim \hat{\theta}$ for $\theta, \hat{\theta} \in \Theta$ if and only if $f(\theta) = f(\hat{\theta})$. In a system theory context, an observable could be the evaluation of the transfer function for an input-output system evaluated at a particular frequency or the impulse response at a specific instant of time. Observables are also referred to as *system functions* [8]. In a topological sense, observables are smooth, regular canonical projections which map Θ to some subset of the real numbers. The assumption on the lack of critical points is technical but essential in view of the lemmas discussed below.

We designate an equivalence class on Θ relative to f and corresponding to the value $k \in f(\Theta)$ as

$$M_k(f) = f^{-1}(k), \tag{7}$$

and the set of all such classes, i.e., the quotient space, as $\Theta/\sim$. The quotient space of minimal state space realizations of linear time-invariant systems has received a great deal of attention in the system identification problem (see [13]–[15], [19]). Our interest, however, is not in the quotient space, but rather the geometric and topological aspects of *individual* equivalence classes. The following lemmas give the geometric structure of an equivalence class and the quotient space.

LEMMA 2.3 (see [6], [20]). *Let f be an observable on realization space (Θ, g) and $k \in f(\Theta)$. Then, the equivalence class $M_k(f) = f^{-1}(k)$ is a Riemannian submanifold of Θ with codimension equal to one.*

LEMMA 2.4 (see [6], [20]). *Let f be an observable on realization space (Θ, g). Then, $\Theta/\sim$ is a smooth foliation on Θ with codimension equal to one.*

An observable thus induces a decomposition of the realization space into connected submanifolds of dimension $n-1$, usually called the *leaves* of the foliation, which stack up locally like subsets of $\Re^{n-1} \times \Re$, where the second coordinate (i.e., k in the above discussion) is held fixed. Perturbations in the parameters induce perturbations in the value of the observable. Among all perturbations of fixed norm (induced by the Riemannian metric), the worst-case perturbation is the perturbation that induces the maximal change in the observable. The minimum sensitivity design problem is to determine which point in Θ (i.e., realizations) on a fixed leaf, when perturbed, will have the smallest worst-case perturbation of the corresponding observable value.

2.2. Extremal sensitivity points. Since each leaf of the foliation induced by an observable function f is itself a Riemannian manifold, it makes sense to talk about the tangent space to a fixed leaf. Furthermore, from the point of view that each leaf is really a *level surface* of the observable function f, the tangent space of a leaf $M_k(f)$ at a fixed point θ can be described by the following subset of the tangent space $T_\theta\Theta$:

$$\begin{aligned} T_\theta M_k(f) &= \{v \in T_\theta\Theta : df(\theta)(v) = 0\} \\ &= \{v \in T_\theta\Theta : g(\theta)(\nabla f, v) = 0\}. \end{aligned} \tag{8}$$

In view of the representation in (8), it follows that $\nabla f \mid_{M_k(f)}$, the restriction of ∇f to $M_k(f)$, is a normal vector field defined on the leaf corresponding to value k. Since for any perturbation $\Delta\theta \in T_{\theta_0}\mathcal{H}$ we have

$$f(\theta) = f(\theta_0) + g(\theta_0)(\nabla f(\theta_0), \Delta\theta), \tag{9}$$

it follows that the worst perturbation of f is obtained for $\Delta\theta$ proportional to the gradient ∇f. Clearly the induced *worst-case deviation* for a unit pertubation vector will be $\| \nabla f \|$. Hence, the realizations in $M_k(f)$ which have extremal sensitivity are those which minimize or maximize the norm of the vectors in this vector field, where the norm function is defined as

$$\| \cdot \| : T_\theta\Theta \mapsto \Re : v \mapsto \sqrt{g(\theta)(v,v)}. \tag{10}$$

Consequently, we have the following more convenient definition.

DEFINITION 2.5. *A realization $\theta^* \in M_k(f)$ is an* extremal sensitivity *point of $M_k(f)$ if θ^* extremizes the performance index*

$$L(\theta) = \frac{1}{2} \parallel \nabla f(\theta) \parallel^2 \tag{11}$$

over the manifold $M_k(f)$.

As an immediate consequence of this definition we get the following theorem, specifying a necessary condition for determining an extremal sensitivity point.

THEOREM 2.6 (see [28]). *If a realization $\theta^* \in M_k(f)$ is an extremal sensitivity point, then*

$$\nabla H(\theta^*) = \frac{1}{2}\nabla g(\theta^*)(\nabla f(\theta^*), \nabla f(\theta^*)) - \lambda \nabla f(\theta^*) = 0, \tag{12}$$

where H is the scalar-valued Hamiltonian function

$$H(\theta) = \frac{1}{2} g(\theta)(\nabla f(\theta), \nabla f(\theta)) - \lambda f(\theta). \tag{13}$$

Proof. The stated condition is the Euler–Lagrange equation for the constrained optimization problem. □

In order to use Theorem 2.6 in a specific problem, one needs a local representation of optimality equation (12), which interestingly has considerable structure not immediately apparent. Consider the following definition and related lemmas.

DEFINITION 2.7 (see [7]). *Let f be an observable on realization space (Θ, g). Then, the* Hessian operator *of f at $\theta \in \Theta$ is the linear operator defined by*

$$\nabla^2 f(\theta) : T_\theta \Theta \mapsto T_\theta \Theta : u \mapsto \nabla_u \nabla f(\theta) \triangleq \nabla^2 f(\theta)(u), \tag{14}$$

where ∇ denotes the unique, symmetric (Riemannian) affine connection compatible with the metric g, i.e., for $\theta \in \Theta$,

$$\nabla : T_\theta \Theta \times T_\theta \Theta \mapsto T_\theta \Theta : (u, v) \mapsto \nabla_u v = (u^i v^j \Gamma^k_{ij} + u(v^k)) \frac{\partial}{\partial \theta^k} \tag{15}$$

with $u = u^i(\partial/\partial\theta^i)$, $v = v^i(\partial/\partial\theta^i)$, and the Christoffel symbols of the Riemannian connection specified by

$$\Gamma^k_{ij} = \frac{1}{2}\left\{\frac{\partial}{\partial \theta^i} g_{jm} + \frac{\partial}{\partial \theta^j} g_{mi} - \frac{\partial}{\partial \theta^m} g_{ij}\right\} g^{mk}. \tag{16}$$

LEMMA 2.8 (see [7]). *The Hessian operator has the following properties:*

(i) *$\nabla^2 f(\theta)$ is self-adjoint;*

(ii) *the bilinear form $\nabla^2 f(\theta)(u, v) \triangleq g(\theta)(\nabla^2 f(\theta)(u), v)$ is symmetric;*

(iii) *in local coordinates,*

$$\nabla^2 f(\theta)(v) = g^{jk}\left[\frac{\partial^2 f}{\partial \theta^i \partial \theta^j} - \Gamma^\ell_{ij} \frac{\partial f}{\partial \theta^\ell}\right] v^i \frac{\partial}{\partial \theta^k}. \tag{17}$$

LEMMA 2.9. *The optimality equation in Theorem 2.6 is equivalent to*

$$(\nabla^2 f(\theta^*) - \lambda I)(\nabla f(\theta^*)) = 0, \tag{18}$$

where I is the identity operator on $T_{\theta^*}\mathcal{H}$. That is, the gradient of f at θ^* must be in the eigenspace of the Hessian operator of f at θ^*.

Proof. Assuming all vector fields below are evaluated at $\theta = \theta^*$, and using the property

$$u(g(\theta)(v, w)) = g(\theta)(\nabla_u v, w) + g(\theta)(v, \nabla_u w) \tag{19}$$

for $u, v, w \in T_\theta\Theta$, it follows that:

$$\begin{aligned}
\nabla H &= \frac{1}{2}\nabla g(\nabla f, \nabla f) - \lambda\nabla f \\
&= \frac{1}{2}(dg(\nabla f, \nabla f))^{\#} - \lambda\nabla f \\
&= \frac{1}{2}\left(\frac{\partial g}{\partial\theta^i}(\nabla f, \nabla f)d\theta^i\right)^{\#} - \lambda\nabla f \\
&= \frac{1}{2}((g(\nabla_{\frac{\partial}{\partial\theta^i}}\nabla f, \nabla f) + g(\nabla f, \nabla_{\frac{\partial}{\partial\theta^i}}\nabla f))d\theta^i)^{\#} - \lambda\nabla f \\
&= (g(\nabla f, \nabla_{\frac{\partial}{\partial\theta^i}}\nabla f)d\theta^i)^{\#} - \lambda\nabla f \\
&= \left(g\left(\nabla_{\nabla f}\nabla f, \frac{\partial}{\partial\theta^i}\right)d\theta^i\right)^{\#} - \lambda\nabla f \\
&= \nabla_{\nabla f}\nabla f - \lambda\nabla f \\
&= \nabla^2 f(\nabla f) - \lambda\nabla f = 0. \qquad \square
\end{aligned}$$

Using the local representation of $\nabla^2 f(\theta)$ given in Lemma 2.8, it follows immediately that equation (18) has the local form

$$\left(g^{jk}\left[\frac{\partial^2 f}{\partial\theta^i\partial\theta^j} - \Gamma^\ell_{ij}\frac{\partial f}{\partial\theta^\ell}\right] - \lambda\delta^k_i\right)g^{mi}\frac{\partial f}{\partial\theta^m}\frac{\partial}{\partial\theta^k} = 0. \tag{20}$$

Then by substitution of either equation (16) for Γ^k_{ij} above or a direct computation of $\nabla g(\nabla f, \nabla f)$ in local coordinates, it also follows that an equivalent expression is

$$\left(g^{jk}\left[\frac{\partial^2 f}{\partial\theta^i\partial\theta^j} + \frac{1}{2}g_{is}\frac{\partial g^{ts}}{\partial\theta^j}\frac{\partial f}{\partial\theta^t}\right] - \lambda\delta^k_i\right)g^{mi}\frac{\partial f}{\partial\theta^m}\frac{\partial}{\partial\theta^k} = 0. \tag{21}$$

This equation (in terms of only the observable, the metric tensor, and its inverse) yields a matrix representation which is far more amenable for computations and thus motivates the following definition.

DEFINITION 2.10. *Let f be an observable on realization space (Θ, g). Then, the* nonsymmetric Hessian operator *of f at $\theta \in \Theta$ is the linear operator defined in local coordinates by*

$$\hat{\nabla}^2 f(\theta) : T_\theta\Theta \mapsto T_\theta\Theta : u \mapsto g^{jk}\left[\frac{\partial^2 f}{\partial\theta^i\partial\theta^j} + \frac{1}{2}g_{is}\frac{\partial g^{ts}}{\partial\theta^j}\frac{\partial f}{\partial\theta^t}\right]u^i\frac{\partial}{\partial\theta^k}, \tag{22}$$

where $u = u^i(\partial/\partial\theta^i)$.

LEMMA 2.11. *Let f be an observable on realization space (Θ, g). Then for any $\theta \in \Theta$,*

$$\hat{\nabla}^2 f(\theta)(\nabla f(\theta)) = \nabla^2 f(\theta)(\nabla f(\theta)). \tag{23}$$

Proof. The lemma is proved by direct substitution. □

Equation (21) expressed in terms of the Vetter derivative and standard matrix calculus operators [2], [3], [4], [5] has the form

$$\left(G^{-1}\frac{\partial^2 f}{\partial\theta^2}+\frac{1}{2}G^{-1}\left(I_n\otimes\frac{\partial f}{\partial\theta^T}\right)\left(\frac{\partial}{\partial\theta}\otimes G^{-1}\right)G-\lambda I_n\right)G^{-1}\frac{\partial f}{\partial\theta}=0, \tag{24}$$

where G denotes a matrix with entries g_{ij}, I_n is an $n\times n$ identity matrix, and $\otimes$ represents the Kronecker product. We denote by $\partial f/\partial\theta$ and $\partial^2 f/\partial\theta^2$, respectively, the column vector of first-order partial derivatives of f with respect to the components of θ, and the matrix formed by the second-order partial derivatives. These are indeed the gradient and Hessian if the metric tensor is the identity, but in the general context they have no direct meaning. For this special case of the uniform metric $G = I_n$, we get a simplified version of the optimality condition which has a matrix notation representation

$$\left(\frac{\partial^2 f}{\partial\theta^2}-\lambda I_n\right)\frac{\partial f}{\partial\theta}=0. \tag{25}$$

Another matrix form may also be obtained which is entirely equivalent to equation (21), but avoids the use of Kronecker products.

DEFINITION 2.12. *Let f be an observable on realization space (Θ, g). Define the* pseudogradient *as*

$$df \triangleq \left(\frac{\partial f}{\partial\theta}\right)^T\cdot G^{-\frac{1}{2}} \tag{26}$$

and the pseudo-Hessian *matrix as*

$$d^2 f \triangleq G^{-\frac{1}{2}}\frac{\partial^2 f}{\partial\theta^2}G^{-\frac{1}{2}}+\frac{1}{2}G^{\frac{1}{2}}\left[\frac{\partial G^{-1}}{\partial\theta^1}\cdot\frac{\partial f}{\partial\theta},\ldots,\frac{\partial G^{-1}}{\partial\theta^n}\cdot\frac{\partial f}{\partial\theta}\right]G^{-\frac{1}{2}}. \tag{27}$$

The extremal sensitivity criterion has a form analogous to equation (24), but uses the pseudogradient and the pseudo-Hessian rather than the gradient and the nonsymmetric Hessian. If the realization space (Θ, g) is such that the parameters are functionally independent of each other, except for the constraint $f(\theta) = k$, then the perturbation of the parameters should be considered independently of each other. Hence, in this case we set $g_{ij} = 0$ whenever $i \neq j$. Moreover, if the perturbation metric for one parameter depends only on the nominal value of that parameter and not the others, then we call the realization space (Θ, g) a *realization space of independent parameter design.* It turns out that in this case the problem is *symmetric*, and the pseudogradient and pseudo-Hessian are, respectively,

$$df = \left(\frac{\partial f}{\partial\theta}\right)^T \operatorname{diag}\left(\frac{1}{\sqrt{g_{ii}}}\right), \tag{28}$$

and

$$d^2 f = \operatorname{diag}\left(\sqrt{g^{ii}}\right)\frac{\partial^2 f}{\partial\theta^2}\operatorname{diag}\left(\sqrt{g^{ii}}\right)+\operatorname{diag}\left(\frac{\partial\sqrt{g^{ii}}}{\partial\theta^i}\right)\operatorname{diag}(df), \tag{29}$$

where, now, $g^{ii} = 1/g_{ii}$.

While studying the sensitivity of realizations in the next section, it will be desirable not only to characterize extremal sensitivity realizations, but also to show explicitly that certain sets have *minimum* sensitivity. One can easily test for this property by examining the definiteness of the bilinear form associated with the Hessian of the Hamiltonian function H defined in (13). This test is posed as a matrix definiteness problem in the following lemma.

LEMMA 2.13. *An extremal sensitivity point $\theta^* \in M_k(f)$ has the minimum sensitivity property if the matrix*

$$\frac{\partial^2 H}{\partial \theta^2} + \frac{1}{2}\left(I_n \otimes \frac{\partial H}{\partial \theta^T}\right)\left(\frac{\partial}{\partial \theta} \otimes G^{-1}\right) G \tag{30}$$

is positive definite on the tangent space of $M_k(f)$ at θ^*, where $H = (1/2)g^{st}(\partial f/\partial \theta^s)$ $(\partial f/\partial \theta^t) - \lambda f$.

Proof. The definiteness of the Hessian of H always determines the nature of each extremal. Equation (30) is simply a matrix representation of the bilinear form $\nabla^2 H(\theta^*)(u, v)$ in local coordinates. □

Writing (30) in fully expanded form is simple to do in principle, but tedious in any notation. The special case of the uniform metric $g_{ij} = \delta_{ij}$, however, is significantly simpler and will be used in later development:

$$[\nabla^2 H] = \left[\sum_m \frac{\partial^3 f}{\partial \theta^m \partial \theta^i \partial \theta^j}\frac{\partial f}{\partial \theta^m}\right] + \left[\frac{\partial^2 f}{\partial \theta^i \partial \theta^j}\right]^2 - \lambda\left[\frac{\partial^2 f}{\partial \theta^i \partial \theta^j}\right]. \tag{31}$$

2.3. Multiple observables. We now consider extensions of the notion of an observable. In the applications which follow, the need arises for a method which will allow one to consider the sensitivity of several observables *simultaneously*. For example, one may wish to define as observables the impulse response of a system at r specific time instances where $r > 1$. The following extension of Definition 2.2 addresses this requirement.

DEFINITION 2.14. *Let (Θ, g) be a realization space and $F : \Theta \mapsto \Re^{p \times m}$ be a smooth function whose differential $dF(\theta)$ is surjective for all $\theta \in \Theta$. Then, F will be called a* (matrix) observable *over Θ, and the matrix $F(\theta)$ will be referred to as its value at θ.*

We have the following results, which are analogous to Lemmas 2.3 and 2.4.

LEMMA 2.15 (see [6], [20]). *Let $F : \Theta \mapsto \Re^{p \times m}$ be a matrix observable on realization space (Θ, g) and $K \in F(\Theta)$. Then, the equivalence class $M_K(F) = F^{-1}(K)$ is a Riemannian submanifold of Θ with codimension equal to pm.*

LEMMA 2.16 (see [6], [20]). *An observable $F : \Theta \mapsto \Re^{p \times m}$ induces a smooth foliation of codimension pm on realization space (Θ, g).*

Another useful generalization of the scalar-valued observable is given in the following definition.

DEFINITION 2.17. *Let $f : Q \times \Theta \mapsto \Re$ be a function such that for each Λ in some admissible set Q it follows that the marginal map $f_\Lambda : \Theta \mapsto \Re$ is an observable function. Then, the set $\{f_\Lambda : \Lambda \in Q\}$ is called a* family of observables.

A simple example of such a family is the linear combination of two scalar-valued observables f_1 and f_2 given by

$$f_\Lambda(\theta) = \lambda_1 f_1(\theta) + \lambda_2 f_2(\theta),$$

where $\Lambda = [\lambda_1 \ \ \lambda_2]$ and $\Lambda\Lambda^T = 1$. A generalization of this example is given in the following lemma.

LEMMA 2.18 (see [10], [28]). *Let $F : \Theta \mapsto \Re^{p\times m}$ be a matrix-valued observable on realization space (Θ, g) and $K \in F(\Theta)$. Define the set $\{f_\Lambda : \Lambda \in Q\}$, where*

$$f_\Lambda(\theta) = \sum_{i,j=1}^{p,m} \lambda_{ji}(f_{ij}(\theta) - k_{ij}) = \operatorname{Tr} \Lambda(F(\theta) - K), \tag{32}$$

$$Q = \{\Lambda \mid \Lambda\Lambda^T = I_m \ \textit{if} \ p \geq m \ \textit{or} \ \Lambda^T\Lambda = I_p \ \textit{if} \ m \geq p\}. \tag{33}$$

(Clearly $Q = O(m)$, the set of $m \times m$ orthogonal matrices, when $p = m$.) Then

(i) $\{f_\Lambda : \Lambda \in Q\}$ is a family of observables;

(ii) $f_\Lambda(\theta) = 0$ for all $\Lambda \in Q$ if and only if $F(\theta) = K$.

Proof. (i) Observe that

$$df_\Lambda(\theta) = \left.\frac{\partial f_\Lambda}{\partial \theta^k}\right|_\theta d\theta^k = \sum_{i,j=1}^{p,m} \lambda_{ji} \left.\frac{\partial f_{ij}(\theta)}{\partial \theta^k}\right|_\theta d\theta^k = 0$$

for some θ if and only if $\Lambda = 0$ since $dF(\theta)$ is surjective. But $0 \notin Q$, and thus $df_\Lambda(\theta) \neq 0$ for all $\Lambda \in Q$ and $\theta \in \Theta$.

(ii) The necessity of the condition $F(\theta) = K$ is the only nontrivial part of the proof. Define $E(\theta) = F(\theta) - K$. For fixed $\theta \in \Theta$, let $E(\theta)$ have the singular value decomposition

$$E(\theta) = \sum_{j=1}^{\min(p,m)} \sigma_j u_j v_j^T.$$

Since Λ can be arbitrarily selected from Q, and any element of Q has all of its singular values equal to unity, it follows that

$$\begin{aligned}
\operatorname{Tr}\Lambda E(\theta) &= \operatorname{Tr}\left(\sum_{i=1}^{\min(p,m)} v_i u_i^T \cdot \sum_{j=1}^{\min(p,m)} \sigma_j u_j v_j^T\right) \\
&= \operatorname{Tr}\left(\sum_{i=1}^{\min(p,m)} \sum_{j=1}^{\min(p,m)} \sigma_j \delta_{ij} v_i v_j^T\right) \\
&= \operatorname{Tr}\left(\sum_{i=1}^{\min(p,m)} \sigma_i v_i v_i^T\right) \\
&= \sum_{i=1}^{\min(p,m)} \sigma_i \\
&= 0.
\end{aligned}$$

However, by the definition of the singular value decomposition, $\sigma_i \geq 0$ for all $i = 1 \ldots \min(p, m)$, thus $E(\theta) = 0$. □

While a matrix-valued observable and a family of scalar-valued observables can each be used to describe a given equivalence class, an important advantage in working with the latter is the relative ease with which the sensitivity theory developed in §2.2 can be adapted. This is illustrated by the following example.

Example. The goal is to determine the state space realizations of a second-order linear system with the property that the coefficients of the characteristic equation have extremal sensitivity with respect to fixed point and floating point quantization of the realization coefficients. Both fixed point and floating point problems are independent parameter design problems. Fixed point parameter quantization is modelled by introducing absolute perturbations $\Delta\theta^i$ of fixed magnitude such that $\hat{\theta}^i = \theta^i + \Delta\theta^i$. Floating point parameter quantization, on the other hand, is modelled by the introduction of a relative perturbation $\Delta\theta^i/\theta^i$ of fixed magnitude since, now, $\hat{\theta}^i = \theta^i[1 + \Delta\theta^i/\theta^i]$. In a geometric context we view these models as an endowment of Θ with two different metrics, an absolute metric $g^{\alpha}_{ij}(\theta) = \delta_{ij}$ and a relative metric $g^{\rho}_{ij}(\theta) = (\theta^i)^{-2}$ if $i = j$ and zero otherwise.

The transfer function of the system is assumed to have the form

$$G(s) = \frac{\omega_n^2}{s^2 + 2\zeta\omega_n s + \omega_n^2}.$$

Thus, we are interested in the set of state space realizations (A, b, c) such that $\det(sI - A) = s^2 + 2\zeta\omega_n s + \omega_n^2$. In the fixed point case we take $\Theta \equiv \Re^4$ ($\theta = [a_{11}\ a_{21}\ a_{12}\ a_{22}]^T$) and in the floating point case we assume Θ is equivalent to $\Re^4$ minus the points along the coordinate axes, where the relative metric is unbounded. (This is always an implicit assumption when the floating point metric is used.) The *two* algebraic constraints on the realization coefficients yield the vector observable

$$F(\theta) = \begin{bmatrix} a_{11} + a_{22} + 2\zeta\omega_n \\ a_{11}a_{22} - a_{21}a_{12} - \omega_n^2 \end{bmatrix},$$

which induces a foliation on Θ with codimension equal to two. Clearly, the leaf of the foliation in which we are interested has an observable value equal to zero. The family of observables is defined by

$$f_\Lambda(\theta) = \mathrm{Tr}\ \Lambda F(\theta) = \Lambda F(\theta),$$

where $\Lambda = [\lambda_1\ \lambda_2]$ and $\Lambda\Lambda^T = 1$. Note from Lemma 2.18 it follows that $f_\Lambda(\theta) = 0$ for all Λ with $\Lambda\Lambda^T = 1$ if and only if $F(\theta) = 0$. To solve the optimization problems, apply the criterion given by equation (24) using the absolute and relative metrics. Using the absolute metric it follows easily from equation (25) that the optimality equations are as follows:

$$\lambda_1 + \lambda_2 a_{11} = \bar{\lambda}(\lambda_1 + \lambda_2 a_{22}), \tag{34}$$

$$a_{21} = -\bar{\lambda} a_{12}, \tag{35}$$

$$a_{12} = -\bar{\lambda} a_{21}, \tag{36}$$

$$\lambda_1 + \lambda_2 a_{22} = \bar{\lambda}(\lambda_1 + \lambda_2 a_{11}), \tag{37}$$

where $\lambda = \bar{\lambda}\lambda_2$. The above equations together with the constraints $F(\theta) = 0$ and $\Lambda\Lambda^T = 1$ define a system of seven (not necessarily independent) equations in seven variables. It can be easily shown that the extremal sensitivity realizations must have an A matrix of the form

$$A = \begin{cases} \begin{bmatrix} -\sigma & \pm\omega_d \\ \mp\omega_d & -\sigma \end{bmatrix} & : \quad 0 \le \zeta \le 1, \\ \begin{bmatrix} -\sigma \pm \omega_d \cos\theta & \omega_d \sin\theta \\ \omega_d \sin\theta & -\sigma \mp \omega_d \cos\theta \end{bmatrix} & : \quad \zeta \ge 1, \end{cases}$$

where $\sigma = \zeta\omega_n$, $w_d = \omega_n\sqrt{|1-\zeta^2|}$, and θ is an arbitrary real parameter. In fact, using Lemma 2.13 it can be shown that these realizations are of minimum sensitivity when $\zeta \neq 1$.

Using the relative metric, it follows directly that the optimality equations are as follows:

$$a_{11}(\lambda_1 + \lambda_2 a_{22})^2 + a_{22}^2\lambda_2(\lambda_1 + \lambda_2 a_{11}) = \lambda(\lambda_1 + \lambda_2 a_{22}), \tag{38}$$

$$2a_{21}a_{12} = -\bar{\lambda}, \tag{39}$$

$$2a_{12}a_{21} = -\bar{\lambda}, \tag{40}$$

$$a_{11}^2\lambda_2(\lambda_1 + \lambda_2 a_{22}) + a_{22}(\lambda_1 + \lambda_2 a_{11})^2 = \lambda(\lambda_1 + \lambda_2 a_{11}). \tag{41}$$

These equations, along with the constraints mentioned above, define a system of six independent equations in seven unknowns. Hence the general set of extremal sensitivity floating point A matrices when $\zeta \neq 0$ and $\zeta \neq 1$ is a one-parameter family of the form

$$A = \begin{cases} -\omega_n \begin{bmatrix} \cos(\theta_o) & t\ \sin(\theta_o) \\ -\frac{1}{t}\ \sin(\theta_o) & \cos(\theta_o) \end{bmatrix} & : \quad 0 < \zeta = \cos(\theta_o) < 1, \\ -\omega_n \begin{bmatrix} \cosh(\theta_o) & t\ \sinh(\theta_o) \\ \frac{1}{t}\sinh(\theta_o) & \cosh(\theta_o) \end{bmatrix} & : \quad \zeta = \cosh(\theta_o) > 1, \end{cases}$$

where t is an arbitrary nonzero real parameter. There is no solution for the undamped and critically damped cases.

3. An application to the design of discrete-time linear time-invariant systems.

In this section, the synthesis of minimum sensitivity state space realizations for linear time-invariant multivariable systems is considered [10], [28]. It is shown that these realizations are related to the balanced realizations of Moore [22]. This problem has been solved in a purely algebraic framework in [21], [24], [25], [26] and in a stochastic framework in [17], [18]. Unlike these approaches, however, the geometric method naturally extends to other classes of dynamical systems, e.g., singular linear systems [9], [12], and bilinear systems [9], [11].

3.1. Realization spaces.

Consider the linear time-invariant system

$$x_{k+1} = Ax_k + Bu_k, \qquad x_k \in \Re^n, u_k \in \Re^m, \tag{42}$$

$$y_k = Cx_k, \qquad y_k \in \Re^p. \tag{43}$$

The state space realization (A, B, C) is specified by $n(n+m+p)$ real numbers. Therefore, we make the natural identification of the space of all such triples, denoted by $\Sigma_{m,n,p}(\Re)$, with the cartesian product space $\Re^{n(n+m+p)}$ with the usual topology. Furthermore, since the addition and scalar multiplication of such triples are of little significance globally, the realization space is assumed to have the algebraic structure of an affine space with vector space $\mathcal{V}^{n(n+m+p)}$. For fixed m, n, and p, it is clear from Definition 2.1 that $\Sigma_{m,n,p}(\Re)$ is a realization space.

Now consider the natural C^∞ group action of $GL_n(\Re)$ on $\Sigma_{m,n,p}(\Re)$ defined by

$$\phi : GL_n(\Re) \times \Sigma_{m,n,p}(\Re) \mapsto \Sigma_{m,n,p}(\Re) : T \times (A, B, C) \mapsto (TAT^{-1}, TB, CT^{-1}), \tag{44}$$

which corresponds to a change of basis in the state space $z = Tx$. The orbit of a point $s \in \Sigma_{m,n,p}(\Re)$ for the action ϕ is defined to be the subset

$$\mathcal{O}_s(\phi) = \{\phi(T, s) \in \Sigma_{m,n,p}(\Re) \mid T \in GL_n(\Re)\}. \tag{45}$$

The isotropy subgroup of $GL_n(\Re)$ at s is defined as

$$G_s(\phi) = \{T \in GL_n(\Re) \mid \phi(T, s) = s\}. \tag{46}$$

In general, the action of a Lie group G on a manifold M is said to be a *foliated action* if for every $s \in M$ the tangent space to the orbit of ϕ passing through s has fixed dimension. Equivalently, a foliated action is characterized by the property that the dimension of the isotropy subgroup $G_s(\phi)$ has fixed dimension independent of s. It is well known that the orbits of a foliated action define the leaves of a foliation [6], [20]. However, observe that for the particular realization space described above, the group action ϕ is *not* a foliated action. For example, if we let $s_1 = (\lambda I_n, 0, 0)$, where $\lambda \in \Re$, then clearly $G_{s_1}(\phi) = GL_n(\Re)$, and thus $\dim(G_{s_1}) = n^2$. On the other hand, for $s_2 = (\lambda I_n, e_1, 0)$, where $e_1^T = (1, 0, \ldots, 0)$, $G_{s_2}(\phi) = \{T \in GL_n(\Re) \mid$ first row of $T = e_1\}$ such that $\dim(G_{s2}) = n^2 - n$. Consequently, the orbits of ϕ do not foliate the realization space $\Sigma_{m,n,p}(\Re)$. However, if we define $\Sigma^{cr}_{m,n,p}(\Re)$, $\Sigma^{co}_{m,n,p}(\Re)$, and $\Sigma^{cr,co}_{m,n,p}(\Re)$ as the open subsets of $\Sigma_{m,n,p}(\Re)$ containing realizations that are reachable, observable, or both, respectively, then we have the following lemma.

LEMMA 3.1 (see [28]). *The group action ϕ restricted to either $\Sigma^{cr}_{m,n,p}(\Re)$, $\Sigma^{co}_{m,n,p}(\Re)$, or $\Sigma^{cr,co}_{m,n,p}(\Re)$ is a foliated action.*

Proof. The lemma follows immediately from the fact that the isotropy subgroup has constant dimension on $\Sigma^{cr}_{m,n,p}(\Re)$, $\Sigma^{co}_{m,n,p}(\Re)$, and $\Sigma^{cr,co}_{m,n,p}(\Re)$ (precisely zero). □

We shall also be interested in another type of realization space for linear time-invariant systems, specifically, sets whose elements comprise the components of the reachability and observability matrices

$$\mathcal{O}_i(A, C) = \left[\begin{array}{cccc} C^T & A^TC^T & \ldots & (A^T)^iC^T \end{array}\right]^T, \tag{47}$$

$$\mathcal{R}_j(A, B) = \left[\begin{array}{cccc} B & AB & \ldots & A^jB \end{array}\right], \tag{48}$$

where $i, j \geq n$. As will be shown shortly, such a realization space will be very natural for our applications. Such realizations will be referred to as $(\mathcal{O}, \mathcal{R})$ realizations in contrast to the usual state space realizations. As with the realization space $\Sigma_{m,n,p}(\Re)$, it is tempting to identify the space of all $(\mathcal{O}, \mathcal{R})$ realizations, say $\Omega_{m,n,p}[i, j](\Re)$, with an affine space modelled on the vector space $\mathcal{V}^{n(p(i+1)+(j+1)m)}$. This identification is fallacious, however, since the matrices as defined above have definite structure, and not every point in $\Re^{n(p(i+1)+(j+1)m)}$ has an associated $(\mathcal{O}, \mathcal{R})$ matrix pair. So consider instead the following mapping:

$$\omega : \Sigma_{m,n,p}(\Re) \mapsto \Re^{p(i+1)\times n} \times \Re^{n\times(j+1)m} : (A, B, C) \mapsto (\mathcal{O}_i(A, C), \mathcal{R}_j(A, B)), \tag{49}$$

where i and j are assumed to be fixed a priori. Observe that ω is a well-defined and smooth mapping since the components of $\mathcal{O}_i(A, C)$ and $\mathcal{R}_j(A, B)$ are smooth functions of the components of (A, B, C). Define the following subsets of $\Re^{p(i+1)\times n} \times \Re^{n\times(j+1)m}$:

$$\begin{aligned} \Omega_{m,n,p}(\Re) &= \omega(\Sigma_{m,n,p}(\Re)), \\ \Omega^{cr}_{m,n,p}(\Re) &= \omega(\Sigma^{cr}_{m,n,p}(\Re)), \\ \Omega^{co}_{m,n,p}(\Re) &= \omega(\Sigma^{co}_{m,n,p}(\Re)), \\ \Omega^{cr,co}_{m,n,p}(\Re) &= \omega(\Sigma^{cr,co}_{m,n,p}(\Re)). \end{aligned}$$

LEMMA 3.2. *The subset $\Omega^{\rho}_{m,n,p}(\Re)$ is a realization space, where either $\rho = cr$, $\rho = co$, or $\rho = cr, co$.*

Proof. The conclusion follows by virtue of the fact that $\Sigma^{\rho}_{m,n,p}$ is known to be a realization space and ω restricted to $\Sigma^{\rho}_{m,n,p}$ is a diffeomorphism. □

It should be noted that since $\Sigma^{\rho}_{m,n,p}(\Re)$ and $\Omega^{\rho}_{m,n,p}(\Re)$ are diffeomorphic, $\omega(\Sigma^{\rho}_{m,n,p}(\Re))$ is a submanifold imbedded in $\Re^{p(i+1)\times n} \times \Re^{n\times(j+1)m}$ with the *same* dimension as $\Sigma^{\rho}_{m,n,p}(\Re)$. This property is of limited utility for the applications considered here, however, since sensitivity is *not* a topological property.

As was the case for realization spaces $\Sigma^{\rho}_{m,n,p}(\Re)$, we are also interested in foliations of $\Omega^{\rho}_{m,n,p}(\Re)$ induced by a group action. Consider the following C^{∞} action of $GL_n(\Re)$ on $\Omega^{\rho}_{m,n,p}(\Re)$ defined by

$$(50) \qquad \psi : GL_n(\Re) \times \Omega^{\rho}_{m,n,p}(\Re) \mapsto \Omega^{\rho}_{m,n,p}(\Re) : T \times (\mathcal{O}, \mathcal{R}) \mapsto (\mathcal{O}T^{-1}, T\mathcal{R}).$$

We have the following analogous lemma.

LEMMA 3.3. *The group action ψ restricted to $\Omega^{\rho}_{m,n,p}(\Re)$, where $\rho = cr, \rho = co$, or $\rho = cr, co$, is a foliated action.*

Proof. The lemma follows immediately from the fact that the isotropy subgroup has constant dimension on $\Omega^{\rho}_{m,n,p}(\Re)$ for $\rho = cr$, $\rho = co$, or $\rho = cr, co$ (precisely zero). □

Now that the realization spaces associated with discrete-time linear systems have been defined, we next consider in detail the corresponding minimum sensitivity design problem via the geometric techniques of §2.

3.2. Minimum sensitivity realizations. Consider the problem of finding a minimum sensitivity state space realization for a discrete-time linear system described by a $p \times m$ strictly proper rational transfer matrix

$$(51) \qquad \mathrm{H}(z) = \sum_{i=1}^{\infty} H_i z^{-i},$$

where for each i, $H_i \in \Re^{p\times m}$. If one knows the minimal system order n a priori, then it is possible to uniquely identify the system from a truncated version of the system Hankel matrix

$$(52) \qquad \mathcal{H}[pi, jm] = \begin{bmatrix} H_1 & H_2 & H_3 & \dots & H_{j+1} \\ H_2 & H_3 & H_4 & \dots & H_{j+2} \\ H_3 & H_4 & H_5 & \dots & H_{j+3} \\ \vdots & \vdots & \vdots & & \vdots \\ H_{i+1} & H_{i+2} & H_{i+3} & \dots & H_{i+j+1} \end{bmatrix}.$$

In the worst case, the transfer matrix is completely parameterized by $2npm$ parameters, so the truncation must retain at least this number of entries. An especially convenient truncated form is $\mathcal{H}[pi, jm]$, where $i, j \geq n$. Any rank n factorization

$$(53) \qquad \mathcal{H}[pi, jm] = \mathcal{O}_i \mathcal{R}_j$$

corresponds directly to a minimal state space realization. Thus we can take a finite subset of the Markov parameters as observables in our geometric sensitivity theory.

Define ℓ as the least common multiple of p and m, and integers $s = \ell/p$ and $t = \ell/m$. Then, any *square* Hankel matrix of the form $\mathcal{H}[\ell i, i\ell]$, where $i \geq n$, will uniquely identify a linear system of order n. Furthermore, any corresponding rank n $(\mathcal{O}, \mathcal{R})$ factorization

$$(54) \qquad \mathcal{H}[\ell i, i\ell] = \mathcal{O}_{si} \mathcal{R}_{it}$$

defines a matrix observable on $\Omega^{cr,co}_{m,n,p}[si,it]$. So, in light of Lemma 2.18 define the family of observables

$$f_\Lambda(\mathcal{O}_{si},\mathcal{R}_{it}) = \operatorname{Tr}\Lambda(\mathcal{H}[\ell i, i\ell] - \mathcal{O}_{si}\mathcal{R}_{it}), \tag{55}$$

where $\Lambda \in \mathcal{O}((i+1)\ell)$, the set of all orthogonal matrices of dimension $(i+1)\ell \times (i+1)\ell$. Then, it follows from the conclusion of the lemma that $f_\Lambda(\mathcal{O}_{si},\mathcal{R}_{it}) = 0$ for all $\Lambda \in \mathcal{O}((i+1)\ell)$ if and only if the equality in equation (54) holds. Thus, f_Λ induces a foliation on $\Omega^{cr,co}_{m,n,p}[si,it]$, where each leaf of the foliation corresponds to a collection of $(\mathcal{O},\mathcal{R})$ realizations with the same corresponding Hankel matrix. The leaf with an associated observable value of zero consists of all $(\mathcal{O},\mathcal{R})$ factorizations of the *given* Hankel matrix. We can also define a related observable on $\Sigma^{cr,co}_{m,n,p}(\Re)$ as follows:

$$\hat{f}_\Lambda = f_\Lambda \circ \omega(A,B,C) \tag{56}$$

$$= \operatorname{Tr}\Lambda(\mathcal{H}[\ell i, i\ell] - \mathcal{O}(A,C)_{si}\mathcal{R}(A,B)_{it}). \tag{57}$$

This observable foliates $\Sigma^{cr,co}_{m,n,p}(\Re)$ into equivalence classes characterized by the corresponding Hankel matrix.

From a systems perspective the sensitivity of state space realizations is the most natural problem to pursue. Unfortunately, this problem is difficult to attack directly since it is highly nonlinear. Thus, we first consider the essentially linear problem of finding the least-sensitive fixed point $(\mathcal{O},\mathcal{R})$ realizations (i.e., factorizations) of a given Hankel matrix, and then return to the state space realization problem later. The main theorem of this section is given below.

THEOREM 3.4 (extremal sensitivity theorem). *Given an nth order linear system with m inputs and p outputs characterized by a square Hankel matrix* $\mathcal{H}[\ell i, i\ell]$,

(i) *extremal sensitivity points under the uniform metric on the leaf of the foliation induced by the observable family*

$$f_\Lambda(\mathcal{O}_{si},\mathcal{R}_{it}) = \operatorname{Tr}\Lambda(\mathcal{H}[\ell i, i\ell] - \mathcal{O}_{si}\mathcal{R}_{it}) \tag{58}$$

have the property that

$$\mathcal{R}_{it}\mathcal{R}^T_{it} = \mathcal{O}^T_{si}\mathcal{O}_{si}, \tag{59}$$

where ℓ is the least common multiple of p and m, $s = \ell/p$, $t = \ell/m$, and $i \geq n$;

(ii) *an extremal sensitivity point under the uniform metric is also an extremal under the floating point metric if*

$$\operatorname{vec}(\mathcal{R}^T_{it})^2 + \operatorname{vec}(\mathcal{O}_{si})^2 = \amalg, \tag{60}$$

where vec *denotes the column stacking operator, $[x^m]_i = x_i^m$ for any vector x and integer m, and* $\amalg = [1\ \ 1 \dots 1]^T$.

Proof. (i) Use optimality equation (25) from the previous section. With the coefficient vector taken to be

$$\theta = \begin{bmatrix} \operatorname{vec}(\mathcal{R}^T_{it}) \\ \operatorname{vec}(\mathcal{O}_{si}) \end{bmatrix}, \tag{61}$$

one can express the observable f_Λ in terms of a quadratic form in θ:

$$f_\Lambda(\theta) = \operatorname{Tr}\Lambda(\mathcal{H}[\ell i, i\ell] - \mathcal{O}_{si}\mathcal{R}_{it}) \tag{62}$$

$$= \operatorname{Tr}\Lambda\mathcal{H}[\ell i, i\ell] - \operatorname{Tr}\mathcal{R}_{it}\Lambda\mathcal{O}_{si} \tag{63}$$

$$= \operatorname{Tr}\Lambda\mathcal{H}[\ell i, i\ell] - (\operatorname{vec}(\mathcal{R}^T_{it}))^T \cdot (I_n \otimes \Lambda) \cdot \operatorname{vec}(\mathcal{O}_{si}). \tag{64}$$

Using this representation, the gradient vector and Hessian matrix are easily computed as

$$\nabla f_\Lambda(\theta) = \frac{\partial f_\Lambda}{\partial \theta} = \begin{bmatrix} -(I_n \otimes \Lambda)\mathrm{vec}(\mathcal{O}_{si}) \\ -(I_n \otimes \Lambda)^T \mathrm{vec}(\mathcal{R}_{it}^T) \end{bmatrix} \tag{65}$$

and

$$\nabla^2 f_\Lambda(\theta) = \frac{\partial^2 f_\Lambda}{\partial \theta^2} = \begin{bmatrix} 0 & -(I_n \otimes \Lambda) \\ -(I_n \otimes \Lambda)^T & 0 \end{bmatrix}. \tag{66}$$

The optimality condition is then

$$\begin{bmatrix} -\lambda I_{nsi} & -(I_n \otimes \Lambda) \\ -(I_n \otimes \Lambda)^T & -\lambda I_{itn} \end{bmatrix} \cdot \begin{bmatrix} -(I_n \otimes \Lambda)\mathrm{vec}(\mathcal{O}_{si}) \\ -(I_n \otimes \Lambda)^T \mathrm{vec}(\mathcal{R}_{it}^T) \end{bmatrix} = 0. \tag{67}$$

Equation (67) gives

$$-\lambda(I_n \otimes \Lambda)\mathrm{vec}(\mathcal{O}_{si}) = \mathrm{vec}(\mathcal{R}_{it}^T), \tag{68}$$

$$\mathrm{vec}(\mathcal{O}_{si}) = -\lambda(I_n \otimes \Lambda)^T \mathrm{vec}(\mathcal{R}_{it}^T), \tag{69}$$

or equivalently,

$$-\lambda\Lambda\mathcal{O}_{si} = \mathcal{R}_{it}^T, \tag{70}$$

$$\mathcal{O}_{si} = -\lambda\Lambda^T\mathcal{R}_{it}^T. \tag{71}$$

Hence, the conclusion follows immediately using the facts that Λ is an orthogonal matrix, and λ, an eigenvalue of the symmetric orthogonal matrix $\nabla^2 f_\Lambda(\theta)$, is equal to ±1.

(ii) Using the general optimality equation (24) with $G(\theta) = \mathrm{diag}(\theta^{-2})$, and assuming property (59) gives

$$\mathrm{vec}(\mathcal{R}_{it}^T)^3 - \lambda(I_n \otimes \Lambda)\mathrm{vec}(\mathcal{O}_{si})^3 = \mathrm{vec}(\mathcal{R}_{it}^T). \tag{72}$$

Substitution using equation (68) proves the theorem's conclusion. □

It should be noted that the condition stated in (ii) is only a sufficient condition for a floating point extremal. Not every system has a realization which lies on the hypersphere defined by equation (60) and satisfies equation (59). The existence of general solutions of the floating point extremal sensitivity problem and the associated computational algorithms are current topics of investigation.

DEFINITION 3.5. *A realization* (A, B, C) *or* $(\mathcal{O}_{si}, \mathcal{R}_{it})$ *is said to be* essentially balanced *if* $\mathcal{R}_{it}\mathcal{R}_{it}^T = \mathcal{O}_{si}^T\mathcal{O}_{si}$.

This definition is motivated by the fact that in the discrete-time case $P_{it} \triangleq \mathcal{R}_{it}\mathcal{R}_{it}^T$ and $Q_{si} \triangleq \mathcal{O}_{si}^T\mathcal{O}_{si}$ are, respectively, the reachability grammian at time $t_1 = it$ and the observability grammian at time $t_2 = si$. Letting

$$P = \lim_{i \to \infty} \mathcal{R}_{it}\mathcal{R}_{it}^T, \tag{73}$$

$$Q = \lim_{i \to \infty} \mathcal{O}_{si}^T\mathcal{O}_{si}, \tag{74}$$

denote the usual steady-state grammians, any essentially balanced state space realization or $(\mathcal{O}, \mathcal{R})$ realization with $i \to \infty$ is only an orthogonal state space transformation away from being a truly balanced realization (in the sense of Moore [22]). Such realizations have also arisen in the context of optimal low noise filter structures [29] and gradient flow techniques for computing balanced realizations [16], [23]. Another important property of an essentially balanced $(\mathcal{O},\mathcal{R})$ realization is given in the following theorem.

THEOREM 3.6 (see [10]). *An essentially balanced $(\mathcal{O},\mathcal{R})$ realization has the minimum sensitivity property under the uniform metric.*

Proof. The conclusion follows directly from an application of Lemma 2.13. For any choice of Λ, $\nabla^2 f_\Lambda(\theta)$ clearly has a spectrum $\{\lambda_i\}$ consisting of n eigenvalues equal to one and n eigenvalues equal to minus one, and a corresponding set of orthonormal eigenvectors $\{v_i\}$. At an extremal we may also assume that v_{2n} corresponds to the normalized version of $\nabla f_\Lambda(\theta^*)$. Now let x be an arbitrary nonzero vector from the tangent space of $M_0 f_\Lambda$ at θ^* (and hence orthogonal to $\nabla f_\Lambda(\theta^*)$). Then from equation (31) it follows that

$$\begin{aligned}\nabla^2 H(\theta^*)(x,x) &= x^T(I-\lambda_{2n}\nabla^2 f_\Lambda(\theta^*))x\\ &= \sum_{i=1}^{2n-1}\alpha_i^2(1-\lambda_{2n}\lambda_i) > 0,\end{aligned}$$

where $x=\sum_{i=1}^{2n-1}\alpha_i v_i$. □

Determining minimum sensitivity state space realizations is significantly more difficult than finding minimum sensitivity $(\mathcal{O},\mathcal{R})$ realizations. The complication arises from the fact that the observables, in this case the Markov parameters, have a nonlinear dependence on the components of the A matrix. The optimality equations resulting from the direct application of Theorem 2.6 are virtually intractable analytically. The fixed point analysis above, however, does lead to a bound-optimal result which is related to the earlier work done using frequency domain techniques [21], [24], [25], [26]. Observe that in light of Definition 2.5, the performance index for the minimum sensitivity $(\mathcal{O},\mathcal{R})$ realization problem can be written in the form

$$L(\mathcal{O},\mathcal{R}) = \frac{1}{2} \parallel \nabla f_\Lambda \parallel^2 \tag{75}$$

$$= \frac{1}{2}\left|\begin{array}{c}\operatorname{vec}(\mathcal{O})\\ \operatorname{vec}(\mathcal{R}^T)\end{array}\right|^2 \tag{76}$$

$$= \frac{1}{2}(\mathrm{Tr}\ \mathcal{R}\mathcal{R}^T + \mathrm{Tr}\ \mathcal{O}^T\mathcal{O}) \tag{77}$$

$$= \frac{1}{2}(\mathrm{Tr}P + \mathrm{Tr}Q), \tag{78}$$

where the subscripts are dropped to indicate the limiting case $i \to \infty$. For a sequence of matrices $\{M_k\}$, define the ℓ_p norm

$$\parallel M_k \parallel_{F,\ell_p} \stackrel{\Delta}{=} \left(\sum_{k=1}^{\infty} \parallel M_k \parallel_F^p\right)^{\frac{1}{p}}, \tag{79}$$

where F denotes the Frobenius norm. Then it follows that

$$\left\|\frac{\partial H_k}{\partial C^T}\right\|_{F,\ell_2}^2 = \left\|\left[\frac{\partial H_1}{\partial C^T} \quad \frac{\partial H_2}{\partial C^T} \quad \frac{\partial H_3}{\partial C^T} \quad \cdots\right]\right\|_F^2 = p\mathrm{Tr}\ P, \tag{80}$$

$$(81)\qquad \left\|\frac{\partial H_k}{\partial B}\right\|_{F,\ell_2}^2 = \left\|\begin{bmatrix}\frac{\partial H_1}{\partial B} & \frac{\partial H_2}{\partial B} & \frac{\partial H_3}{\partial B} & \cdots\end{bmatrix}\right\|_F^2 = m\mathrm{Tr}\ Q.$$

Thus, Tr P and Tr Q are measures of the sensitivity of the impulse response to the components of B and C *per I/O channel.* The performance index $L(\mathcal{O},\mathcal{R})$ is the arithmetic average of the two measures. Furthermore, it can easily be shown that

$$(82)\qquad \left\|\frac{\partial H_k}{\partial A}\right\|_F^2 \le \mathrm{Tr}\ P_{k-1}\mathrm{Tr}\ Q_{k-1},$$

where $k = 1, 2, \ldots$. For stable systems, the real number sequences $\{\mathrm{Tr}\ P_{k-1}\}_{k\ge 1}$ and $\{\mathrm{Tr}\ Q_{k-1}\}_{k\ge 1}$ are bounded and monotone increasing. Therefore, it follows directly that

$$(83)\qquad \left\|\frac{\partial H_k}{\partial A}\right\|_{F,\ell_\infty} = \sup_k \left\|\frac{\partial H_k}{\partial A}\right\|_F \le \sqrt{\mathrm{Tr}\ P\ \mathrm{Tr}\ Q}.$$

However, for positive numbers Tr P and Tr Q,

$$(84)\qquad \sqrt{\mathrm{Tr}\ P\ \ \mathrm{Tr}\ Q} \le \frac{1}{2}(\mathrm{Tr}\ P + \mathrm{Tr}\ Q)$$

in general with equality if and only if Tr P = Tr Q. Thus, minimizing the performance index $L(\mathcal{O},\mathcal{R})$ will also minimize an upper bound on the sensitivity of the block Markov parameters (the observables) with respect to the components of A. In particular, when we assume $P = Q$ we have an upper bound (in the single-input, single-output case) on the sensitivity measure

$$(85)\qquad M_T \triangleq \left\|\frac{\partial H_k}{\partial A}\right\|_{F,\ell_\infty}^2 + \left\|\frac{\partial H_k}{\partial b}\right\|_{F,\ell_2}^2 + \left\|\frac{\partial H_k}{\partial c^T}\right\|_{F,\ell_2}^2$$

equal to

$$(86)\qquad M^* = \mathrm{Tr}\ P\ \ \mathrm{Tr}\ Q + \mathrm{Tr}\ P + \mathrm{Tr}\ Q.$$

This expression is identical to the general upper bound on the sensitivity of the frequency response

$$(87)\qquad M_F \triangleq \left\|\frac{\partial H(z)}{\partial A}\right\|_{F,L_1}^2 + \left\|\frac{\partial H(z)}{\partial b}\right\|_{F,L_2}^2 + \left\|\frac{\partial H(z)}{\partial c^T}\right\|_{F,L_2}^2$$

defined in [21], [24], [25], [26], where

$$(88)\qquad \|\ F(z)\ \|_{F,L_p} \triangleq \left(\frac{1}{2\pi j}\oint_\Gamma \|\ F(z)\ \|_F^p\ z^{-1}\,dz\right)^{\frac{1}{p}} = \left(\frac{1}{2\pi}\int_0^{2\pi} \|\ F(e^{j\omega})\ \|_F^p\right)^{\frac{1}{p}},$$

and Γ denotes the contour along the unit circle. A fundamental difference, however, between the time domain and frequency approaches is that M_T will not achieve the upper bound M^* in the optimal case while it is known that $M_F = M^*$ when Tr P = Tr Q [26]. This can be demonstrated by a simple first-order example as can the fact that M^* will not in general be an upper bound on M_T when the ℓ_∞ norm is replaced by either the ℓ_1 norm or the ℓ_2 norm. While using mixed norms as in equations (85) and (87) may seem somewhat artificial, at present no analytical technique has been found for direct optimization using a single norm in either the time or frequency domain approaches (see [30]).

4. Conclusions. Minimum sensitivity designs were defined as points in realization space, which preserved model behavior and minimized a geometrically motivated sensitivity measure. Such a characterization yielded a general optimality equation with a simple eigen-condition interpretation. The method was then applied to the problem of synthesizing minimum sensitivity state space realizations of discrete-time linear time-invariant systems under fixed point and floating point metrics. Optimal fixed point realizations were related to so-called balanced realizations.

Acknowledgment. The authors wish to sincerely thank the reviewers for their valuable comments and insights. Their diligence has served to significantly improve this paper.

REFERENCES

[1] A. Bohm, *Quantum Mechanics: Foundations and Applications*, 2nd ed., Springer-Verlag, New York, 1986.

[2] J. W. Brewer, *Matrix calculus and the sensitivity analysis of linear dynamic systems*, IEEE Trans. Automat. Control, AC-23 (1978), pp. 748–751.

[3] ———, *Kronecker products and matrix calculus in system theory*, IEEE Trans. Circuits Syst., CAS-25 (1978), pp. 772-781.

[4] ———, *Correction to "Kronecker products and matrix calculus in system theory"*, IEEE Trans. Circuits Systems, CAS-26 (1979), p. 360.

[5] ———, *Derivatives of the characteristic polynomial, trace, and determinant with respect to a matrix*, IEEE Trans. Automat. Control, AC-24 (1979), pp. 787–790.

[6] C. Camacho and A. L. Neto, *Geometric Theory of Foliations* (translation of *Teoria Geometrica das Folheacoes*), Birkhäuser, Boston, 1985.

[7] M. P. do Carmo, *Riemannian Geometry* (Translation of *Geometria Riemanniana*), Birkhäuser, Boston, 1992.

[8] P. M. Frank, *Introduction to System Sensitivity Theory*, Academic Press, New York, 1978.

[9] W. S. Gray, *A geometric approach to the parametric sensitivity of dynamical systems*, Ph.D. thesis, School of Electrical Engineering, Georgia Institute of Technology, Atlanta, GA, 1989.

[10] W. S. Gray and E. I. Verriest, *Optimality properties of balanced realizations: Minimum sensitivity*, in Proc. 26th IEEE Conference on Decision and Control, Los Angeles, 1987.

[11] ———, *The parametric sensitivity of bilinear dynamical systems*, in Progress in Systems and Control Theory, Vol. 1, M. A. Kaashoek, A. C. M. Ran, and J. H. van Schuppen, eds., Birkhäuser, Boston, 1990, pp. 369–378.

[12] W. S. Gray, E. I. Verriest, and F. L. Lewis, *A Hankel matrix approach to singular system realization theory*, in Proc. 29th IEEE Conference on Decision and Control, Honolulu, HI, 1990.

[13] M. Hazewinkel, *Moduli and canonical forms for linear dynamical systems* II: *The topological case*, Math. Systems Theory, 10 (1977), pp. 363–385.

[14] ———, (*Fine*) *moduli* (*spaces*) *for linear systems: What are they good for?* in Geometric Methods for the Theory of Linear Systems, C. I. Byrnes and C. Martin, eds., D. Reidel, Dordrecht, the Netherlands, 1980, pp. 125–194.

[15] ———, *On families of linear systems: Degeneration phenomena*, in Algebraic and Geometric Methods in Linear System Theory, C. I. Byrnes and C. F. Martin, eds., Lecture Notes in Mathematics 18, American Mathematical Society, Providence, RI, 1980, pp. 157–189.

[16] U. Helmke, *Balanced realizations for linear systems: A variational approach*, SIAM J. Control Optim., 31 (1993), pp. 1–15.

[17] M. Iwatsuki, M. Kawamata, and T. Higuchi, *Synthesis of minimum sensitivity structures in linear systems using controllability and observability measures*, in Proc. Internat. Conference on Acoustics, Speech, and Signal Processing 86, Tokyo, Japan, Institute for Electrical and Electronics Engineers, 1986.

[18] M. Kawamata and T. Higuchi, *A unified approach to the optimal synthesis of fixed-point state-space digital filters*, IEEE Trans. Acoust., Speech, Signal Proces., ASSP-33 (1985), pp. 911–920.

[19] A. S. Khadr and C. F. Martin, *On the GL(n,R) action on linear systems: The orbit closure problem*, in Algebraic and Geometric Methods in Linear System Theory C. I. Byrnes and

C. F. Martin, eds., Lecture Notes in Applied Mathematics 18, American Mathematical Society, Providence, RI, 1980, pp. 239–251.

[20] H. B. LAWSON, JR., *The Quantitative Theory of Foliations*, in Regional Conference Series in Mathematics No. 27, American Mathematical Society, Providence, RI, 1977.

[21] W. J. LUTZ AND S. L. HAKIMI, *Design of multi-input multi-output systems with minimum sensitivity*, IEEE Trans. Circuits Systems, CAS-35 (1988), pp. 1114–1122.

[22] B. C. MOORE, *Principal component analysis in linear systems: controllability, observability, and model reduction*, IEEE Trans. Automat. Control, AC-26 (1981), pp. 17–32.

[23] J. E. PERKINS, U. HELMKE, AND J. B. MOORE, *Balanced realizations via gradient flow techniques*, Systems Control Lett., 14 (1990), pp. 369–380.

[24] V. TAVSANOĞLU AND L. THIELE, *Optimal design of state-space digital filters by simultaneous minimization of sensitivity and roundoff noise*, IEEE Trans. Circuits Systems, CAS-31 (1984), pp. 884–888.

[25] L. THIELE, *Design of sensitivity and round-off noise optimal state-space discrete systems*, Internat. J. Circuit Theory Appl., 12 (1984), pp. 39–46.

[26] ———, *On the sensitivity of linear state-space systems*, IEEE Trans. Circuits Systems, CAS-33 (1986), pp. 502–510.

[27] E. I. VERRIEST, *Minimum sensitivity implementations for multi-mode systems*, in Proc. 27th IEEE Conference on Decision and Control, Austin, TX, 1988.

[28] E. I. VERRIEST AND W. S. GRAY, *Robust design problems: a geometric approach*, in Linear Circuits, Systems and Signal Processing: Theory and Application, C. I. Byrnes, C. F. Martin, and R. E. Saeks, eds., Elsevier–North Holland, Amsterdam, 1988, pp. 321–328.

[29] D. WILLIAMSON, *A property of internally balanced and low noise structures*, IEEE Trans. Automat. Control, AC-31 (1986), pp. 633–634.

[30] W.-Y. YAN AND J. B. MOORE, *On L^2-sensitivity minimization of linear state-space systems*, IEEE Trans. Circuits Systems I Fund. Theory Appl., CAS-39 (1992), pp. 641–648.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 33, No. 3, pp. 882–915, May 1995

© 1995 Society for Industrial and Applied Mathematics
011

DISCRETE APPROXIMATIONS AND REFINED EULER–LAGRANGE CONDITIONS FOR NONCONVEX DIFFERENTIAL INCLUSIONS*

BORIS S. MORDUKHOVICH[†]

This paper is dedicated to Terry Rockafeller on the occasion of his 60*th birthday.*

Abstract. This paper deals with the Bolza problem (P) for differential inclusions subject to general endpoint constraints. We pursue a twofold goal. First, we develop a finite difference method for studying (P) and construct a discrete approximation to (P) that ensures a strong convergence of optimal solutions. Second, we use this direct method to obtain necessary optimality conditions in a refined Euler–Lagrange form without standard convexity assumptions. In general, we prove necessary conditions for the so-called intermediate relaxed local minimum that takes an intermediate place between the classical concepts of strong and weak minima. In the case of a Mayer cost functional or boundary solutions to differential inclusions, this Euler–Lagrange form holds without any relaxation. The results obtained are expressed in terms of nonconvex-valued generalized differentiation constructions for nonsmooth mappings and sets.

Key words. discrete approximations, differential inclusions, nonsmooth analysis, generalized differentiation, Euler–Lagrange conditions

AMS subject classifications. 49K24, 49J52, 49M25

1. Introduction.

This paper is mainly concerned with the following problem of dynamic optimization: minimize the Bolza functional

$$J[x] := \varphi(x(a), x(b)) + \int_a^b f(x(t), \dot{x}(t), t)dt \tag{1.1}$$

over absolutely continuous trajectories $x : [a, b] \to \mathbf{R}^n$ for the differential inclusion

$$\dot{x}(t) \in F(x(t), t) \ \text{ a.e. } \ t \in [a, b] \tag{1.2}$$

subject to general endpoint constraints

$$(x(a), x(b)) \in \Omega \subset \mathbf{R}^{2n}. \tag{1.3}$$

Here $T := [a, b]$ is a fixed time interval and F is a set-valued mapping (multifunction). We label this problem (P) and call it *the Bolza problem for differential inclusions.*

The formulated problem covers a broad range of other problems in dynamic optimization, in particular, both standard and nonstandard models in optimal control for open-loop and closed-loop control systems (see, e.g., Clarke [8]). On the other hand, problem (P) can be imbedded in the so-called Generalized Problem of Bolza [39], [8] where the function f is allowed to take values in $\bar{\mathbf{R}} := \mathbf{R} \cup \{+\infty\}$. In this paper we

* Received by the editors March 16, 1993; accepted for publication (in revised form) January 3, 1994. The results and techniques of this paper were presented at the Institute for Mathematics and Its Applications Workshop "Nonsmooth Analysis and Geometric Methods in Deterministic Optimal Control" on February 15, 1993. This research was partially supported by National Science Foundation grant DMS–9206989 and the Career Development Chairs Award at Wayne State University and also by the Institute for Mathematics and Its Applications with funds provided by the National Science Foundation. A preliminary version of this paper has been published in Institute for Mathematics and Its Applications Preprint Series #1115, March 1993.

[†] Department of Mathematics, Wayne State University, Detroit, Michigan 48202 (boris@math.wayne.edu).

prefer to consider the problem in form (1.1)–(1.3) and prove results depending on the specific character of differential inclusions.

The mainstream in studying optimization problems for differential inclusions consists of obtaining necessary conditions for optimality (global or strong local minima). There are different approaches and various results in this area using one or another tool in nonsmooth analysis; we refer the reader to [5]–[10], [16], [22]–[25], [27]–[29], [35], [37], [38], [46], [51] and the bibliography therein. Most of the results are obtained for the *Mayer problem*, which corresponds to (1.1)–(1.3) in the case where $f = 0$.

In [24], Loewen and Rockafellar consider the Bolza problem (P) (with additional state constraints). Assuming that the function $f(x, \cdot, t)$ and the sets $F(x,t)$ of admissible velocities are *convex*, they obtain necessary optimality conditions under usual boundedness and Lipschitzness hypotheses but without imposing any constraint qualification such as calmness (cf. Clarke [5]–[10]). They prove that if $\bar{x}(t)$ solves problem (1.1)–(1.3), then there exist a number $\lambda \geq 0$ and an absolutely continuous function $p : [a,b] \to \mathbf{R}^n$, not both zero, such that

$$(1.4)\,(\dot{p}(t), p(t)) \in \lambda\partial_C f(\bar{x}(t), \dot{\bar{x}}(t), t) + N_C((\bar{x}(t), \dot{\bar{x}}(t)); \mathrm{gph}F(\cdot, t)) \text{ a.e. } t \in [a,b],$$

$$(1.5)\qquad (-\dot{p}(t), \dot{\bar{x}}(t)) \in \partial_C H_\lambda(\bar{x}(t), p(t), t) \text{ a.e. } t \in [a,b],$$

$$(1.6)\qquad \langle p(t), \dot{\bar{x}}(t)\rangle = H_\lambda(\bar{x}(t), p(t), t) \text{ a.e. } t \in [a,b],$$

$$(1.7)\qquad (p(a), -p(b)) \in \lambda\partial_C\varphi(\bar{x}(a), \bar{x}(b)) + N_C((\bar{x}(a), \bar{x}(b)); \Omega)$$

where $\mathrm{gph}\, F(\cdot, t) := \{(x, v) \in \mathbf{R}^{2n} | v \in F(x,t)\}$,

$$(1.8)\qquad H_\lambda(x, p, t) := \max\{\langle p, v\rangle - \lambda f(x, v, t) | v \in F(x,t)\},$$

and notation N_C and ∂_C, respectively, stand for Clarke's normal cone to a closed set at a given point and the generalized gradient of a locally Lipschitz function [8].

Condition (1.4), which is called the Euler–Lagrange inclusion, was first obtained by Clarke [5] for the Mayer problem under the calmness assumption, which ensures normality ($\lambda = 1$) in (1.4), (1.7). The Hamiltonian inclusion (1.5) was proved by Clarke first under the calmness hypothesis and then without it; see [8], [9]. Observe that (1.6) is the Weierstrass–Pontryagin maximum condition, which is implied by each of the conditions (1.4) and (1.5) under the convexity assumptions imposed. Note also that, in general, conditions (1.4) and (1.5) are independent; see examples in [22], [25].

Another version of necessary optimality conditions was obtained by Mordukhovich for the Mayer problem (1.1)–(1.3) with $f = 0$ under the convexity of $F(x,t)$ and usual boundedness and Lipschitzness assumptions but without any calmness hypotheses or something similar; see [27]–[29]. The conditions obtained are stated in the form:

$$(1.9)\qquad (\dot{p}(t), \dot{\bar{x}}(t)) \in \mathrm{co}\{(u,v) | (u, p(t)) \in N((\bar{x}(t), v); \mathrm{gph}F(\cdot, t)),$$

$$v \in M(\bar{x}(t), p(t), t)\} \text{ a.e. } t \in [a,b],$$

$$(1.10)\qquad (p(a), -p(b)) \in \lambda\partial\varphi(\bar{x}(a), \bar{x}(b)) + N((x(a), x(b)); \Omega)$$

where "co" means the convex hull of a set,

$$M(x,p,t) := \{v \in F(x,t) | \langle p, x\rangle = H(x,p,t)\}, \tag{1.11}$$

and $H(x,p,t)$ coincides with the Hamiltonian (1.8) for $f = 0$. Here N and ∂ are not Clarke's normal cone and generalized gradient but their *nonconvex* counterparts whose convex closures coincide with the corresponding constructions of Clarke; see §4 for more details. These nonconvex constructions were first used in Mordukhovich [26] for obtaining transversality conditions like (1.10) in nonsmooth problems of optimal control.

Observe that condition (1.9) implies both the maximum condition (1.6) and an analogue of the Euler–Lagrange inclusion (1.4) in the form

$$\begin{aligned} \dot p(t) \in \mathrm{co}\{u | (u, p(t)) \in N((\bar x(t), v); \mathrm{gph} F(\cdot, t)),\\ v \in M(\bar x(t), p(t), t)\} \text{ a.e. } t \in [a,b]. \end{aligned} \tag{1.12}$$

In comparison with (1.4) for $f = 0$, condition (1.12) requires less convexification: only to the components involving derivatives of the adjoint function instead of to all components at once. This makes (1.12) essentially stronger than (1.4) in certain situations. In particular, if the maximum set (1.11) is a singleton along $(\bar x(t), p(t))$ for a.e. $t \in [a,b]$ (it happens, for instance, if the sets $F(x,t)$ are strictly convex along $\bar x(t)$), then (1.12) is reduced to

$$\dot p(t) \in \mathrm{co}\{u | (u, p(t)) \in N((\bar x(t), \dot{\bar x}(t)); \mathrm{gph} F(\cdot, t))\} \text{ a.e. } t \in [a,b], \tag{1.13}$$

which is strictly better than (1.4).

So (1.9) turns out to be an advanced version of the Euler–Lagrange inclusion and the maximum condition for the Mayer problem involving convex (i.e., *convex-valued*) differential inclusions (1.2). What relationships exist between (1.9) and the Hamiltonian inclusion (1.5) under the usual convexity, boundedness, and Lipschitzness assumptions?

It follows from Rockafellar's dualization result [43] that (1.5) implies (1.9). On the other hand, it has been recently proved by Ioffe (personal communication; see also [19, §3.5]) that (1.9) implies (1.5) under the mentioned assumptions. Therefore, version (1.9) of the Euler–Lagrange condition is *equivalent* to the Hamiltonian condition in Clarke's form (with the same adjoint function) for convex differential inclusions. This prolongs the line of equivalency between the Hamiltonian and Euler–Lagrange conditions, which is well known for smooth and fully convex problems (see, e.g., [39], [43]). Now one can conclude that any improvement of the necessary optimality conditions in form (1.9) provides a strengthening of the Hamiltonian conditions in Clarke's form for convex differential inclusions.

In the recent paper [25], Loewen and Rockafellar establish that the Mayer problem for convex differential inclusions can actually be reduced to the situation where the sets $F(x,t)$ are *strictly convex* along the optimal trajectory. In the latter case, conditions (1.12) and (1.13) are equivalent while (1.9) is equivalent to the simultaneous fulfilment of (1.6) and (1.13). In the general convex case, (1.13) always implies the maximum condition in (1.8) with $f = 0$; see Proposition 4.7 stated below. Therefore, (1.13) also implies (1.12) as well as (1.9) where the improvement may be proper; see [25].

Using the mentioned strict convexification procedure and a Hamiltonian calculus, Loewen and Rockafellar prove in [25] that an optimal solution to the Mayer problem for convex differential inclusions always satisfies conditions (1.10) and (1.13) as well as (1.5) and (1.6) where $f = 0$ with the same adjoint function $p(t)$. Moreover, they also consider the case of unbounded differential inclusions, truncating it to the bounded case under suitable Lipschitzian assumptions and the convexity of $F(x,t)$. Their consideration of the unbounded case leads to improvements of necessary conditions for the Bolza problem with convex velocities.

It follows from the discussion above that the refined Euler–Lagrange and transversality conditions (1.13) and (1.10) provide the best results for convex differential inclusions. On the other hand, convexity assumptions appear to be restrictive in certain important situations, so it makes sense to release them as much as possible. Note that if the admissible velocity sets $F(x,t)$ are convex and the multifunction $F(\cdot,t)$ is Lipschitz continuous, then the differential inclusion (1.2) admits a control representation $F(x,t) = g(x,U,t)$ with a Lipschitzian function g and a control set U independent on x; see, for example, [3]. This is no longer the case when $F(x,t)$ are not convex. So when considering nonconvex differential inclusions, one should definitely study them for their own sakes.

The primary goal of this paper is to develop the theory of necessary optimality conditions in the *refined Euler–Lagrange form* for Mayer and Bolza problems involving *nonconvex* bounded differential inclusions. The results obtained below achieve the following advancements in the state of art.

1. We study a new (to the best of our knowledge) concept of local minimum for the considered variational problems involving differential inclusions. Previous results for such problems were concerned with strong (or global) minima. In contrast to the strong minimum, we compare a reference feasible trajectory $\bar{x}(\cdot)$ with other feasible ones close to it not only in the C-norm for arcs but also in the L^p-norm ($1 \leq p < \infty$) for derivatives. This means that we consider a neighborhood of $\bar{x}(\cdot)$ in the Sobolev space $W^{1,p}$ equipped with a natural topology. Such a local minimum takes an intermediate place between the classical weak and strong minima; we call it the *intermediate local minimum.* Note that the results obtained in this paper provide new information even for convex differential inclusions. In particular, they imply the maximum condition for an intermediate local minimum, which may not be strong.

2. We obtain refined necessary conditions for the Bolza problem (P) stated above with the Euler–Lagrange inclusion

$$(1.14)\quad \dot{p}(t) \in \mathrm{co}\{u | (u, p(t)) \in \lambda \partial f(\bar{x}(t), \dot{\bar{x}}(t), t) + N((\bar{x}(t), \dot{\bar{x}}(t)); \mathrm{gph} F(\cdot, t))\}$$

and the transversality inclusion (1.10) where an absolutely continuous function $p(\cdot)$ and a number $\lambda \geq 0$ are not equal to zero simultaneously.

In general, we prove the refined Euler–Lagrange inclusion (1.14) for any trajectory $\bar{x}(t)$ that is feasible for the original problem (P) and provides an intermediate local minimum for the so-called *relaxed* problem obtained from (P) by some convexification procedure. Note that in this case, condition (1.14) is expressed in terms of the original data F, f and may be quite different from its counterpart in terms of the convexifications. We discuss effective sufficient conditions when an optimal solution to (P) solves the relaxed problem as well, so the conditions obtained characterize solutions to the original problem of Bolza without any convexity.

3. In the case of a Mayer functional in (1.1)–(1.3), we prove that the refined Euler–Lagrange inclusion (1.14), coinciding with (1.13) in this case, is fulfilled for every *strong minimum* $\bar{x}(t)$ *without any relaxation.* This implies that if $\bar{x}(t)$ solves the original Mayer problem but may not solve the relaxed one, conditions (1.10) and (1.13)

still hold. Moreover, these conditions are proved to be necessary for the *weak minimum* to the Mayer problem under an additional Riemann integrability assumption that makes the technique used more transparent and in principle can be omitted.

We also establish the refined Euler–Lagrange inclusion (1.13) for any *boundary trajectory* of (1.2) without either convexity or relaxation. The latter result essentially strengthens the recent one of Kaskosz and Lojasiewicz [22] who proved the Euler–Lagrange inclusion in Clarke's form (1.4) with $f = 0$ for boundary trajectories.

So we obtain that the refined Euler–Lagrange conditions, providing strongest results for convex differential inclusions, also hold true for nonconvex problems. This is proved under the relaxability assumption (so far) for the general Bolza problem and without the latter assumption for the Mayer problem as well as for boundary trajectories. Note, however, that the maximum condition no longer follows from (1.13) or (1.14) in the nonconvex case.

The Hamiltonian inclusion (1.5) always implies the maximum condition (1.6), but for now (1.5) has been justified (as a necessary condition for the strong minimum or boundary trajectories) only in the convex case. This implies that the Hamiltonian condition also holds under the relaxability assumption because (1.5), in contrast to the Euler–Lagrange inclusions, is obviously invariant with respect to convexification. It follows from [25, Example 5.2] that the refined Euler–Lagrange inclusion may be essentially better than the Hamiltonian inclusion in the convex case. A series of examples in [22, §2] shows that the Hamiltonian inclusion does not imply even Clarke's form (1.4) of the Euler–Lagrange inclusion in nonconvex and/or convexified problems. Therefore, the results obtained in this paper sharpen known conditions under the relaxability assumption and especially in the fully nonconvex setting.

Now let us explain the principal method that we use to obtain the mentioned results. This is a direct method based on *finite difference* (*discrete*) *approximations.* Such an approach to variational problems goes back to Euler, who used it in 1744 to prove the classical Euler–Lagrange equation in the calculus of variations. (Actually Leibnitz was the first to employ a similar direct method to find the brachistochrone in the very beginning of the calculus of variations; see, for example, [1]). The basic idea is as follows: (1) to replace (approximate) the original continuous-time variational problem by a "correct" sequence of finite-dimensional optimization problems that can be solved (studied) effectively, and then (2) passing to the limit with respect to approximation parameters to obtain desirable characteristics of the original variational problem.

Finite difference methods turn out to be a powerful tool for numerical solutions of infinite-dimensional variational problems. We refer the reader to the book of Polak [36] and the survey paper of Dontchev and Lempio [13], which are devoted to numerical aspects of consistent discrete approximations in optimal control. Some results in this paper are also concerned with numerical questions. In §3 we develop a discrete approximation algorithm for nonconvex differential inclusions with strong convergence properties and error estimates. But our main interest here is to use finite difference approximations as a direct vehicle for obtaining necessary optimality conditions in infinite-dimensional problem (P) via a variational analysis of nonsmooth problems in finite dimensions. Two issues are important in this approach:

(1) to construct a correct discrete approximation of problem (P) that ensures a desirable convergence of optimal solutions for discrete problems to a given local minimum for (P);

(2) to choose "right" generalized derivative (normal) constructions for nonsmooth

mappings and sets that are suitable to the method. Such constructions need to be appropriate for obtaining optimality conditions in discrete problems and also possess robustness and calculus properties for passing to the limit in the approximation procedure. Let us observe that problem (P) and its discrete counterparts are definitely objects of *nonsmooth analysis and optimization* because of a special nature of dynamic constraints like (1.2) even under smooth data in (1.1) and (1.3).

Note that not all differentiation constructions in nonsmooth analysis fit these requirements. For example, Pshenichnyi in [38] employed some tangentially generated constructions related to the contingent cone. Such constructions possess the required properties only in special situations. This allowed him to prove necessary conditions for global (actually strong) minima in autonomous differential inclusions under some restrictive assumptions close to the graph convexity of $F(\cdot, t)$. He used a discrete approximation ensuring the uniform (C-) convergence of optimal trajectories.

In Mordukhovich [27]–[29], we used generalized normals and derivatives of another nature and somewhat different algorithms to approximate Lipschitzian differential inclusions. These generalized constructions appear in (1.9), (1.10) and possess required robustness and calculus properties that are reviewed in §4. Note that if one employs the convexification of the normal cone in (1.9), i.e., uses Clarke's normal cone to the graph of a Lipschitzian mapping, then such a construction does not ensure the convergence of adjoint functions in discrete approximations (see Remark 4.6).

Although the approximation algorithm in [28] is used to prove the C-convergence of optimal trajectories in discrete approximations, its slight modification provides the *strong* L^2-convergence of the velocities. It was first observed by Smirnov [48], who obtained the refined condition (1.13) for optimal solutions to a Mayer problem involving convex autonomous differential inclusions under some additional assumptions.

In this paper, we develop the method of discrete approximations to obtain the results mentioned above in the general setting under consideration. The remainder of the paper is organized as follows.

Section 2 is devoted to the concept of intermediate local minimum for the original and relaxed problems of Bolza. We consider sufficient conditions that ensure the relaxability of (P) when a given minimum for the original problem solves the relaxed problem as well.

Section 3 deals with discrete approximations of problem (P). We provide a construction of discrete approximations and natural assumptions that ensure the strong convergence of optimal solutions with respect to the value function, trajectories, and velocities.

In §4 we describe the tools of the generalized differentiation for nonsmooth and set-valued mappings used in the paper to obtain necessary optimality conditions for discrete and differential inclusions. The reader can find there a brief review of the basic differentiation properties that are important in the method of discrete approximations.

Section 5 is concerned with necessary optimality conditions for nonsmooth finite-dimensional problems. We obtain discrete analogues of the refined Euler–Lagrange and tranversality conditions (1.14), (1.10) without the convexity operation in (1.14) and any Lipschitzian assumptions on F, f. These results turn out to be direct consequences of the Lagrange multiplier rule in nondifferentiable programming with many geometric constraints.

Section 6 is devoted to the limiting procedure in discrete approximations that allows us to prove conditions (1.10) and (1.14) for an intermediate relaxed local minimum in (P). Under relaxation stability, the results obtained characterize optimal solutions to the original problem without imposing convexity.

In §7, we study a general Mayer problem for nonconvex differential inclusions without relaxation. We prove the refined Euler–Lagrange and transversality conditions for strong (as well as for weak) local minima on the base of results in §6 and an approximation procedure involving Ekeland's variational principle. The same approach works to prove (1.13) for any boundary trajectory.

The notation in this paper is standard. The adjoint (transposed) matrix of A is denoted by A^*; the set B is always the unit closed ball of the space in question. Some special symbols are introduced and explained in §4.

2. Intermediate local minimum and relaxation.

Recall that we consider problem (P) stated above in the class of absolutely continuous functions $x : [a,b] \to \mathbf{R}^n$ (arcs) satisfying constraints (1.2)–(1.3). Any solution to (1.2) is called an (*original*) *trajectory* for the differential inclusion, and any trajectory satisfying constraints (1.3) is called a *feasible solution* to problem (P). Let us introduce a notion of local minimum for (P) studied in the paper.

DEFINITION 2.1. *The arc $\bar{x}(\cdot)$ is called an* intermediate local minimum (*i.l.m.*) *of rank $p \in [1,\infty)$ for (P) if $\bar{x}(\cdot)$ is a feasible solution to (P) and there exist numbers $\varepsilon > 0$ and $\alpha \geq 0$ such that $J[\bar{x}] \leq J[x]$ for any other feasible solution $x(\cdot)$ to (P) satisfying*

$$|x(t) - \bar{x}(t)| < \varepsilon \;\; \forall\, t \in [a,b], \tag{2.1}$$

$$\alpha \int_a^b |\dot{x}(t) - \dot{\bar{x}}(t)|^p dt < \varepsilon. \tag{2.2}$$

If (2.2) is fulfilled, then instead of (2.1) one can obviously use $|x(a) - \bar{x}(a)| < \varepsilon$. Relationships (2.1), (2.2) mean that we consider a neighborhood of $\bar{x}(\cdot)$ in the Sobolev space $W^{1,p}$ of absolute continuous functions $x : [a,b] \to \mathbf{R}^n$ equipped with a natural norm. If there is only requirement (2.1) in Definition 2.1 (i.e., $\alpha = 0$), then one gets a *strong* local minimum (with respect to the C-norm). This actually corresponds to the L^1-weak topology for derivatives instead of the strong (L^p-norm) topology in (2.2). Obviously any *optimal solution* to problem (P) (global minimum) provides a strong local minimum (and, therefore, an i.l.m.) for (P). As we know, most necessary optimality conditions for differential inclusions are obtained for strong local minima, but it is not clear a priori whether they hold for i.l.m.

If instead of (2.2) one sets the more restrictive requirement

$$|\dot{x}(t) - \dot{\bar{x}}(t)| < \varepsilon \;\; \text{a.e.} \;\; t \in [a,b],$$

then we have a *weak* local minimum in the framework of Definition 2.1. This corresponds to considering a neighborhood of $\bar{x}(\cdot)$ in the space $W^{1,\infty}$ with the L^∞-norm for derivatives (or the C^1-norm for continuously differentiable functions in classical variational problems). Therefore, the notion of i.l.m. that we introduced takes (for any rank $p \in [1,\infty)$) an intermediate place between the familiar concepts of strong ($\alpha = 0$) and weak ($p = \alpha$) minima. Note that some aspects of this setting are related to the (local) *Lavrentiev phenomenon* in the calculus of variations.

The following example shows that the class of intermediate local minimizers differs from that of weak local minimizers in classical variational problems.

Example 2.2. Consider the simplest problem of the calculus of variations:

$$\text{minimize } J[x] = \int_0^1 \dot{x}^3(t)dt \text{ subject to } x(0) = 0, \ \ x(1) = 1.$$

It is easy to conclude that the function $\bar{x}(t) = t$ provides a weak local minimum for this problem; see [21, §2.2.2]. Now taking the functions $x_k(t) = \bar{x}(t) + y_k(t)$ with $y_k(0) = y_k(1) = 0$ and

$$\dot{y}_k(t) = \begin{cases} -\sqrt{n} & \text{if } 0 \le t \le 1/n, \\ \sqrt{n}(n-1)^{-1} & \text{if } 1/n < t \le 1, \end{cases}$$

one can check that

$$J[x_k] = \sqrt{n} + O(1) \to -\infty$$

and

$$\int_0^1 |\dot{x}_k(t) - \dot{\bar{x}}(t)|^p dt \to 0 \text{ as } k \to \infty$$

for each $p \in [1, \infty)$. Therefore, the extremal $\bar{x}(t)$ does not provide an intermediate local minimum of any rank $p \in [1, \infty)$ for the example considered.

On the other hand, intermediate local minimizers may not be strong local minimizers even for *convex* and Lipschitzian differential inclusions. Such examples are constructed by Vinter and Woodford [49] in both bounded (autonomous) and unbounded cases. They also distinguish intermediate local minimizers of different rank for some unbounded (but integrably bounded) differential inclusions.

Now we consider an extension of the original problem (P) in the line well known in the calculus of variations and optimal control (cf., e.g., [4], [6], [15], [20], [50], [52]). Let

$$f_F(x, v, t) := f(x, v, t) + \delta(v, F(x, t)) \tag{2.3}$$

where $\delta(v, \Lambda) = 0$ if $v \in \Lambda$ and $\delta(v, \Lambda) = \infty$ if $v \notin \Lambda$ (the indicator function). Denote by $\hat{f}_F(x, v, t)$ the *convexification* (the biconjugate function) for f_F in the v variable, i.e., the largest convex function majorized by $f_F(x, \cdot, t)$ for each x and t. Along with the original problem (P), we consider its *relaxation* (R) as follows:

$$\text{minimize } \hat{J}[x] := \varphi(x(a), x(b)) + \int_a^b \hat{f}_F(x(t), \dot{x}(t), t)dt \tag{2.4}$$

over absolutely continuous functions on $[a, b]$ under endpoint constraints (1.3). Note that if $\hat{J}[x] < \infty$, then $x(\cdot)$ satisfies the convexified differential inclusion

$$\dot{x}(t) \in \text{co}F(x(t), t) \ \text{ a.e. } t \in [a, b]. \tag{2.5}$$

Any trajectory for (2.5) is called a *relaxed trajectory* for (1.2). It is well known that under natural assumptions involving Lipschitzness of F in x, the following *approximation property* holds: Every relaxed trajectory $x(\cdot)$ can be uniformly approximated in $[a, b]$ by original trajectories $x_k(\cdot)$ starting with the same initial state (but may not satisfy endpoint constraints) such that

$$\underline{\lim} \int_a^b f(x_k(t), \dot{x}_k(t), t)dt \le \int_a^b \hat{f}_F(x(t), \dot{x}(t), t)dt \text{ as } k \to \infty. \tag{2.6}$$

DEFINITION 2.3. *The arc* $\bar{x}(\cdot)$ *is called an* intermediate relaxed local minimum *(i.r.l.m.) of rank* $p \in [1, \infty)$ *for the original problem* (P) *if* $\bar{x}(\cdot)$ *is a feasible solution of* (P) *and provides an intermediate local minimum of rank p for the relaxed problem* (R) *with* $J[\bar{x}] = \hat{J}[\bar{x}]$.

It is essential in Definition 2.3 that $\bar{x}(\cdot)$ is an original trajectory for (1.2). Obviously, there is no difference between i.r.l.m. and i.l.m. if problem (P) is convex in the following sense: for each $t \in [a, b]$ and x around $\bar{x}(t)$, the function f is convex in v on the convex set $F(x, t)$. One can see that the refined Euler–Lagrange inclusion that is obtained for i.r.l.m. may be different from its counterpart for arbitrary i.l.m. in the relaxed problem. This happens because the normal cone to the graph of F and to the graph of co$F$ is not the same. On the other hand, we cannot guarantee, in principle, that necessary conditions for i.r.l.m. will work for arbitrary i.l.m. (or strong minima) for the original problem. Nevertheless, the latter holds in rather general settings without any convexity assumptions. Actually this is related to the property of "hidden convexity" inherent in continuous-time systems like (1.2). In this paper, we use only one result in this direction going back to the classical Bogoljubov theorem [4].

PROPOSITION 2.4. *Let* $\bar{x}(\cdot) \in W^{1,\infty}[a, b]$ *be a strong minimum for problem* (1.1), (1.3) *where the integrand* $f(x, v, t)$ *is continuous in* (x, v) *around* $(\bar{x}(t), \dot{\bar{x}}(t))$ *uniformly in* $[a, b]$ *and measurable in t. Then* $\bar{x}(\cdot)$ *is a strong minimum for the relaxed problem* (1.3), (2.4) *with* $\hat{f}_F = \hat{f}$ *and* $J[\bar{x}] = \hat{J}[\bar{x}]$.

Proof. According to the version of Bogoljubov's theorem in [21, §9.2.4], for any $x(\cdot) \in W^{1,\infty}[a, b]$ one can find a sequence of $x_k(\cdot) \in W^{1,\infty}[a, b]$ such that $x_k(a) = x(a)$, $x_k(b) = x(b)$, $x_k(\cdot)$ converge to $x(\cdot)$ uniformly in $[a, b]$, and (2.6) is fulfilled in the case where $\hat{f}_F = \hat{f}$. If $\bar{x}(\cdot)$ is not a strong minimum for the relaxed problem (2.4), (1.3) or/and $\hat{J}[\bar{x}] < J[\bar{x}]$, then there exists a function $x(\cdot) \in W^{1,\infty}[a, b]$ with $\hat{J}[x] < J[\bar{x}]$ such that $x(\cdot)$ belongs to a C-neighborhood of $\bar{x}(\cdot)$ and satisfies constraints (1.3). This contradicts the strong minimality of $\bar{x}(\cdot)$ in the original problem, thanks to the Bogoljubov approximation for $x(\cdot)$. □

There are several generalizations and analogues of Bogoljubov's theorem that have many important applications to optimal control systems and differential inclusions; see, for example, [6], [15], [20], [28], [50] and references therein. They lead to the property of *relaxation stability* (or proper relaxation) when an optimal solution to the original problem solves the relaxed problem as well with the same optimal value. For problems (P) involving differential inclusions with endpoint constraints at either $t = a$ or $t = b$, such a relaxation stability follows directly from the approximation property for relaxed trajectories stated above.

For problems with general endpoint constraints, the relaxation stability is ensured by the *calmness* property in Clarke [6], [8]. The latter property is fulfilled for "almost all" endpoint constraints (at least of inequality type) and shows that the relaxation stability may fail only for ill-posed problems where small perturbations of boundary conditions produce proportionally unbounded variations of the minimum. According to Clarke [8], the calmness hypothesis implies that corresponding necessary optimality conditions can be taken to be *normal.* A general result that "normality implies relaxation stability" for optimal control systems has been obtained by Warga [51].

For special classes of problems (P) with arbitrary endpoint constraints, the relaxation stability holds without any calmness or normality assumptions. In particular, let differential inclusion (1.2) be represented in the *linear* form

$$\dot{x}(t) \in F_1(t)x(t) + F_2(t) \text{ a.e. } t \in [a, b]$$

where the multifunctions F_1 and F_2 are integrable in $[a,b]$, F_1 is convex-valued while F_2 is not. If, moreover, the function f in (1.1) is convex in v, then any of such problems (P) possesses the property of relaxation stability. This can be proved by using Aumann's theorem about set-valued integrals; cf. arguments in [28, Thm. 19.7]. The same situation holds for problems (P) involving nonlinear one-dimensional differential inclusions; see Remark 19.2 in [28].

3. Discrete approximations. In this section we construct a sequence of discrete approximations for the original problem of Bolza such that optimal solutions to discrete approximations converge in $W^{1,p}$ to a given i.r.l.m. for problem (P).

First let us consider a fixed original trajectory $\bar{x}(\cdot)$ for (1.2) and prove that it can be approximated by trajectories for corresponding discrete inclusions. To do this, we assume that the multifunction $F(x,t)$ is bounded and locally Lipschitzian in x around $\bar{x}(\cdot)$ and is Hausdorff continuous in t a.e. on $[a,b]$. More precisely, we impose the following hypotheses:

(H1) There are an open set $U \subset \mathbf{R}^n$ and positive numbers l_F, m_F such that $\bar{x}(t) \in U$ for any $t \in [a,b]$, the sets $F(x,t)$ are closed for all $(x,t) \in U \times [a,b]$, and one has

$$F(x,t) \subset m_F B \quad \forall (x,t) \in U \times [a,b] \tag{3.1}$$

and

$$F(x_1,t) \subset F(x_2,t) + l_F|x_1 - x_2|B \quad \forall x_1, x_2 \in U, \quad t \in [a,b]. \tag{3.2}$$

(H2) The multifunction $F(x,\cdot)$ is Hausdorff continuous for a.e. $t \in [a,b]$ uniformly in $x \in U$.

Following Dontchev and Farkhi [12], let us consider the so-called *averaged modulus of continuity* for the multifunction $F(x,t)$ in $t \in [a,b]$ when $x \in U$. This modulus $\tau(F;h)$ that depends on the parameter $h > 0$ is defined as follows:

$$\tau(F;h) := \int_a^b \sigma(F;t,h)dt \tag{3.3}$$

where $\sigma(F;t,h) := \sup\{\omega(F;x,t,h) | x \in U\}$, where

$$\omega(F;x,t,h) := \sup\{\text{haus}(F(x,t'),F(x,t'')) | t',t'' \in [t-h/2, t+h/2] \cap [a,b]\},$$

and where haus$(\cdot,\cdot)$ is the Hausdorff distance between compact sets.

It is proved in [12] that if $F(x,\cdot)$ is Hausdorff continuous for a.e. $t \in [a,b]$ uniformly in $x \in U$, then $\tau(F;h) \to 0$ as $h \to 0$.

Note that in the case of single-valued functions $f(t)$ not depending on x, the construction $\tau(f;h)$ in (3.3) was developed in Sendov and Popov [47] under the name of "averaged modulus of smoothness." It was proved in [47] that $\tau(f;h) \to 0$ as $h \to 0$ if and only if f is Riemann integrable on $[a,b]$. The latter is equivalent to f being continuous for a.e. $t \in [a,b]$. In this paper, we use the name "averaged modulus of continuity" for both single-valued and multivalued cases.

Now let us construct a finite difference (discrete) approximation for the given differential inclusion using the replacement of the derivative in (1.2) by the *Euler finite difference*

$$\dot{x}(t) \approx [x(t+h) - x(t)]/h.$$

For any natural number $k = 1, 2, \ldots$, we consider a uniform grid $T_k := \{t_j | j = 0, 1, \ldots, k\}$ with $t_0 = a$, $t_k = b$ and stepsize

$$h_k := (b-a)/k = t_{j+1} - t_j \quad (j = 0, \ldots, k-1).$$

The associate discrete inclusion is as follows:

$$x_{j+1}^k \in x_j^k + h_k F(x_j^k, t_j), \quad j = 0, \ldots, k-1. \tag{3.4}$$

THEOREM 3.1. *Let $\bar{x}(\cdot)$ be a trajectory for* (1.2) *under hypotheses* (H1) *and* (H2). *Then there is a sequence $\{z_j^k | j = 0, \ldots, k\}$, $k = 1, 2, \ldots$, of solutions to discrete inclusions* (3.4) *such that $z_0^k = \bar{x}(a)$ and the functions*

$$v^k(t) := (z_{j+1}^k - z_j^k)/h_k, \quad t \in [t_j, t_{j+1}), \quad j = 0, \ldots, k-1, \tag{3.5}$$

converge to $\dot{\bar{x}}(\cdot)$ as $k \to \infty$ in the norm topology of $L^1[a,b]$.

Proof. Let $\{w^k(\cdot)\}$, $k = 1, 2, \ldots$, be an arbitrary sequence of functions in $[a,b]$ such that $w^k(t)$ are constant in $[t_j, t_{j+1})$ for every $j = 0, \ldots, k-1$ and $w^k(t)$ converge to $\dot{\bar{x}}(t)$ as $k \to \infty$ in the norm of $L^\infty[a,b]$. Such a sequence always exists because of the density of step-functions in $L^\infty[a,b]$. Employing (3.1), one gets

$$|w^k(t)| \le m_F + 1 \;\; \forall t \in [a,b] \;\text{ as }\; k \to \infty. \tag{3.6}$$

In the arguments and estimates below, we use the number

$$\xi_k := \int_a^b |\dot{\bar{x}}(t) - w^k(t)| dt \to 0 \;\text{ as }\; k \to \infty. \tag{3.7}$$

Let us define the discrete functions $\{y_j^k | j = 0, \ldots, k\}$ as follows:

$$y_{j+1}^k = y_j^k + h_k w_j^k, \quad j = 0, \ldots, k-1, \quad y_0^k = \bar{x}(a) \tag{3.8}$$

where $w_j^k := w^k(t_j)$, $j = 0, \ldots, k-1$. Note that the functions

$$y^k(t) := \bar{x}(a) + \int_a^t w^k(s) ds, \quad a \le t \le b,$$

are piecewise linear extensions of (3.8) on the interval $[a,b]$ and

$$|y^k(t) - \bar{x}(t)| \le \xi_k \;\; \forall t \in [a,b]. \tag{3.9}$$

Therefore, $y^k(t) \in U$ for all $t \in [a,b]$ if k is big enough.

Denote by $\operatorname{dist}(w, F)$ the *Euclidean distance* between the point w and the closed set F. It is well known that the Lipschitz condition (3.2) is equivalent to

$$\operatorname{dist}(w, F(x_1, t)) \le \operatorname{dist}(w, F(x_2, t)) + l_F |x_1 - x_2| \;\; \forall w \in \mathbf{R}^n, \;\; x_1, x_2 \in U, \;\; t \in [a,b].$$

For any $w, x \in \mathbf{R}^n$ and $t_1, t_2 \in [a,b]$, one obviously has

$$\operatorname{dist}(w, F(x, t_1)) \le \operatorname{dist}(w, F(x, t_2)) + \operatorname{haus}(F(x, t_1), F(x, t_2)).$$

Now using (3.3), we get

$$\varsigma_k := \sum_{j=0}^{k-1} h_k \mathrm{dist}(w_j^k, F(y_j^k, t_j)) = \sum_{j=0}^{k-1} \int_{t_j}^{t_{j+1}} \mathrm{dist}(w_j^k, F(y_j^k, t_j))dt$$
$$\leq \sum_{j=0}^{k-1} \int_{t_j}^{t_{j+1}} \mathrm{dist}(w_j^k, F(y_j^k, t))dt + \tau(F; h_k).$$

It follows from (3.2), (3.6), and (3.9) that

$$\mathrm{dist}(w_j^k, F(y_j^k, t)) \leq \mathrm{dist}(w^k(t), F(y^k(t), t)) + l_F(m_F + 1)(t - t_j) \;\; \forall t \in [t_j, t_{j+1})$$

and

$$\mathrm{dist}(w^k(t), F(y^k(t), t)) \leq \mathrm{dist}(w^k(t), F(\bar{x}(t), t)) + l_F|y^k(t) - \bar{x}(t)|$$
$$\leq |w^k(t) - \dot{\bar{x}}(t)| + l_F \xi_k \;\; \text{a.e. } t \in [a, b].$$

Therefore, we have the estimate

$$(3.10) \qquad \varsigma_k \leq \gamma_k := (1 + l_F)\xi_k + l_F(b - a)(m_F + 1)h_k/2 + \tau(F; h_k).$$

Note that functions (3.8) are not trajectories for (3.4) because one does not get $w_j^k \in F(y_j^k, t_j)$. Now we use w_j^k to define trajectories for (3.4), which are close to y_j^k and have the convergence property stated in this theorem.

Let us construct the desirable trajectories $\{z_j^k | j = 0, \ldots, k\}$ using the following *proximal algorithm*:

$$(3.11) \qquad z_0^k = \bar{x}(a), \; v_j^k \in F(z_j^k, t_j) \text{ with } |v_j^k - w_j^k| = \mathrm{dist}(w_j^k, F(z_j^k, t_j)),$$

$$z_{j+1}^k = z_j^k + h_k v_j^k, \qquad j = 0, \ldots, k - 1.$$

Note that in (3.11) we take *projections of velocities* as in [28], [48] instead of projections of states as in [38], [12]. This will allow us to prove a strong convergence of discrete approximations with respect to velocities.

First we prove that algorithm (3.11) keeps $\{z_j^k | j = 0, \ldots, k\}$ inside the neighborhood U from (H1) if k is big enough. Indeed, we consider any number k such that $\bar{x}(t) + \eta_k B \subset U$ for all $t \in [a, b]$ where

$$\eta_k := \gamma_k \exp[l_F(b - a)] + \xi_k,$$

where ξ_k and γ_k are defined in (3.7) and (3.10), respectively. One can see that $\eta_k \to 0$ as $k \to \infty$ because $\tau(F; h_k) \to 0$ under assumption (H2).

By induction, let us show that if $z_m^k \in U$ for all $m = 0, \ldots, j$, then this also holds for $m = j + 1$. Using (3.2), (3.10), and (3.11), one gets

$$|z_{j+1}^k - y_{j+1}^k| \leq |z_j^k - y_j^k| + h_k \mathrm{dist}(w_j^k, F(z_j^k, t_j)) \leq |z_j^k - y_j^k| + h_k(\mathrm{dist}(w_j^k, F(y_j^k, t_j))$$
$$+ l_F|z_j^k - y_j^k|) \leq \cdots \leq h_k \sum_{m=0}^{j} (1 + l_F h_k)^{j-m} \mathrm{dist}(w_m^k, F(y_m^k, t_m))$$
$$\leq \exp[l_F(b - a)] \sum_{m=0}^{j} h_k \mathrm{dist}(w_m^k, F(y_m^k, t_m)) \leq \gamma_k \exp[l_F(b - a)].$$

Due to (3.9), the latter implies

$$(3.12)\qquad |z_{j+1}^k - \bar{x}(t_{j+1})| \le \gamma_k \exp[l(b-a)] + \xi_k := \eta_k,$$

which proves that $z_j^k \in U$ for all $j \in \{0,\dots,k\}$. Taking this into account, one can extract from the previous arguments the following estimate:

$$(3.13)\qquad \sum_{j=0}^{k} |z_j^k - y_j^k| \le (b-a)\exp[l_F(b-a)] \sum_{j=0}^{k-1} \operatorname{dist}(w_j^k, F(y_j^k, t_j)).$$

Now let us estimate the quantity $\vartheta_k := \int_a^b |v^k(t) - w^k(t)|dt$ where the functions $v^k(t)$ are defined in (3.5). Employing (3.10), (3.11), and (3.13), we get

$$\vartheta_k = \sum_{j=0}^{k-1} h_k|v_j^k - w_j^k| = \sum_{j=0}^{k-1} h_k \operatorname{dist}(w_j^k, F(z_j^k, t_j)) \le \sum_{j=0}^{k-1} h_k \operatorname{dist}(w_j^k, F(y_j^k, t_j))$$

$$+ l_F \sum_{j=0}^{k-1} h_k |z_j^k - y_j^k| \le \gamma_k(1 + l_F(b-a)\exp[l_F(b-a)]).$$

Thus we obtain the final estimate

$$(3.14)\quad \alpha_k := \int_a^b |v^k(t) - \dot{\bar{x}}(t)|dt \le \beta_k := \xi_k + \gamma_k(1 + l_F(b-a)\exp[l_F(b-a)]).$$

This ensures the L^1-convergence $v^k(\cdot) \to \dot{\bar{x}}(\cdot)$ due to (3.7) and $\tau(F; h_k) \to 0$ as $k \to \infty$ under (H2). □

Remark 3.2. The result obtained provides a *strong approximation* with respect to velocities of any absolutely continuous trajectory for the differential inclusion (1.2) by discrete trajectories for its Euler difference counterparts (3.4). Note that the error estimate for velocities (3.14) immediately implies the following estimate

$$|z^k(t) - \bar{x}(t)| \le \beta_k \;\; \forall t \in [a,b]$$

for the corresponding motions $z^k(t) := \bar{x}(a) + \int_a^t v^k(s)ds$, which are piecewise linear extensions of discrete trajectories (3.11). One can see that the *numerical efficiency* of the estimates obtained depends on the evaluation of $\tau(F;h)$ and the approximation accuracy in (3.7).

It has been proved in [12] that $\tau(F;h) = O(h)$ if $F(x,\cdot)$ is of *bounded variation* on $[a,b]$ uniformly in $x \in U$. Using the technique for averaged moduli of continuity (smoothness) developed in [47], one can obtain effective estimates for ξ_k in (3.7). Indeed, if $\dot{\bar{x}}(\cdot)$ is *Riemann integrable* on $[a,b]$, then we always get $\xi_k \le 2\tau(\dot{\bar{x}}; h_k)$, taking $v^k(t) = \dot{\bar{x}}(t_j)$ for $t \in [t_j, t_j + h_k)$ as $j = 0,\dots,k-1$.

Now we consider the given original trajectory $\bar{x}(\cdot)$, which is an i.r.l.m. of some rank $p \in [1,\infty)$ for problem (P). One can easily see that under boundedness assumption (3.1), the notion of i.r.l.m. for (P) *does not depend on rank* p. This means that if $\bar{x}(\cdot)$ is an i.r.l.m. of some rank $p \in [1,\infty)$, then it will be an i.r.l.m. of any other rank from $[1,\infty)$. In what follows we always take $p = 2$ and set $\alpha = 1$ in (2.2) for simplicity.

Let us construct a sequence of optimization problems (P_k) for discrete inclusions (3.4) such that optimal solutions to (P_k) *strongly* (in $W^{1,2}[a,b]$) converge to $\bar{x}(\cdot)$ as $k \to \infty$. Take a number ε in (2.1), (2.2) for the given i.r.l.m; and assume (H1), (H2) along $\bar{x}(\cdot)$. One can always suppose that $\bar{x}(t) + \varepsilon/2 \in U$ for all $t \in [a,b]$. Using Theorem 3.1, we approximate $\bar{x}(\cdot)$ by discrete trajectories $\{z_j^k | j = 0, \dots, k\}$ and compute the numbers η_k in (3.12). Now we define a sequence of discrete approximation problems (P_k), $k = 1, 2, \dots$, as follows:

$$\text{(3.15)} \qquad \text{minimize } J_k[x^k] := \varphi(x_0^k, x_k^k) + |x_0^k - \bar{x}(a)|^2$$

$$+h_k \sum_{j=0}^{k-1} f(x_j^k, (x_{j+1}^k - x_j^k)/h_k, t_j) + \sum_{j=0}^{k-1} \int_{t_j}^{t_{j+1}} |(x_{j+1}^k - x_j^k)/h_k - \dot{\bar{x}}(t)|^2 dt$$

over discrete trajectories $x^k = (x_0^k, x_1^k, \dots, x_k^k)$ for the difference inclusion (3.4) subject to the constraints

$$\text{(3.16)} \qquad (x_0^k, x_k^k) \in \Omega_k := \Omega + \eta_k B,$$

$$\text{(3.17)} \qquad |x_j^k - \bar{x}(t_j)| \le \varepsilon/2 \text{ for } j = 0, \dots, k,$$

and

$$\text{(3.18)} \qquad \sum_{j=0}^{k-1} \int_{t_j}^{t_{j+1}} |(x_{j+1}^k - x_j^k)/h_k - \dot{\bar{x}}(t)|^2 dt \le \varepsilon/2.$$

Let $x^k(\cdot)$ be the piecewise linear extension of the discrete trajectory $\{x_j^k | j = 0, \dots, k\}$ on $[a,b]$, and let $\dot{x}^k(\cdot)$ denote the piecewise constant extension of the "velocity" $(x_{j+1}^k - x_j^k)/h_k$. One has

$$\dot{x}^k(t) = (x_{j+1}^k - x_j^k)/h_k = \dot{x}^k(t_j) \text{ for any } t \in [t_j, t_{j+1}).$$

We are going to consider the (strong) $W^{1,2}$-convergence of $x^k(\cdot)$ to some absolutely continuous function $x(\cdot)$ in $[a,b]$. This means that $x^k(a) = x_0^k \to x(a)$ and $\dot{x}^k(\cdot) \to \dot{x}(\cdot)$ in $L^2[a,b]$ as $k \to \infty$. The latter obviously implies that $x^k(\cdot)$ converge to $x(\cdot)$ uniformly in $[a,b]$.

In addition to (H1) and (H2), now we impose the following hypotheses on f, φ, and Ω:

(H3) $f(x,v,\cdot)$ is continuous for a.e. $t \in [a,b]$ and bounded uniformly in $(x,v) \in U \times (m_F B)$.

(H4) There exists $\nu > 0$ such that the function $f(\cdot,\cdot,t)$ is continuous on the set

$$\text{(3.19)} \quad A_\nu(t) := \{(x,v) \in U \times (m_F + \nu)B | \; v \in F(x,t') \text{ for some } t' \in (t-\nu, t]\}$$

uniformly in $t \in [a,b]$.

(H5) φ is continuous on $U \times U$ and Ω is closed around $(\bar{x}(a), \bar{x}(b))$.

THEOREM 3.3. *Let $\bar{x}(\cdot)$ be an i.r.l.m. for problem (P), and let hypotheses* (H1)–(H5) *be fulfilled. Then any sequence $\{\bar{x}^k(\cdot)\}$, $k = 1, 2, \dots$, of optimal solutions to (P_k) converges to $\bar{x}(\cdot)$ in the space $W^{1,2}[a,b]$ as $k \to \infty$.*

Proof. First let us prove that for each k big enough, the discrete trajectory $\{z_j^k | j = 0, \ldots, k\}$ constructed in Theorem 3.1 is a feasible solution of (P_k). We need to check that this trajectory satisfies constraints (3.16)–(3.18). For the case of (3.16), it follows directly from (3.12). Taking k such that $\eta_k \leq \varepsilon/2$, we also get (3.17) from (3.12). By virtue of (3.1) and (3.14), constraint (3.18) for $x^k = z^k$ is reduced to

$$\int_a^b |v^k(t) - \dot{\bar{x}}(t)|^2 dt \leq 2m_F\alpha_k \leq 2m_F\beta_k \leq \varepsilon/2.$$

The latter is fulfilled for big numbers k due to the expression for β_k in (3.14). Therefore, x^k is a feasible solution to (P_k) for all k big enough. According to the classical Weierstrass theorem, we can conclude that there is an optimal solution $\bar{x}^k$ to (P_k) for such k under the assumptions made.

Let us prove that for any sequence of optimal solutions $\bar{x}^k$ to (P_k) one has

$$\overline{\lim}\, J_k[\bar{x}^k] \leq J[\bar{x}] \text{ as } k \to \infty. \tag{3.20}$$

To accomplish this, it suffices to show that

$$J_k[z_k] \to J[\bar{x}] \text{ as } k \to \infty \tag{3.21}$$

for the sequence of discrete trajectories z^k approximating $\bar{x}(\cdot)$ by virtue of Theorem 3.1.

Let us consider expression (3.15) for $J_k[z^k]$. Due to continuity of φ one has

$$\varphi(z_0^k, z_k^k) \to \varphi(\bar{x}(a), \bar{x}(b)) \text{ as } k \to \infty.$$

Further, the second term in this expression vanishes; the fourth term tends to zero as $k \to \infty$ because of (3.5) and (3.14). To justify (3.21), it remains to prove that

$$\sigma_k := h_k \sum_{j=0}^{k-1} f(z_j^k, (z_{j+1}^k - z_j^k)/h_k, t_j) \to \int_a^b f(\bar{x}(t), \dot{\bar{x}}(t), t)\, dt \text{ as } k \to \infty$$

under assumptions (H1)–(H4). Note that (H3) implies $\tau(f; h_k) \to 0$ as $k \to \infty$ for modulus (3.3). In what follows we use the sign "$\sim$" for expressions that are equivalent as $k \to \infty$. Due to (3.5), (3.12), and (3.14) one gets

$$\sigma_k = \sum_{j=0}^{k-1} \int_{t_j}^{t_{j+1}} f(z_j^k, v^k(t), t_j) dt \sim \sum_{j=0}^{k-1} \int_{t_j}^{t_{j+1}} f(z_j^k, v^k(t), t) dt + \tau(f; h_k)$$

$$\sim \sum_{j=0}^{k-1} \int_{t_j}^{t_{j+1}} f(\bar{x}(t), v^k(t), t) dt = \int_a^b f(\bar{x}(t), v^k(t), t) dt \sim \int_a^b f(\bar{x}(t), \dot{\bar{x}}(t), t) dt.$$

The last statement holds by virtue of the classical Lebesgue limiting theorem because $\{v^k(\cdot)\}$ contains a subsequence converging for a.e. $t \in [a, b]$. Therefore, we obtain (3.21), which implies (3.20).

In the arguments above we have not actually used the property of $\bar{x}(\cdot)$ to be an i.r.l.m. for (P). Now let us prove that in the latter case, inequality (3.20) implies

$$\lim_{k\to\infty} [c_k := |\bar{x}^k(a) - \bar{x}(a)|^2 + \int_a^b |\dot{\bar{x}}^k(t) - \dot{\bar{x}}(t)|^2 dt] = 0, \tag{3.22}$$

i.e., $\bar{x}^k(\cdot)$ converge to $\bar{x}(\cdot)$ in the norm of $W^{1,2}[a,b]$. Suppose that it is not true, and consider a limiting point $c > 0$ of the sequence $\{c_k\}$ in (3.22). Let, for simplicity, $c = \lim c_k$ for all $k \to \infty$.

Because of (3.17) and (3.18) we claim the existence of an absolutely continuous function $\tilde{x}(t)$ in $[a,b]$ such that $\bar{x}^k(\cdot) \to \tilde{x}(\cdot)$ uniformly in $[a,b]$ and $\dot{\bar{x}}^k(\cdot) \to \dot{\tilde{x}}(\cdot)$ weakly in $L^2[a,b]$ as $k \to \infty$ (we take all k without loss of generality). According to the classical Mazur theorem, there is a sequence of convex combinations of $\dot{\bar{x}}^k(\cdot)$ that converges to $\dot{\tilde{x}}(\cdot)$ in the norm topology of $L^2[a,b]$. Hence it contains a subsequence converging to $\dot{\tilde{x}}(\cdot)$ for a.e. $t \in [a,b]$.

Using these facts and taking into account that

$$h_k \sum_{j=0}^{k-1} f(\bar{x}_j^k, (\bar{x}_{j+1}^k - \bar{x}_j^k)/h_k, t_j) \sim \int_a^b f(\bar{x}^k(t), \dot{\bar{x}}^k(t), t)dt \text{ as } k \to \infty$$

and also the definition of $\hat{f}_F$ for (2.3), we get

$$\text{(3.23)} \quad \int_a^b \hat{f}_F(\tilde{x}(t), \dot{\tilde{x}}(t), t)dt \le \underline{\lim}\, h_k \sum_{j=0}^{k-1} f(\bar{x}_j^k, (\bar{x}_{j+1}^k - \bar{x}_j^k)/h_k, t_j) \text{ as } k \to \infty,$$

where $\tilde{x}(\cdot)$ satisfies the convexified differential inclusion (2.5).

Observe that the integral functional

$$I[v] := \int_a^b |v(t) - \dot{\bar{x}}(t)|^2 dt$$

is lower semicontinuous in the weak topology of $L^2[a,b]$ due to the convexity of the integrand in v. Since

$$\sum_{j=0}^{k-1} \int_{t_j}^{t_{j+1}} |(\bar{x}_{j+1}^k - \bar{x}_j^k)/h_k - \dot{\bar{x}}(t)|^2 dt = \int_a^b |\dot{\bar{x}}^k(t) - \dot{\bar{x}}(t)|^2 dt,$$

the latter implies that

$$\text{(3.24)} \int_a^b |\dot{\tilde{x}}(t) - \dot{\bar{x}}(t)|^2 dt \le \underline{\lim} \sum_{j=0}^{k-1} \int_{t_j}^{t_{j+1}} |(\bar{x}_{j+1}^k - \bar{x}_j^k)/h_k - \dot{\bar{x}}(t)|^2 dt \text{ as } k \to \infty.$$

Now passing to the limit in (3.16)–(3.18) as $k \to \infty$ and using (3.24) as well as (H5), we get that $\tilde{x}(\cdot)$ satisfies constraints (1.3) and

$$|\tilde{x}(t) - \bar{x}(t)| \le \varepsilon/2 \text{ for } t \in [a,b], \quad \int_a^b |\dot{\tilde{x}}(t) - \dot{\bar{x}}(t)|^2 dt \le \varepsilon/2.$$

The latter means that $\tilde{x}(\cdot)$ belongs to the given neighborhood of $\bar{x}(\cdot)$ in $W^{1,2}[a,b]$. Moreover, (3.23) implies

$$\text{(3.25)} \qquad \varphi(\tilde{x}(a), \tilde{x}(b)) + \int_a^b \hat{f}_F(\tilde{x}(t), \dot{\tilde{x}}(t), t)dt + c \le \underline{\lim}\, J_k[\bar{x}^k] \text{ as } k \to \infty.$$

Due to (3.20), (3.25), and $c > 0$ we get $\hat{J}[\tilde{x}] < J[\bar{x}]$. But this is impossible because $\bar{x}(\cdot)$ is an i.r.l.m. for (P). Therefore, one has $c = 0$, which establishes (3.22) and ends the proof of the theorem. □

Remark 3.4. In the convergence result of Theorem 3.3, one can avoid the continuity hypothesis (H3) on f in t by changing the approximation

$$h_k \sum_{j=0}^{k-1} f(x_j^k, (x_{j+1}^k - x_j^k)/h_k, t_j) \text{ for } \sum_{j=0}^{k-1} \int_{t_j}^{t_{j+1}} f(x_j^k, (x_{j+1}^k - x_j^k)/h_k, t)\, dt.$$

Indeed, we can handle the latter approximations in the same way as the last term in (3.15) under the measurability assumption on f in t.

The convergence theorem that we proved allows us to make a bridge between variational problems for differential inclusions and dynamic optimization problems in finite dimensions. The latter can be reduced to finite-dimensional problems of mathematical programming with *many geometric constraints.* The mathematical programming problems obtained in this way turn out to be objects of *nonsmooth optimization* even in the case of smooth initial data in the original problem (P).

For variational analysis of these problems and then for passing to the limit in optimality conditions as $k \to \infty$, we need to use generalized differential constructions with special properties. They are considered in the next section.

4. Tools of variational analysis. This section is concerned with tools of generalized differentiation that are appropriate for the main objectives of the research. The results reviewed are mostly connected with the approach in Mordukhovich [26]–[28] and recent developments in [30]–[34] dealing with *nonconvex*-valued generalized differential constructions. The reader can also consult Clarke [8], [10], Ioffe [17]–[19], Loewen [23], Rockafellar [40], [44], and Rockafellar and Wets [45] for related and additional material.

Let Ω be a nonempty set in $\mathbf{R}^n$, and let

$$\Pi(x, \Omega) := \{\omega \in \mathrm{cl}\Omega \text{ such that } |x - \omega| = \mathrm{dist}(x, \Omega)\} \tag{4.1}$$

be the Euclidean projector of x on clΩ. In the following definition, "cone" stands for the conic hull of a set and "Limsup" denotes the well-known Kuratowski–Painlevé upper limit for multifunctions.

DEFINITION 4.1. *Given $\bar{x} \in \mathrm{cl}\Omega$, the closed cone*

$$N(\bar{x}; \Omega) := \mathrm{Limsup}_{x \to \bar{x}}[\mathrm{cone}(x - \Pi(x, \Omega))] \tag{4.2}$$

is called the normal cone *to the set Ω at the point $\bar{x}$.*

If Ω is convex, then (4.2) is reduced to the normal cone of convex analysis. In general, the convex closure of (4.2) coincides with the *Clarke normal cone*:

$$N_C(\bar{x}; \Omega) = \mathrm{clco}N(\bar{x}; \Omega). \tag{4.3}$$

Note that, in contrast to (4.3), the normal cone (4.2) is always *robust* with respect to perturbations of $\bar{x}$, i.e., the multifunction $N(\cdot; \Omega)$ has closed graph.

DEFINITION 4.2. *Let $f : \mathbf{R}^n \to \bar{\mathbf{R}}$ be an extended-real-valued function, and let $|f(\bar{x})| < \infty$. The set*

$$\partial f(\bar{x}) := \{x^* \in \mathbf{R}^n | (x^*, -1) \in N((\bar{x}, f(\bar{x})); \mathrm{epi}f)\} \tag{4.4}$$

is called the subdifferential *of f at $\bar{x}$. If $|f(\bar{x})| = \infty$, we put $\partial f(\bar{x}) = \emptyset$.*

Observe that for continuous functions, the subdifferential (4.4) turns out to be the upper limit (*robust regularization*) of the subdifferential mapping used in the theory of *viscosity solutions* [11]. It is well known that if f is locally Lipschitzian around $\bar{x}$ with modulus l_f, then

$$\partial f(\bar{x}) \neq \emptyset, \quad |x^\star| \leq l_f \ \ \forall x^\star \in \partial f(\bar{x}). \tag{4.5}$$

Moreover, in this case $\partial_C f(\bar{x}) = \text{co}\partial f(\bar{x})$ for the generalized gradient of Clarke. Note also that

$$\partial\delta(\bar{x}, \Omega) = N(\bar{x}; \Omega) \ \text{ if } \ \bar{x} \in \Omega \tag{4.6}$$

for the indicator function $\delta(\cdot, \Omega)$. We also use another representation of the normal cone (4.2) in terms of the subdifferential (4.4) subdifferential of the distance function (see [28, Prop. 2.7]):

$$N(\bar{x}; \Omega) = \text{cone}[\partial\text{dist}(\bar{x}, \Omega)] \ \text{ if } \ \ \bar{x} \in \text{cl}\Omega. \tag{4.7}$$

Among the most important advantages of constructions (4.2) and (4.4), one has a *rich calculus* under general assumptions. We refer to [10], [17]–[19], [23], [28]–[33], [40], [45] for various results in this direction. For applications in this paper, we need the two following basic rules:

$$\partial(f_1 + f_2)(\bar{x}) \subset \partial f_1(\bar{x}) + \partial f_2(\bar{x}) \tag{4.8}$$

if one of the functions f_i is Lipschitz continuous around $\bar{x}$, and

$$N(\bar{x}; \Omega_1 \cap \Omega_2) \subset N(\bar{x}; \Omega_1) + N(\bar{x}; \Omega_2) \ \text{ if } \ N(\bar{x}; \Omega_1) \cap (-N(\bar{x}; \Omega_2)) = \{0\} \tag{4.9}$$

for any closed sets Ω_1 and Ω_2. Note also the useful chain rule equality

$$\partial(\varphi \circ g)(\bar{x}) = \partial\langle\nabla\varphi(\bar{y}), g\rangle(\bar{x}) \tag{4.10}$$

where $(\varphi \circ g)(x) := \varphi(g(x))$ with $\varphi : \mathbf{R}^m \to \mathbf{R}$ strictly differentiable at $\bar{y} := g(\bar{x})$ and $g : \mathbf{R}^n \to \mathbf{R}^m$ Lipschitz continuous around $\bar{x}$.

DEFINITION 4.3. *Let $F : \mathbf{R}^n \rightrightarrows \mathbf{R}^m$ be a multifunction of nonempty graph* gph$F$, *and let $(\bar{x}, \bar{y}) \in \text{cl}(\text{gph}F)$. The multifunction $D^\star F(\bar{x}, \bar{y})$ from $\mathbf{R}^m$ into $\mathbf{R}^n$ defined by*

$$D^\star F(\bar{x}, \bar{y})(y^\star) := \{x^\star \in \mathbf{R}^n | (x^\star, -y^\star) \in N((\bar{x}, \bar{y}); \text{gph}F)\} \tag{4.11}$$

is called the coderivative *of F at $(\bar{x}, \bar{y})$. The symbol $D^\star F(\bar{x})$ is used when F is single-valued at $\bar{x}$ and $\bar{y} = F(\bar{x})$.*

Note that because it is nonconvex valued, the coderivative (4.11) is *not dual* to any tangentially generated derivatives of multifunctions (see, e.g., [3, Chap. 5]). Now we review some properties of the coderivative (4.11) that are significant for applications in this paper. First we consider the multifunction F of a special form whose graph is

$$\text{gph}F := \{(x, y) \in \mathbf{R}^n \times \mathbf{R}^m | x \in \Omega, \ g(x) - y \in \Delta\} \tag{4.12}$$

where $g : \mathbf{R}^n \to \mathbf{R}^m$. The following result is proved in [28, Thm. 3.3].

PROPOSITION 4.4. (i) *Let the set* (4.12) *be closed around the point* $(\bar{x}, \bar{y}) \in \text{gph}F$. *Then*

$$[D^\star F(\bar{x}, \bar{y})(y^\star) \neq \emptyset] \Longrightarrow y^\star \in N(g(\bar{x}) - \bar{y}; \Delta).$$

(ii) *If g is Lipschitz continuous around $\bar{x}$, then*

$$D^\star F(\bar{x},\bar{y})(y^\star) = \begin{cases} \partial_x s(\bar{x},y^\star) & \textit{for } y^\star \in N(g(\bar{x})-\bar{y};\Delta), \\ \emptyset & \textit{otherwise} \end{cases} \tag{4.13}$$

where $s(x,y^\star) := \langle y^\star, g(x)\rangle + \delta(x,\Omega)$.

Setting $\Omega = \mathbf{R}^n$ and $\Delta = \{0\}$, we get from (4.13) the *scalarization formula* obtained in [17, Prop. 8]. Moreover, one has the following relation with Clarke's generalized Jacobian:

$$\mathrm{co} D^\star g(\bar{x})(y^\star) = \mathrm{co}\partial\langle y^\star, g\rangle(\bar{x}) = (J_C g(\bar{x}))^\star y^\star \quad \forall y^\star \in \mathbf{R}^m. \tag{4.14}$$

It turns out that the coderivative (4.11) enjoys a rich calculus under natural (e.g., Lipschitzian) assumptions in general multivalued settings; see [33] for various results in this direction. Furthermore, the coderivative construction allows us to get *complete dual characterizations* of Lipschitzian properties of multifunctions that are of crucial importance for applications below (see Corollary 5.3 and the proofs of Theorems 6.1 and 7.1). Now we present results for the classical locally Lipschitz property that were proved in [31, Thm. 5.11]. We refer the reader to [2], [31], [32], [34], [41], [45] for studies and applications of more general Lipschitzian properties and related topics.

PROPOSITION 4.5. *Let F be of closed graph and bounded around $\bar{x}$ with $F(\bar{x}) \neq \emptyset$. Then each of the following conditions is necessary and sufficient for F to be locally Lipschitzian around this point:*

(i) *there exist a neighborhood U of $\bar{x}$ and a constant $l \geq 0$ such that*

$$\sup\{|x^\star| : x^\star \in D^\star F(x,y)(y^\star)\} \leq l|y^\star| \quad \forall x \in U,\ y \in F(x),\ y^\star \in \mathbf{R}^m; \tag{4.15}$$

(ii) $D^\star F(\bar{x},\bar{y})(0) = \{0\}$ $\forall \bar{y} \in F(\bar{x})$.

Remark 4.6. The estimate (4.15) is crucial to ensure the convergence of adjoint functions in the approximation procedures of §§6 and 7. If one replaces the normal cone (4.2) in the coderivative construction (4.11) by the Clarke normal cone (4.3), then such a counterpart $D_C^\star F$ of the coderivative (actually appearing in Clarke's version of the Euler–Lagrange inclusion) does not provide estimate (4.15) and the "null-condition" (ii) for Lipschitzian multifunctions in many important situations. This is related to the fact that Clarke's normal cone to any *Lipschitzian manifold* (which is a set locally representable as the graph of a Lipschitz continuous vector function) is a *linear subspace*; see Rockafellar [42]. It turns out that Lipschitzian manifolds include not only Lipschitz continuous functions but also graphs of maximal monotone operators, in particular, subdifferential mappings for convex and saddle functions. For such objects, estimate (4.15) in terms of $D_C^\star F$ and the corresponding "null-condition" are fulfilled in fact only for "strictly smooth" multifunctions. We refer to [42] and [34] for more information about these and related properties.

In conclusion of this section, we present a useful result for convex-valued multifunctions [28, Thm. 3.1], hence showing that the considered Euler–Lagrange conditions for differential and discrete inclusions automatically imply the *maximum* (*minimum*) *conditions* in problems with convex velocities.

PROPOSITION 4.7. *Let F be convex-valued around $\bar{x}$ and lower semicontinuous at $\bar{x}$. Then for any $\bar{y} \in F(\bar{x})$ one has*

$$[D^\star F(\bar{x},\bar{y})(y^\star) \neq \emptyset] \Longrightarrow [\langle y^\star, \bar{y}\rangle = \min\{\langle y^\star, y\rangle |\ y \in F(\bar{x})\}].$$

5. Necessary conditions for discrete approximations. In this section we obtain necessary optimality conditions in discrete approximation problems (P_k) for each $k = 1, 2, \ldots$. These conditions will be derived from a *generalized Lagrange multiplier rule* for finite dimensional problems in mathematical programming with many geometric constraints.

Let $\phi_j : \mathbf{R}^d \to \bar{\mathbf{R}}$ for $j = 0, \ldots, s$ and $g_j : \mathbf{R}^d \to \mathbf{R}^n$ for $j = 0, \ldots, m$. Consider the following problem (MP):

$$\text{minimize } \phi_0(z) \text{ for } z \in \mathbf{R}^d \text{ subject to} \tag{5.1}$$

$$\phi_j(z) \le 0 \text{ for } j = 1, \ldots, s, \tag{5.2}$$

$$g_j(z) = 0 \text{ for } j = 0, \ldots, m, \tag{5.3}$$

$$z \in \Delta_j \text{ for } j = 0, \ldots, l. \tag{5.4}$$

PROPOSITION 5.1. *Let $\bar{z}$ be an optimal solution to problem (MP). Assume that the functions ϕ_j are Lipschitz continuous, the functions g_j are smooth, and the sets Δ_j are closed around $\bar{z}$. Then there exist real numbers $\{\mu_j | j = 0, \ldots, s\}$ as well as vectors $\{\psi_j \in \mathbf{R}^n | j = 0, \ldots, m\}$ and $\{z_j^* \in \mathbf{R}^d | j = 0, \ldots, l\}$, not all zero, such that*

$$z_j^* \in N(\bar{z}; \Delta_j) \textit{ for } j = 0, \ldots, l, \tag{5.5}$$

$$\mu_j \ge 0 \textit{ for } j = 0, \ldots, s, \tag{5.6}$$

$$\mu_j \phi_j(\bar{z}) = 0 \textit{ for } j = 1, \ldots, s, \tag{5.7}$$

$$-z_0^* - \cdots - z_l^* \in \partial \left(\sum_{j=0}^{s} \mu_j \phi_j \right)(\bar{z}) + \sum_{j=0}^{m} (\nabla g_j(\bar{z}))^* \psi_j. \tag{5.8}$$

The proof of this result, which is based on the *metric approximation method*, can be found in Mordukhovich [27, Thm. 1] and [28, Cor. 7.5.1]. Note that this method facilitates obtaining necessary conditions for (MP) in more general forms in the presence of nonsmooth equality and inequality constraints without Lipschitzian assumptions; see [28, §7] and §7 below.

Now we employ Proposition 5.1 and calculus rules for the generalized differential constructions in §4 to prove necessary optimality conditions for finite difference problems (P_k) in the following *Euler–Lagrange form*. Considering problem (P_k) in (3.4), (3.15)–(3.18) for any fixed $k = 1, 2, \ldots$, we denote

$$F_j(\cdot) := F(\cdot, t_j) \text{ and } f_j(\cdot, \cdot) := f(\cdot, \cdot, t_j) \text{ as } j = 0, \ldots, k-1.$$

THEOREM 5.2. *Let $\bar{x}^k = (\bar{x}_0^k, \ldots, \bar{x}_k^k)$ be an optimal solution to problem (P_k). Assume that the sets Ω and* gphF_j *are closed and the functions φ and f_j are Lipschitz continuous around the points $(\bar{x}_0^k, \bar{x}_k^k)$ and $(\bar{x}_j^k, (\bar{x}_{j+1}^k - \bar{x}_j^k)/h_k)$ respectively, for all*

$j = 0, \dots, k-1$. *Then there exist a number* $\lambda^k \geq 0$ *and a vector* $p^k = (p_0^k, \dots, p_k^k) \in \mathbf{R}^{(k+1)n}$, *not both zero, such that*

$$(p_0^k + 2\lambda^k(\bar{x}(a) - \bar{x}_0^k), -p_k^k) \in \lambda^k \partial\varphi(\bar{x}_0^k, \bar{x}_k^k) + N((\bar{x}_0^k, \bar{x}_k^k); \Omega_k), \tag{5.9}$$

$$\begin{aligned} ((p_{j+1}^k - p_j^k)/h_k, p_{j+1}^k - \lambda^k \theta_j^k/h_k) \in \lambda^k \partial f_j(\bar{x}_j^k, (\bar{x}_{j+1}^k - \bar{x}_j^k)/h_k) \\ + N((\bar{x}_j^k, (\bar{x}_{j+1}^k - \bar{x}_j^k)/h_k); \mathrm{gph} F_j) \ \ \textit{for}\ j = 0, \dots, k-1 \end{aligned} \tag{5.10}$$

where

$$\theta_j^k := -2\int_{t_j}^{t_{j+1}} (\dot{\bar{x}}(t) - (\bar{x}_{j+1}^k - \bar{x}_j^k)/h_k)dt. \tag{5.11}$$

Proof. Let us introduce a new variable $z = (x_0, \dots, x_k, y_0, \dots, y_{k-1}) \in \mathbf{R}^{(2k+1)n}$ and consider the following problem of mathematical programming:

$$\begin{aligned} &\text{minimize } \phi_0(z) := \varphi(x_0, x_k) + |x_0 - \bar{x}(a)|^2 + h_k \sum_{j=0}^{k-1} f_j(x_j, y_j) \\ &\quad + \sum_{j=0}^{k-1} \int_{t_j}^{t_{j+1}} |y_j - \dot{\bar{x}}(t)|^2 dt \ \text{ subject to} \end{aligned} \tag{5.12}$$

$$\phi_j(z) := |x_{j-1} - \bar{x}(t_{j-1})| - \varepsilon/2 \leq 0 \ \text{ for } j = 1, \dots, k+1, \tag{5.13}$$

$$\phi_{k+2}(z) := \sum_{j=0}^{k-1} \int_{t_j}^{t_{j+1}} |y_j - \dot{\bar{x}}(t)|^2 dt - \varepsilon/2 \leq 0, \tag{5.14}$$

$$g_j(z) := x_{j+1} - x_j - h_k y_j = 0 \ \text{ for } j = 0, \dots, k-1, \tag{5.15}$$

$$z \in \Delta_j := \{(x_0, \dots, y_{k-1}) \in \mathbf{R}^{(2k+1)n} | y_j \in F_j(x_j)\} \ \text{ for } j = 0, \dots, k-1, \tag{5.16}$$

$$z \in \Delta_k := \{(x_0, \dots, y_{k-1}) \in \mathbf{R}^{(2k+1)n} | (x_0, x_k) \in \Omega_k\}. \tag{5.17}$$

It is easy to see that problem (5.12)–(5.17) as defined is equivalent to the discrete approximation problem (P_k) in (3.4), (3.15)–(3.18). On the other hand, (5.12)–(5.17) is a problem (MP) in (5.1)–(5.4) with $d = (2k+1)n$, $s = k+2$, $m = k-1$, $l = k$, and the specified functions ϕ_j, g_j and sets Δ_j. Now we employ Proposition 5.1 for the optimal solution $\bar{z} = \bar{z}^k := (\bar{x}_0^k, \dots, \bar{x}_k^k, (\bar{x}_1^k - \bar{x}_0^k)/h_k, \dots, (\bar{x}_k^k - \bar{x}_{k-1}^k)/h_k)$ of (5.12)–(5.17) where $\bar{x}^k = (\bar{x}_0^k, \dots, \bar{x}_k^k)$ is a given optimal solution of (P_k).

According to this result, one gets real numbers $(\mu_0, \dots, \mu_{k+2})$ as well as vectors $\psi_j \in \mathbf{R}^n$ $(j = 0, \dots, k-1)$ and $z_j^\star = (x_{0j}^\star, \dots, x_{kj}^\star, y_{0j}^\star, \dots, y_{k-1\,j}^\star) \in \mathbf{R}^{(2k+1)n}$ $(j = 0, \dots, k)$, not all zero, such that conditions (5.5)–(5.8) are fulfilled for the initial data in (5.12)–(5.17). Note that these μ_j, ψ_j, and $z_j^\star$ depend on k but we omit the index "k" for simplicity, considering k big enough.

First let us observe that thanks to Theorem 3.3, $\phi_j(\bar{z}^k) < 0$ as $j = 1, \ldots, k+2$ for all big k. This implies $\mu_j = 0$ for $j = 1, \ldots, k+2$ by virtue of the corresponding complementary slackness conditions in (5.7). Now denoting $\lambda^k := \mu_0 \geq 0$, we ensure that λ^k, $(\psi_0, \ldots, \psi_{k-1})$, and $(z_0^\star, \ldots, z_k^\star)$ are not equal to zero simultaneously for k big enough.

Further, it follows from the structure of the sets Δ_j in (5.16) and (5.17) that conditions (5.5) are equivalent to

$$(x_{jj}^\star, y_{jj}^\star) \in N((\bar{x}_j^k, (\bar{x}_{j+1}^k - \bar{x}_j^k)/h_k); \mathrm{gph} F_j) \text{ and } \tag{5.18}$$

$$x_{ij}^\star = y_{ij}^\star = 0 \text{ if } i \neq j \ \forall j = 0, \ldots, k-1;$$

$$(x_{0k}^\star, x_{kk}^\star) \in N((\bar{x}_0^k, \bar{x}_k^k); \Omega_k) \text{ and } x_{ik}^\star = y_{ik}^\star = 0 \text{ otherwise.} \tag{5.19}$$

Taking this into account and using calculus rule (4.8), we get from (5.7), (5.8), (5.12), and (5.15) the following relationships:

$$-x_{00}^\star - x_{0k}^\star = \lambda^k u_0 + 2\lambda^k(\bar{x}_0^k - \bar{x}(a)) + \lambda^k h_k \vartheta_0 - \psi_0, \tag{5.20}$$

$$-x_{jj}^\star = \lambda^k h_k \vartheta_j + \psi_{j-1} - \psi_j \text{ for } j = 1, \ldots, k-1, \tag{5.21}$$

$$-x_{kk}^\star = \lambda^k u_k + \psi_{k-1}, \tag{5.22}$$

$$-y_{jj}^\star = \lambda^k h_k w_j + \lambda^k \theta_j^k - h_k \psi_j \text{ for } j = 0, \ldots, k-1 \tag{5.23}$$

where θ_j^k is defined in (5.11),

$$(u_0, u_k) \in \partial\varphi(\bar{x}_0^k, \bar{x}_k^k), \tag{5.24}$$

and

$$(\vartheta_i, w_j) \in \partial f_j(\bar{x}_j^k, (\bar{x}_{j+1}^k - \bar{x}_j^k)/h_k) \text{ for } j = 0, \ldots, k-1. \tag{5.25}$$

Now denoting

$$p_0^k := x_{0k}^\star + \lambda^k u_0 + 2\lambda^k(\bar{x}_0^k - \bar{x}(a)) \text{ and } p_j := \psi_{j-1} \text{ for } j = 1, \ldots, k,$$

one can conclude that relationships (5.18)–(5.25) imply conditions (5.9) and (5.10) where λ^k and $(p_0^k, \ldots, p_k^k)$ are not equal to zero simultaneously. This ends the proof of the theorem. □

COROLLARY 5.3. *In addition to the assumptions of Theorem* 5.2, *let us suppose that for each* $j = 0, \ldots, k-1$, *the multifunction* F_j *is bounded and Lipschitz continuous around* $\bar{x}_j^k$. *The conditions* (5.9) *and* (5.10) *are fulfilled with* $(\lambda^k, p_k^k) \neq 0$, *i.e., one can set*

$$\lambda^k + |p_k^k| = 1 \ \forall k = 1, 2, \ldots. \tag{5.26}$$

Proof. If $\lambda^k = 0$, then (5.10) is represented as

$$(p_{j+1}^k - p_j^k)/h_k \in D^\star F_j((\bar{x}_j^k, (x_{j+1}^k - \bar{x}_j^k)/h_k)(-p_{j+1}^k) \text{ for } j = 0, \ldots, k-1 \tag{5.27}$$

in terms of the coderivative (4.11). By virtue of (5.27), $p_k^k = 0$ implies that $p_j^k = 0$ for all $j = 0, \ldots, k-1$, according to Proposition 4.5. This proves the corollary. □

6. Necessary conditions for the Bolza problem. Now we come back to the original Bolza problem (P) and prove necessary optimality conditions for an i.r.l.m. in the refined Euler–Lagrange form. To accomplish it, it remains to pass to the limit in the necessary conditions for discrete approximation problems (P_k) as $k \to \infty$ (§5), taking into account the $W^{1,2}$-convergence of discrete optimal solutions (§3) and some properties of the generalized differential constructions in §4.

Here we keep assumptions (H1)–(H3), but instead of the continuity hypotheses in (H4) and (H5) we assume the corresponding Lipschitz continuity. Namely:

(H4$'$) There exist numbers $\nu > 0$ and $l_f \geq 0$ such that $f(\cdot,\cdot,t)$ is locally Lipschitzian with modulus l_f around any point of the set $A_\nu(t)$ in (3.19).

(H5$'$) φ is Lipschitz continuous on $U \times U$ and Ω is closed around $(\bar{x}(a), \bar{x}(b))$.

In what follows, we denote by $\partial f = \partial f(\cdot,\cdot,t)$ the subdifferential (4.4) of the function $f(x,v,t)$ with respect to (x,v) under fixed t. Similarly, $N((\cdot,\cdot);\mathrm{gph}F(\cdot,t))$ means the normal cone (4.2) to the set $\mathrm{gph}F(\cdot,t)$ at a given point $(\cdot,\cdot)$ when t is fixed. Note that the normal cone to the graph of F is related to the generalized derivative (coderivative) of F according to (4.11).

One of the fundamental properties of the generalized differential constructions under consideration is their robustness (upper semicontinuity) with respect to variables of differentiation; see §4. This is of principal importance for the method of discrete approximations. In the limiting procedure below, we also need such a robustness of $\partial f(\cdot,\cdot,t)$ and $N((\cdot,\cdot);\mathrm{gph}F(\cdot,t))$ with respect to the parameter t. More precisely, we impose the following technical assumptions:

(H6) For a.e. $t \in [a,b]$ one has

$$\limsup_{\substack{(x',v')\to(\bar{x}(t),\dot{\bar{x}}(t))\\ t'\to t,\ t'<t}} \partial f(x,'v',t') = \partial f(\bar{x}(t),\dot{\bar{x}}(t),t).$$

(H7) For a.e. $t \in [a,b]$ one has

$$\limsup_{\substack{(x',v')\to(\bar{x}(t),\dot{\bar{x}}(t))\\ t'\to t,\ t'<t}} N((x',v');\mathrm{gph}F(\cdot,t')) = N((\bar{x}(t),\dot{\bar{x}}(t));\mathrm{gph}F(\cdot,t)).$$

Properties (H6) and (H7) are obviously fulfilled if f and F do not depend on t and also if $f = f_1(x,v) + f_2(t)$, $F = F_1(x) + F_2(t)$ with F_2 satisfying (H2). Actually, (H6) and (H7) mean that the continuity in t holds under the generalized differentiation of f and F with respect to the other variables. In particular, this takes place when f and F are represented as compositions of mappings separated in t and (x,v). Note also that for functions f smooth in (x,v), (H6) means the classical continuity of $\partial f/\partial x$ and $\partial f/\partial v$ at $(\bar{x}(t),\dot{\bar{x}}(t),t)$.

Now we prove the refined Euler–Lagrange conditions for the original Bolza problem (P).

THEOREM 6.1. *Let $\bar{x}(\cdot)$ be an i.r.l.m. for problem (P) under assumptions* (H1)–(H3), (H4$'$), (H5$'$), (H6), *and* (H7). *Then there exist a number $\lambda \geq 0$ and an absolutely continuous function $p : [a,b] \to \mathbf{R}^n$, not both zero, such that*

$$\dot{p}(t) \in \mathrm{co}\{u | (u,p(t)) \in \lambda\partial f(\bar{x}(t),\dot{\bar{x}}(t),t) + N((\bar{x}(t),\dot{\bar{x}}(t));\mathrm{gph}F(\cdot,t))\} \tag{6.1}$$

for a.e. $t \in [a,b]$ and

$$(p(a), -p(b)) \in \lambda\partial\varphi(\bar{x}(a),\bar{x}(b)) + N((\bar{x}(a),\bar{x}(b));\Omega). \tag{6.2}$$

Proof. Let us construct a sequence of discrete approximations (P_k) of problem (P), which approximates $\bar{x}(\cdot)$ in the sense of Theorem 3.3. Now employing Theorem 5.2 for optimal solutions $x^k = (x_0^k, \dots, x_k^k)$ to (P_k) as $k \to \infty$, we find sequences of numbers $\lambda^k \geq 0$ and vectors $p^k = (p_0^k, \dots, p_k^k)$ satisfying conditions (5.9), (5.10), and (5.26). One can always suppose that $\lambda^k \to \lambda \geq 0$ as $k \to \infty$.

In what follows we use the notation $\bar{x}^k(t)$ and $p^k(t)$ for piecewise linear extensions of the corresponding discrete functions on $[a,b]$ with their piecewise constant derivatives $\dot{\bar{x}}^k(t)$ and $\dot{p}^k(t)$. Let us consider a sequence of the functions

$$\theta^k(t) := \theta_j^k/h_k \text{ for } t \in [t_j, t_{j+1}), \ j = 0, \dots, k-1,$$

generated by (5.11). Theorem 3.3 implies that

$$\begin{aligned} \int_a^b |\theta^k(t)|dt &= \sum_{j=0}^{k-1} |\theta_j^k| \leq 2 \sum_{j=0}^{k-1} \int_{t_j}^{t_{j+1}} |\dot{\bar{x}}(t) - (\bar{x}_{j+1}^k - \bar{x}_j^k)/h_k|dt \\ &= 2 \int_a^b |\dot{\bar{x}}(t) - \dot{\bar{x}}^k(t)|dt := \nu_k \to 0 \text{ as } k \to \infty. \end{aligned} \tag{6.3}$$

Without loss of generality we can suppose that

$$\dot{\bar{x}}^k(t) \to \dot{\bar{x}}(t) \text{ and } \theta^k(t) \to 0 \text{ a.e. } t \in [a,b] \text{ as } k \to \infty. \tag{6.4}$$

Let us estimate the adjoint functions $p^k(\cdot)$ for big k. According to (5.10) and Definition 4.3 of the coderivative D^*F_j, there exist vectors $(\vartheta_j^k, w_j^k) \in \partial f_i(\bar{x}_j^k, (\bar{x}_{j+1}^k - \bar{x}_j^k)/h_k)$ such that

$$(p_{j+1}^k - p_j^k)/h_k - \lambda^k \vartheta_j^k \in D^*F_j(\bar{x}_j^k, (\bar{x}_{j+1}^k - \bar{x}_j^k)/h_k)(\lambda^k w_j^k + \lambda^k \theta_j^k/h_k - p_{j+1}^k) \tag{6.5}$$

for all $j = 0, \dots, k-1$. Now using (6.5), (3.2), and Proposition 4.5, one has

$$|(p_{j+1}^k - p_j^k)/h_k - \lambda^k \vartheta_j^k| \leq l_F |\lambda^k w_j^k + \lambda^k \theta_j^k/h_k - p_{j+1}^k| \text{ for } j = 0, \dots, k-1. \tag{6.6}$$

It follows from (H4′) and estimate (4.5) that

$$|\vartheta_j^k| \leq l_f \text{ and } |w_j^k| \leq l_f \text{ for } j = 0, \dots, k-1. \tag{6.7}$$

Using (5.26), (6.3), (6.6), and (6.7), we get

$$\begin{aligned} |p_j^k| &\leq (1 + h_k l_F)|p_{j+1}^k| + h_k l_f (1 + l_F) + l_F |\theta_j^k| \leq \cdots \\ &\leq \exp[l_F(b-a)] + l_f(b-a)(1+l_F) + l_F \nu_k \ \forall j = 0, \dots, k-1 \text{ as } k \to \infty. \end{aligned} \tag{6.8}$$

This means that the adjoint functions $p^k(t)$ are uniformly bounded in $[a,b]$. Employing (6.6) and (6.7) to estimate the derivatives $\dot{p}^k(t)$, one has

$$|\dot{p}^k(t)| = |(p_{j+1}^k - p_j^k)/h_k| \leq l_f + l_F(l_f + |\theta^k(t)| + |p_{j+1}^k|) \text{ for } t_j \leq t < t_{j+1}.$$

By virtue of (6.4) and (6.8) this implies that the sequence $\{\dot{p}^k(\cdot)\}$ is weakly compact in $L^1[a,b]$. Therefore, we can find an absolutely continuous function $p(\cdot)$ such that $p^k(\cdot) \to p(\cdot)$ uniformly in $[a,b]$ and $\dot{p}^k(\cdot) \to \dot{p}(\cdot)$ weakly in $L^1[a,b]$ for $k \to \infty$ (as usual we take all $k = 1, 2, \dots$).

Let us rewrite (5.10) as follows:

$$\begin{aligned}\dot{p}^k(t) \in \{u | (u, p^k(t_{j+1}) - \lambda^k \theta^k(t)) \in \lambda^k \partial f(\bar{x}^k(t_j), \dot{\bar{x}}^k(t), t_j) \\ + N((\bar{x}^k(t_j), \dot{\bar{x}}^k(t)); \mathrm{gph} F(\cdot, t_j))\} \text{ for } t \in [t_j, t_{j+1}),\ j = 0, \dots, k-1.\end{aligned} \tag{6.9}$$

According to the classical results, there is a sequence of convex combinations of $\dot{p}^k(t)$ that converges to $\dot{p}(t)$ for a.e. $t \in [a, b]$. Now passing to the limit in (6.9) as $k \to \infty$ and using (6.4) as well as hypotheses (H6) and (H7), we obtain the Euler–Lagrange inclusion (6.1).

Taking the limit in (5.26), one has the normalization condition

$$\lambda + |p(b)| = 1,$$

which implies that λ and $p(\cdot)$ are not equal to zero simultaneously. It follows from (6.8) that if $p(t_0) = 0$ at some point $t_0 \in [a, b]$, then $p(t) \equiv 0$ in $[a, b]$.

It remains to establish the transversality inclusion (6.2). First note that

$$\lambda^k \partial \varphi(\bar{x}_0^k, \bar{x}^k) \to \lambda \partial \varphi(\bar{x}(a), \bar{x}(b)) \text{ as } k \to \infty$$

due to robustness of the subdifferential (4.4). Then observe that the set Ω_k in (3.16) is represented as

$$\Omega_k = \{(x_0, x_k) \in \mathbf{R}^{2n} | \operatorname{dist}((x_0, x_k), \Omega) \le \eta_k\}. \tag{6.10}$$

Now passing to the limit in (5.9) as $k \to \infty$ and using (4.7), we obtain (6.2). This completes the proof of the theorem. □

Remark 6.2. The Euler–Lagrange inclusion (6.1) can be expressed in terms of the coderivative $D_x^* F$ of the multifunction $F(\cdot, t)$ under fixed t. Indeed, one has

$$\dot{p}(t) \in \mathrm{co}\left\{ \bigcup_{(\vartheta, w) \in \partial f(\bar{x}(t), \dot{\bar{x}}(t), t)} [\lambda \vartheta + D_x^* F(\bar{x}(t), \dot{\bar{x}}(t), t)(\lambda w - p(t))] \right\} \tag{6.11}$$

for a.e. $t \in [a, b]$, which is equavalent to (6.1) by virtue of (4.11). Inclusion (6.11) implies the following:

$$\dot{p}(t) \in \mathrm{co}[\lambda \partial_x f(\bar{x}(t), \dot{\bar{x}}(t), t) + D_x^* F(\bar{x}(t), \dot{\bar{x}}(t), t)(\lambda \partial_v f(\bar{x}(t), \dot{\bar{x}}(t), t) - p(t))]$$

in the case when $\partial f(\bar{x}(t), \dot{\bar{x}}(t), t) \subset \partial_x f(\bar{x}(t), \dot{\bar{x}}(t), t) \times \partial_v f(\bar{x}(t), \dot{\bar{x}}(t), t)$ for a.e. $t \in [a, b]$. In particular, it happens when either $f(x, v, t) = f_1(x, t) + f_2(v, t)$ or f possesses some regularity (e.g., f is smooth or convex in (x, v)).

Remark 6.3. Theorem 6.1 provides necessary optimality conditions for the Bolza problem (P) in the basic case where endpoint constraints are given in the abstract (geometric) form (1.3). Using calculus rules available for the nonconvex subdifferential constructions in §4, we can obtain refined transversality conditions for special representations of the set Ω in (1.3). In particular, for the case of inequality and equality type constraints

$$\varphi_i(x(a), x(b)) \le 0 \text{ if } i = 1, \dots, q, \tag{6.12}$$

$$\varphi_i(x(a), x(b)) = 0 \text{ if } i = q+1, \dots, q+r \tag{6.13}$$

defined by locally Lipschitz functions φ_i, one has

$$(6.14)\qquad (p(a),-p(b))\in\partial\left(\sum_{i=0}^{q+r}\lambda_i\varphi_i\right)(\bar{x}(a),\bar{x}(b)),$$

$$(6.15)\qquad \lambda_i\geq 0 \text{ for } i=0,1,\ldots,q,$$

$$(6.16)\qquad \lambda_i\varphi_i(\bar{x}(a),\bar{x}(b))=0 \text{ for } i=1,\ldots,q$$

where $\varphi_0:=\varphi$ and $(\lambda_0,\ldots,\lambda_{q+r},p(\cdot))\neq 0$; cf. [27]–[29]. Due to the subdifferential sum rule, the transversality inclusion (6.14) implies the following "separated" one:

$$(p(a),-p(b))\in\sum_{i=0}^{q}\lambda_i\partial\varphi_i(\bar{x}(a),\bar{x}(b))+\sum_{i=q+1}^{q+r}\lambda_i\partial^0\varphi_i(\bar{x}(a),\bar{x}(b))$$

where $\partial^0\varphi(\bar{x}):=\partial\varphi(\bar{x})\cup[-\partial(-\varphi)(\bar{x})]$ is a symmetric subdifferential construction. In the next section we prove unified transversality conditions in the presence of both functional and geometric constraints for the case of Lipschitz as well as non-Lipschitz functions φ_i.

Remark 6.4. Developing the discrete approximation approach, we can extend necessary optimality conditions in Theorem 6.1 to the case of "Bolza constraints"

$$\varphi_i(x(a),x(b))+\int_a^b f_i(x(t),\dot{x}(t),t)dt\leq 0 \text{ for } i=1,\ldots,q,$$

$$\varphi_i(x(a),x(b))+\int_a^b f_i(x(t),\dot{x}(t),t)dt=0 \text{ for } i=q+1,\ldots,q+r,$$

which unify endpoint constraints (6.12) and (6.13) with isoperimetric type constraints in variational problems. In this case, the set of necessary conditions consists of (6.14), (6.15) and

$$\lambda_i\left[\varphi_i(\bar{x}(a),\bar{x}(b))-\int_a^b f_i(\bar{x}(t),\dot{\bar{x}}(t),t)dt\right]=0 \text{ for } i=1,\ldots,q,$$

$$\dot{p}(t)\in\operatorname{co}\left\{u\,\middle|\,(u,p(t))\in\partial\left(\sum_{i=0}^{q+r}\lambda_i f_i\right)(\bar{x}(t),\dot{\bar{x}}(t),t)+N((\bar{x}(t),\dot{\bar{x}}(t));\operatorname{gph}F)\right\}$$

for a.e. $t\in[a,b]$ where $f_0:=f$ and $(\lambda_0,\ldots,\lambda_{q+r},p(\cdot))\neq 0$.

Remark 6.5. The approach and results obtained can be extended to the Bolza problem for nonconvex differential inclusions with *free time* where a varying time interval is involved in constraints and optimization. We refer the reader to [35] for more details and discussions.

The necessary optimality conditions in Theorem 6.1 are proved for any i.r.l.m. in the Bolza problem (P). In particular, they hold for any feasible solution to (P), which turns out to be a strong local minimum for the relaxed problem. As we pointed out in §2, an optimal solution of (P) automatically solves the relaxed problem as well in some common settings. Let us present such a corollary of Theorem 6.1, which is used in the next section to obtain the refined Euler–Lagrange conditions without any relaxation.

COROLLARY 6.6. *Let the arc $\bar{x}(\cdot)$ provide a strong local minimum for the Bolza problem* (1.1) *and* (1.3) *obtained by ignoring the differential inclusion* (1.2). *Suppose that for some number $\mu > 0$ and open set $U \subset \mathbf{R}^n$ one has*:

$$\bar{x}(t) \in U \;\; \forall t \in [a,b] \;\; \text{and} \;\; |\dot{\bar{x}}(t)| < \mu \;\; \text{a.e.} \;\; t \in [a,b]; \tag{6.17}$$

$f(x,v,\cdot)$ is bounded and continuous for a.e. $t \in [a,b]$ uniformly in $(x,v) \in U \times (\mu B)$; $f(\cdot,\cdot,t)$ is Lipschitz continuous on $U \times (\mu B)$ uniformly in $t \in [a,b]$; and (H5′), (H6) *hold. Then there exists an absolutely continuous function $p : [a,b] \longrightarrow \mathbf{R}^n$ such that one has* (6.2) *with $\lambda = 1$ and*

$$\dot{p}(t) \in \operatorname{co}\{u | (u,p(t)) \in \partial f(\bar{x}(t),\dot{\bar{x}}(t),t)\} \;\; \text{a.e.} \;\; t \in [a,b]. \tag{6.18}$$

Proof. The boundedness of $\dot{\bar{x}}(t)$ in (6.17) means that $\bar{x}(\cdot) \in W^{1,\infty}[a,b]$. According to Proposition 2.4, $\bar{x}(\cdot)$ is a strong minimum for the relaxed problem (1.3), (2.4). Let us consider the (trivial) differential inclusion

$$\dot{x} \in F(x,t) := \mu B \;\; \text{a.e.} \;\; t \in [a,b]. \tag{6.19}$$

It is obvious that $\bar{x}(\cdot)$ is an i.r.l.m. for problem (P) in (1.1), (1.3), and (6.19) where all the assumptions of Theorem 6.1 are fulfilled. Moreover, $(\bar{x}(t),\dot{\bar{x}}(t)) \in (\operatorname{int} \operatorname{gph} F)$ for a.e. $t \in [a,b]$. In this case, (6.1) is equivalent to (6.18). □

Remark 6.7. Using the modified discrete approximation of the Bolza functional in Remark 3.4 and developing the procedure in the proof of Theorem 6.1, one can extend the results obtained for the case of integrands f measurable in t.

7. The Euler–Lagrange inclusion without relaxation.

In the concluding section of the paper we consider the following Mayer problem (P_{M}) for differential inclusions:

$$\text{minimize} \;\; J[x] := \varphi_0(x(a),x(b))$$

over absolutely continuous trajectories of (1.2) under geometric, inequality, and equality type endpoint constraints (1.3), (6.12), and (6.13). Obviously, problem (P_{M}) is a special case of the Bolza problem (P) with $f = 0$. On the other hand, the Bolza problem can be reduced to the Mayer form involving *unbounded* differential inclusions; see, for example, [8], [24]. Here we consider problem (P_{M}) under the boundedness assumption which is used in our technique.

The main objective of this section is to prove the refined Euler–Lagrange conditions for the nonconvex Mayer problem (P_{M}) *without any relaxability assumption.* To accomplish it, we employ the results in the previous section (namely, Corollary 6.6) and an additional approximation procedure combining ideas in [7], [22], [28]. The latter procedure allows us to approximate the Mayer problem under consideration by a sequence of nonsmooth Bolza problems without differential inclusions and endpoint constraints. In this way, we prove the refined Euler–Lagrange conditions for any *strong* local minimum in (P_{M}) and also for any *boundary* trajectory of a nonconvex differential inclusion. Moreover, these conditions are justified for a *weak* local minimum in (P_{M}) under an additional Riemann integrability assumption that simplifies the technique employed.

Keeping here assumptions (H1), (H2), and (H6) on F around the given trajectory for (1.2), we relax the Lipschitz continuity of φ_i in (H5) and Remark 6.3. Namely, we assume the following.

($\overline{\mathrm{H}5}$) The functions φ_i are lower semicontinuous for $i = 0, \dots, q$ and continuous for $i = q+1, \dots, q+r$; the set Ω is closed around $(\bar{x}(a), \bar{x}(b))$.

Consider the closed set

$$\begin{aligned}(7.1)\qquad \mathcal{E}_\Omega := \{(x_0, x_1, \nu_0, \dots, \nu_{q+r}) \in \mathbf{R}^{2n+q+r+1} | (x_0, x_1) \in \Omega,\\ \varphi_i(x_0, x_1) \le \nu_i \ \text{ for } \ i = 0, \dots, q \ \text{ and}\\ \varphi_i(x_0, x_1) = \nu_i \ \text{ for } \ i = q+1, \dots, q+r\}\end{aligned}$$

and the *essential endpoint Lagrangian*

$$(7.2)\qquad L_\Omega(x_0, x_1, \lambda_0, \dots, \lambda_{q+r}) := \sum_{i=0}^{q+r} \lambda_i \varphi_i(x_0, x_1) + \delta((x_0, x_1), \Omega).$$

THEOREM 7.1. *Let $\bar{x}(\cdot)$ be a strong minimum for the Mayer problem (P_{M}) under assumptions* (H1), (H2), ($\overline{\mathrm{H}5}$), *and* (H7). *Then there exist a vector $y^\star := (\lambda_0, \dots, \lambda_{q+r}) \in \mathbf{R}^{q+r+1}$ and an absolutely continuous function $p : [a, b] \to \mathbf{R}^n$, not both zero, such that*

$$(7.3)\qquad \dot{p}(t) \in \operatorname{co} D_x^\star F(\bar{x}(t), \dot{\bar{x}}(t), t)(-p(t)) \ \text{ a.e. } \ t \in [a, b],$$

$$(7.4)\qquad (p(a), -p(b), -y^\star) \in N((\bar{x}(a), \bar{x}(b), \bar{c}); \mathcal{E}_\Omega)$$

where $\bar{c} := (\varphi_0(\bar{x}(a), \bar{x}(b)), 0 \dots, 0) \in \mathbf{R}^{q+r+1}$. Condition (7.4) *always implies* (6.15) *and* (6.16). *Moreover,* (7.4) *is equivalent to the simultaneous fulfilment of* (6.15), (6.16), *and*

$$(7.5)\qquad (p(a), -p(b)) \in \partial L_\Omega(\cdot, \lambda_0, \dots, \lambda_{q+r})(\bar{x}(a), \bar{x}(b))$$

if all φ_i are Lipschitz continuous around $(\bar{x}(a), \bar{x}(b))$.

Proof. Let $\bar{\gamma} := \varphi_0(\bar{x}(a), \bar{x}(b))$. According to the metric approximation method in Mordukhovich [26–28], we consider the parametric functional

$$(7.6)\ I_\gamma[x] := \operatorname{dist}((x(a), x(b), c), \mathcal{E}_\Omega) \ \text{ with } \ c := (\gamma, 0, \dots, 0) \in \mathbf{R}^{q+r+1}, \quad \gamma \in \mathbf{R},$$

on trajectories for the differential inclusion (1.2). Let $U \subset \mathbf{R}^n$ be a bounded neighborhood of the strong minimum $\bar{x}(\cdot)$ where assumptions (H1), (H2), and ($\overline{\mathrm{H}5}$) are fulfilled. For every $\varepsilon > 0$, one has

$$I_\gamma[\bar{x}] \le |\gamma - \bar{\gamma}| < \varepsilon$$

if γ is close to $\bar{\gamma}$. On the other hand,

$$I_\gamma[x] > 0 \ \text{ for any } \ \gamma < \bar{\gamma}$$

whenever the trajectory $x(\cdot)$ for (1.2) belongs to the neighborhood U of the strong minimum $\bar{x}(\cdot)$.

Following Clarke [7], let us consider the set X of all trajectories $x(\cdot)$ for (1.2) satisfying $x(t) \in \mathrm{cl}U$ in $[a,b]$ and define a metric in X as follows:

$$d(x,y) := |x(a) - y(a)| + \int_a^b |\dot{x}(t) - \dot{y}(t)|dt. \tag{7.7}$$

One can easily see that X is a complete metric space with metric (7.7) and functional (7.6) is continuous in X for any γ. According to the constructions above, for every $\varepsilon > 0$ we find $\gamma_\varepsilon < \bar{\gamma}$ such that $\gamma_\varepsilon \to \bar{\gamma}$ as $\varepsilon \to 0$, $I_\varepsilon[\bar{x}] < \varepsilon$, and

$$I_\varepsilon[x] > 0 \text{ for } I_\varepsilon := I_{\gamma_\varepsilon} \tag{7.8}$$

where $x(\cdot)$ is any trajectory for (1.2) with $x(t) \in U$ in $[a,b]$. Now one can apply the Ekeland variational principle [14] and claim the existence of $x_\varepsilon(\cdot) \in X$ such that

$$d(x_\varepsilon, \bar{x}) \leq \sqrt{\varepsilon} \text{ and} \tag{7.9}$$

$$I_\varepsilon[x] + \sqrt{\varepsilon} d(x, x_\varepsilon) \geq I_\varepsilon[x_\varepsilon] \;\; \forall x(\cdot) \in X.$$

Note that (7.9) implies $x_\varepsilon(t) \in U$ for ε small enough, so $I_\varepsilon[x_\varepsilon] > 0$ by virtue of (7.8).

Now for any positive numbers M, ε and the Lipschitz constant l_F in (3.2), we define the functional

$$J_\varepsilon^M[x] := I_\varepsilon[x] + \sqrt{\varepsilon} d(x, x_\varepsilon) + M(1 + l_F^2)^{1/2} \int_a^b \mathrm{dist}((x(t), \dot{x}(t)), \mathrm{gph}F(\cdot, t))dt \tag{7.10}$$

on the set of all arcs $x(\cdot)$ (not necessarily trajectories for (1.2)) satisfying $x(t) \in U$ in $[a,b]$. We omit the proof of the following lemma, which can be furnished by the arguments in Kaskosz and Lojasiewicz [22, Lemmas 1 and 2].

LEMMA 7.2. *There exists a number $M \geq 1$ such that for every $\varepsilon \in (0, 1/M)$ the arc $x_\varepsilon(\cdot)$ provides an unconditional strong local minimum for the functional* (7.10).

Let us continue the proof of Theorem 7.1. Setting $c_\varepsilon := (\gamma_\varepsilon, 0, \ldots, 0)$, we consider any element $(z_{0\varepsilon}, z_{1\varepsilon}, e_\varepsilon) \in \Pi((x_\varepsilon(a), x_\varepsilon(b), c_\varepsilon), \mathcal{E}_\Omega)$ from the Euclidean projector (4.1) of $(x_\varepsilon(a), x_\varepsilon(b), c_\varepsilon)$ on the set $\mathcal{E}_\Omega$. Using this projection and Lemma 7.2, one can conclude that $x_\varepsilon(\cdot)$ provides a strong local minimun for the unconstrained Bolza problem

$$\text{minimize } \varphi_\varepsilon(x(a), x(b)) + \int_a^b f_\varepsilon(x(t), \dot{x}(t), t)dt \tag{7.11}$$

with the endpoint function

$$\varphi_\varepsilon(x_0, x_1) := [|x_0 - z_{0\varepsilon}|^2 + |x_1 - z_{1\varepsilon}|^2 + |c_\varepsilon - e_\varepsilon|^2]^{1/2} + \sqrt{\varepsilon}|x_0 - x_\varepsilon(a)| \tag{7.12}$$

and the integrand

$$f_\varepsilon(x, v, t) := M(1 + l_F^2)^{1/2} \mathrm{dist}((x, v), \mathrm{gph}F(\cdot, t)) + \sqrt{\varepsilon}|v - \dot{x}_\varepsilon(t)|. \tag{7.13}$$

Now we employ Corollary 6.6 in the unconstrained Bolza problem (7.11)–(7.13) taking into account the modified discrete approximation in Remark 3.4 for the last (measurable) term in (7.13); cf. the proof of Theorem 6.1. One can easily check that the assumptions in Theorem 7.1 ensure the fulfilment of the assumptions in

Corollary 6.6 around the solution $x_\varepsilon(\cdot)$ and the first term in (7.12) is smooth around $(x_\varepsilon(a), x_\varepsilon(b))$ by virtue of (7.8).

For each $\varepsilon > 0$, using the result in Corollary 6.6 and also calculus formulas (4.7) and (4.8) for functions (7.12) and (7.13), we find an arc $p_\varepsilon(\cdot)$ such that

$$\dot{p}_\varepsilon(t) \in \mathrm{co}\{u|(u, p(t)) \in N((x_\varepsilon(t), \dot{x}_\varepsilon(t)); \mathrm{gph}F(\cdot, t)) + \sqrt{\varepsilon}(0, B)\} \tag{7.14}$$

for a.e. $t \in [a, b]$ and

$$p_\varepsilon(a) = (x_\varepsilon(a) - z_{0\varepsilon})/\alpha_\varepsilon + \sqrt{\varepsilon}, \quad -p_\varepsilon(b) = (x_\varepsilon(b) - z_{1\varepsilon})/\alpha_\varepsilon \tag{7.15}$$

where $\alpha_\varepsilon := [|x_\varepsilon(a) - z_{0\varepsilon}|^2 + |x_\varepsilon(b) - z_{1\varepsilon}|^2 + |c_\varepsilon - e_\varepsilon|^2]^{1/2} > 0$.

Denoting $y_\varepsilon^\star := (e_\varepsilon - c_\varepsilon)/\alpha\varepsilon$, one has

$$|p_\varepsilon(a)|^2 + |p_\varepsilon(b)|^2 + |y_\varepsilon^\star|^2 = 1 \tag{7.16}$$

and

$$(p_\varepsilon(a), -p_\varepsilon(b), -y_\varepsilon^\star) \in \mathrm{cone}[(x_\varepsilon(a), x_\varepsilon(b), c_\varepsilon) - \Pi((x_\varepsilon(a), x_\varepsilon(b), c_\varepsilon), \mathcal{E}_\Omega)]. \tag{7.17}$$

Now let us consider the limiting procedure in (7.14)–(7.17) as $\varepsilon \to 0$. By virtue of (7.7), relationship (7.9) means that $x_\varepsilon(\cdot) \to \bar{x}(\cdot)$ in $W^{1,\infty}[a, b]$ as $\varepsilon \to 0$. This implies that $x_\varepsilon(t) \to \bar{x}(t)$ uniformly in $[a, b]$ and $\dot{x}_\varepsilon(t) \to \dot{\bar{x}}(t)$ for a.e. $t \in [a, b]$.

Further, using (7.14), (7.16), (4.11), and Proposition 4.5, one can conclude (cf. the proof of Theorem 6.1) that there exist an arc $p(\cdot)$ and a vector $y^\star \in \mathbf{R}^{q+r+1}$, not both zero, such that $y_\varepsilon^\star$ converges to $y^\star$, $p_\varepsilon(t)$ converges to $p(t)$ uniformly in $[a, b]$, and a convex combination of $\dot{p}_\varepsilon(t)$ converges to $\dot{p}(t)$ for a.e. $t \in [a, b]$ as $\varepsilon \to 0$ along some subsequence.

Now passing to the limit in (7.14), (7.16), (7.17) and using Definitions 4.1 and 4.3 as well as robustness of the normal cone (4.2), one gets the main conclusions (7.3) and (7.4) of the theorem. Representing $y^\star = (\lambda_0, \ldots, \lambda_{q+r})$, we obtain (6.15) and (6.16) directly from Proposition 4.4(i) where

$$\mathrm{gph}F = \mathcal{E}_\Omega, \quad g = (\varphi_0, \ldots, \varphi_{q+r}), \quad \bar{x} = (\bar{x}(a), \bar{x}(b)), \quad \bar{y} = (0, \ldots, 0),$$

$$\Delta = \{(\mu_0, \ldots, \mu_{q+r})|\mu_i \leq 0 \text{ for } i = 0, \ldots, q \text{ and } \mu_i = 0 \text{ for } i = q+1, \ldots, q+r\}.$$

If all φ_i are locally Lipschitzian around $(\bar{x}(a), \bar{x}(b)$, then the equivalence of (7.4) to the simultaneous fulfilment of (6.15), (6.16), and (7.5) follows from Proposition 4.4(ii) where the scalarization function $s(\bar{x}, y^\star)$ is reduced to the essential endpoint Lagrangian (7.2) in the case under consideration. This ends the proof of the theorem. □

COROLLARY 7.3. *Let all φ_i be Lipschitz continuous around $(\bar{x}(a), \bar{x}(b))$ in the framework of Theorem 7.1. Then, in addition to (7.3), one has the transversality inclusion*

$$(p(a), -p(b)) \in \partial\left(\sum_{i=0}^{q+r} \lambda_i \varphi_i\right)(\bar{x}(a), \bar{x}(b)) + N((\bar{x}(a), \bar{x}(b)); \Omega) \tag{7.18}$$

where λ_i satisfies (6.15), (6.16) and $(\lambda_0, \ldots, \lambda_{q+r}, p(\cdot)) \neq 0$.

Proof. One can derive (7.18) directly from the sum rule (4.8) and representation (4.6). Note that (7.18) implies transversality conditions of the corresponding "separated" form in Remark 6.3. □

COROLLARY 7.4. *Let $\bar{x}(\cdot)$ be a weak local minimizer for problem (P_{M}). If, in addition to the assumptions of Theorem 7.1, $\dot{\bar{x}}(\cdot)$ is Riemann intregrable on $[a, b]$, then all the conclusions of the theorem hold true.*

Proof. By definition of the weak local minimum for (P_{M}), there exists $\varepsilon > 0$ such that $\bar{x}(\cdot)$ provides the strong (actually global) minimum for the auxiliary problem $(P_{\mathrm{M}}^{\varepsilon})$ of the same structure involving the differential inclusion

$$\dot{x} \in F_{\varepsilon}(x, t) := F(x, t) \cap G_{\varepsilon}(x, t) \tag{7.19}$$

where $G_{\varepsilon}(x, t) := \{v \in \mathbf{R}^n |$ such that $|v - \dot{\bar{x}}(t)| \leq \varepsilon\}$. Due to a.e. continuity of $\dot{\bar{x}}(\cdot)$ on $[a, b]$, the differential inclusion (7.19) satisfies all the assumptions in Theorem 7.1.

Now we can employ the necessary conditions of Theorem 7.1 in problem $(P_{\mathrm{M}}^{\varepsilon})$. Taking into account the structure of G_{ε} in (7.19) and using the intersection rule (4.9) for representing $D_x^{\star} F_{\varepsilon}$, we arrive at the required Euler–Lagrange inclusion (7.3) for the weak local minimizer $\bar{x}(\cdot)$ to the original Mayer problem (P_{M}). □

Remark 7.5. One can see that the Riemann integrability (a.e. continuity) assumption on $\dot{\bar{x}}(\cdot)$ in Corollary 7.4 is required to satisfy the Hausdorff continuity assumption (H2) on $F_{\varepsilon}(x, \cdot)$ for a.e. $t \in [a, b]$. This is essential in the proof of general results in Theorem 6.1 based on discrete approximations. However, for applications to the Mayer problem in the proof of Theorem 7.1, we need to use only the result of Corollary 6.6 that was obtained for the classical (but nonsmooth) Bolza problem without differential inclusions. The latter result could be extended to the case of integrands f measurable in t; see Remark 6.7. In this way, one can obtain the Euler–Lagrange conditions in Theorem 7.1 for Mayer problems involving differential inclusions with the *measurable* t-dependence corresponding to the measurability of the distance function integrands (7.13) in the approximation problems. Now using the procedure in Corollary 7.4, we could justify the refined Euler–Lagrange inclusion (7.3) for *any* (Lipschitz continuous) *weak minimizer* to the nonconvex Mayer problem (P_{M}) under consideration.

Remark 7.6. If F is convex-valued, then the Euler–Lagrange inclusion (7.3) automatically implies the Weierstrass–Pontryagin maximum condition

$$\langle p(t), \dot{\bar{x}}(t)\rangle = \max\{\langle p(t), v\rangle |\; v \in F(\bar{x}(t), t)\} \text{ a.e. } t \in [a, b] \tag{7.20}$$

due to Proposition 4.7. It is no longer true in the nonconvex setting for arbitrary weak or intermediate local minimizers. But we believe that for the case of *strong minimum*, the refined Euler–Lagrange conditions in Theorem 7.1 are fulfilled simultaneously with the maximum condition (7.20) for nonconvex differential inclusions. This statement can be obtained directly from the proof of Theorem 7.1 if one establishes the classical Weierstrass necessary condition for the strong local minimum to the simplest variational problem like (7.11)–(7.13) with no smoothness and/or convexity assumptions.

Now let us consider an analogue of the results obtained for the case of *boundary trajectories*. Given a nonempty closed set $A \subset \mathbf{R}^n$, we denote by $R(A)$ the *reachable set* for (1.2) from A, i.e., the set of all points $x(b)$ where $x(\cdot)$ is a trajectory for (1.2) with $x(a) \in A$.

THEOREM 7.7. *Let $\bar{x}(\cdot)$ be a trajectory for* (1.2) *with $\bar{x}(a) \in A$, and let assumptions* (H1), (H2), (H7) *hold. If $g : \mathbf{R}^n \to \mathbf{R}^m$ is a locally Lipschitzian function* *around $\bar{x}(b)$ and if $g(\bar{x}(b))$ is a boundary point of the set $R(A)$, then there exist an arc $p : [a,b] \to \mathbf{R}^n$ and a unit vector $\psi \in \mathbf{R}^m$ such that $p(\cdot)$ satisfies the refined Euler–Lagrange inclusion* (7.3) *with the following boundary* (*transversality*) *conditions:*

$$p(a) \in N(\bar{x}(a); A), \quad -p(b) \in \partial\langle \psi, g\rangle(\bar{x}(b)). \tag{7.21}$$

Remark 7.8. The result formulated generalizes the recent one in Kaskosz and Lojasiewicz [22] where (7.3) is replaced by Clarke's form of the Euler–Lagrange inclusion (see (1.4) in the case where $f = 0$) and conditions (7.21) are replaced by

$$p(a) \in N_C(\bar{x}(a); A), \quad p(b) \in [J_C g(\bar{x}(b))]^* \psi$$

in terms of Clarke's normal cone and generalized Jacobian; cf. (4.3) and (4.14).

Proof of Theorem 7.7. Following the arguments in [22], for every $\varepsilon > 0$ we can find a vector $c_\varepsilon \in \mathbf{R}^m$ and a trajectory $x_\varepsilon(\cdot)$ of (1.2) with $x_\varepsilon(a) \in A$ such that $|g(x_\varepsilon(b)) - c_\varepsilon| > 0$,

$$c_\varepsilon \to g(\bar{x}(b)), \quad x_\varepsilon(\cdot) \to \bar{x}(\cdot) \ \text{ in } \ W^{1,\infty}[a,b] \ \text{ as } \ \varepsilon \to 0,$$

and $x_\varepsilon(\cdot)$ is an unconditional strong local minimizer for problem (7.11) with integrand (7.13) and the endpoint functional

$$\varphi_\varepsilon(x_0, x_1) := |g(x_1) - c_\varepsilon| + \sqrt{\varepsilon}|x_0 - x_\varepsilon(a)| + M\text{dist}(x_0, A). \tag{7.22}$$

Now employing Corollary 6.6 in problem (7.11), (7.13), (7.22) and using calculus rules (4.7), (4.8), and (4.10) to compute the subdifferential of (7.22), one gets an arc $p_\varepsilon(\cdot)$ and a unit vector $\psi_\varepsilon \in \mathbf{R}^m$ such that (7.14) holds and

$$p_\varepsilon(a) \in \sqrt{\varepsilon}B + N(x_\varepsilon(a); A), \quad -p_\varepsilon(b) \in \partial\langle \psi_\varepsilon, g\rangle(x_\varepsilon(b)). \tag{7.23}$$

Following the proof of Theorem 7.1, we obtain conditions (7.3) and (7.21) by passing to the limit in (7.14) and (7.23) as $\varepsilon \to 0$. □

Remark 7.9. Using the method of metric approximations (as in the proof of Theorem 7.1), one can extend Theorem 7.7 to a more general setting when $g(\bar{x}(b))$ is a *locally extremal point* of the set $R(A)$ relative to other given sets. We refer to [28, §6] and [33, §3] for more details about this concept.

Remark 7.10. Following the proof of Theorem 7.7, we can obtain, in addition to (7.3) and (7.21), the maximum condition (7.20) for boundary trajectories of nonconvex differential inclusions if one has the Weirestrass necessary condition for the strong mimimum to the simplest variational problem (7.11) with nonsmoothness (7.13) and (7.22).

Acknowledgments. The author is indebted to Asen Dontchev, Alexander Ioffe, Richard Vinter, and two anonymous referees for their helpful remarks.

REFERENCES

[1] V. M. Alekseev, V. M. Tikhomirov, and S. V. Fomin, *Optimal Control*, Nauka, Moscow, 1979 (in Russian); English transl. in Consultants Bureau, New York, 1987.

[2] J.-P. Aubin, *Lipschitz behavior of solutions to convex minimization problems*, Math. Oper. Res., 9 (1984), pp. 87–111.

[3] J.-P. Aubin and H. Frankowska, *Set-Valued Analysis*, Birkhäuser, Boston, 1990.

[4] N. N. Bogoljubov, *Sur quelques methodes nouvelles dans le calculus des variations*, Ann. Math. Pura Appl., 7 (1930), pp. 249–271.

[5] F. H. CLARKE, *Optimal solutions to differential inclusions*, J. Optimization Theory Appl., 19 (1976), pp. 469–478.

[6] ———, *The generalized problem of Bolza*, SIAM J. Control Optim., 14 (1976), pp. 682–699.

[7] ———, *Necessary conditions for a general control problem*, in Calculus of Variations and Control Theory, D. Russel, ed., Academic Press, New York, 1976, pp. 257–278.

[8] ———, *Optimization and Nonsmooth Analysis*, Wiley-Interscience, New York, 1983.

[9] ———, *Hamiltonian analysis of the generalized problem of Bolza*, Trans. Amer. Math. Soc., 301 (1987), pp. 385–400.

[10] ———, *Methods of Dynamic and Nonsmooth Optimization*, Society for Industrial and Applied Mathematics, Philadelphia, PA, 1989.

[11] M. G. CRANDALL AND P.-L. LIONS, *Viscosity solutions of Hamilton-Jacobi equations*, Trans. Amer. Math. Soc., 277 (1983), pp. 1–42.

[12] A. L. DONTCHEV AND E. M. FARKHI, *Error estimates for discretized differential inclusions*, Computing, 41 (1989), pp. 349–358.

[13] A. L. DONTCHEV AND F. LEMPIO, *Difference methods for differential inclusions: a survey*, SIAM Rev., 34 (1992), pp. 263–294.

[14] I. EKELAND, *On the variational principle*, J. Math. Anal. Appl. 47 (1974), 324–353.

[15] I. EKELAND AND R. TEMAN, *Convex Analysis and Variational Problems*, North-Holland, Amsterdam, 1976.

[16] H. FRANKOWSKA, *The maximum principle for an optimal solution to a differential inclusion with endpoint constraints*, SIAM J. Control Optim., 25 (1987), pp. 145–157.

[17] A. D. IOFFE, *Approximate subdifferentials and applications*, I: *the finite dimensional theory*, Trans. Amer. Math. Soc., 281 (1984), pp. 389–416.

[18] ———, *Approximate subdifferentials and applications*, III: *the metric theory*, Mathematika, 36 (1989), pp. 1–38.

[19] ———, *Nonsmooth subdifferentials: their calculus and applications*, in the Proceedings of the 1st World Congress of Nonlinear Analysts, W. de Gruyter, Berlin, 1995.

[20] A. D. IOFFE AND V. M. TIKHOMIROV, *Extensions of variational problems*, Trans. Moscow Math. Soc., 18 (1968), pp. 207–273. (in Russian.)

[21] ———, *Theory of Extremal Problems*, Nauka, Moscow, 1974 (in Russian); English transl., North-Holland, Amsterdam, 1979.

[22] B. KASKOSZ AND S. LOJASIEWICZ JR., *Lagrange-type extremal trajectories in differential inclusions*, Systems Control Lett., 19 (1992), pp. 241–247.

[23] P. D. LOEWEN, *Optimal Control via Nonsmooth Analysis*, CRM Proceedings and Lecture Notes, Vol. 2, American Mathematical Society, Providence, RI, 1993.

[24] P. D. LOEWEN AND R. T. ROCKAFELLAR, *The adjoint arc in nonsmooth optimization*, Trans. Amer. Math. Soc., 325 (1991), pp. 39–72.

[25] ———, *Optimal control of unbounded differential inclusions*, SIAM J. Control Optim., 34 (1994), pp. 442–470.

[26] B. S. MORDUKHOVICH, *Maximum principle in problems of time optimal control with nonsmooth constraints*, J. Appl. Math. Mech., 40 (1976), pp. 960–969.

[27] ———, *Metric approximations and necessary optimality conditions for general classes of nonsmooth extremal problems*, Soviet Math. Dokl., 22 (1980), pp. 526–530.

[28] ———, *Approximation Methods in Problems of Optimization and Control*, Nauka, Moscow, 1988; revised English transl. to appear, Wiley-Interscience.

[29] ———, *On variational analysis of differential inclusions*, in Optimization and Nonlinear Analysis, A. Ioffe et al., eds., Pitman Res. Notes Math. Ser. 244, Longman Sci. Tech., Harlow, 1992, pp. 199–213.

[30] ———, *Sensitivity analysis in nonsmooth optimization*, in Theoretical Aspects of Industrial Design, D. A. Field and V. Komkov, eds., Society for Industrial and Applied Mathematics, Philadelphia, PA, 1992, pp. 32–46.

[31] ———, *Complete characterization of openness, metric regularity, and Lipschitzian properties of multifunctions*, Trans. Amer. Math. Soc., 340 (1993), pp. 1–35.

[32] ———, *Lipschitzian stability of constraint systems and generalized equations*, Nonlinear Anal. Theory Methods Appl., 32 (1994), pp. 173–206.

[33] ———, *Generalized differential calculus for nonsmooth and set-valued mappings*, J. Math. Anal. Appl., 182 (1994), pp. 250–288.

[34] ———, *Stability theory for parametric generalized equations and variational inequalities via nonsmooth analysis*, Trans. Amer. Math. Soc., 343 (1994), pp. 609–658.

[35] ———, *Optimization and finite difference approximations of differential inclusions with free time*, in the Proc. IMA Workshop on Nonsmooth Analysis and Geometric Nethods in Deterministic Optimal Control, B. S. Mordukhovich and H. J. Sussmann, eds., Springer-Verlag, New York, 1995.

[36] E. Polak, *Computational Methods in Optimization: A Unified Approach*, Academic Press, New York, 1971; 2nd revised edition, Academic Press, to appear.

[37] E. S. Polovinkin and G. V. Smirnov, *An approach to differentiation of multifunctions and necessary optimality conditions for differential inclusions*, Differential Equations, 22 (1986), pp. 660–668.

[38] B. N. Pshenichnyi, *Convex Analysis and Extremal Problems*, Nauka, Moscow, 1980. (Russian)

[39] R. T. Rockafellar, *Conjugate convex functions in optimal control and the calculus of variations*, J. Math. Anal. Appl., 32 (1970), pp. 174–222.

[40] ———, *Extensions of subgradient calculus with applications to optimization*, Nonlinear Anal. Theory Methods Appl., 9 (1985), pp. 665–698.

[41] ———, *Lipschitzian properties of multifunctions*, Nonlinear Anal. Theory Methods Appl., 9 (1985), pp. 867–885.

[42] ———, *Maximal monotone relations and the second derivatives of nonsmooth functions*, Ann. Inst. H. Poincaré. Anal. Non Linéare, 2 (1985), pp. 167–184.

[43] ———, *Dualization of subgradient conditions for optimality*, Nonlinear Anal. Theory Methods Appl., 20 (1993), pp. 627–646.

[44] ———, *Lagrange multipliers and optimality*, SIAM Rev., 35 (1993), pp. 183–238.

[45] R. T. Rockafellar and R. J.-B. Wets, *Variational Analysis*, Springer-Verlag, New York, to appear.

[46] J. D. L. Rowland and R. B. Vinter, *Dynamic optimization problems with free time and active state constraints*, SIAM J. Control Optim., 31 (1993), 677–691.

[47] B. Sendov and V. A. Popov, *The Averaged Moduli of Smoothness*, Wiley-Interscience, New York, 1988.

[48] G. V. Smirnov, *Discrete approximations and optimal solutions to differential inclusions*, Kibernetika (1991), pp. 76–79. (Russian)

[49] R. B. Vinter and P. D. Woodford, *On the occurence of intermediate local minimizers that are not strong local minimizers*, Preprint, April 1994.

[50] J. Warga, *Optimal Control of Differential and Functional Equations*, Academic Press, New York, 1972.

[51] ———, *Controllability, extremality, and abnormality in nonsmooth optimal control*, J. Optim. Theory Appl., 41 (1983), pp. 239–260.

[52] L. C. Young, *Lectures on the Calculus of Variations and Optimal Control Theory*, Saunders, Philadelphia, PA, 1969.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 33, No. 3, pp. 916–936, May 1995

© 1995 Society for Industrial and Applied Mathematics
012

SINGULAR OPTIMAL STOCHASTIC CONTROLS I: EXISTENCE*

ULRICH G. HAUSSMANN[†] AND WULIN SUO[‡]

Abstract. We apply the compactification method to study the control problem where the state is governed by an Ito stochastic differential equation allowing both classical and singular control. The problem is reformulated as a martingale problem on an appropriate canonical space after the relaxed form of the classical control is introduced. Under some mild continuity hypotheses on the data, it is shown by purely probabilistic arguments that an optimal control for the problem exists. The value function is shown to be Borel measurable.

Key words. singular controls, relaxed controls, control rules, existence theory, pseudopath topology, compactification method, value function

AMS subject classifications. 49J30, 49A55, 60G44, 93E20

1. Introduction.

The class of singular stochastic control problems, which has been studied extensively in recent years, deals with systems described by a stochastic differential equation in which one restricts the cumulative displacement of the state caused by control to be of an additive nature. More precisely, in this paper we study the existence of optimal controls to the problem in which the state evolves according to the d-dimensional stochastic differential equation

$$x_t = x + \int_s^t b(\theta, x_\theta, u_\theta)d\theta + \int_s^t \sigma(\theta, x_\theta, u_\theta)dB_\theta + \int_s^t g(\theta)dv_\theta$$

on some filtered probability space $(\Omega, \mathcal{F}, \mathcal{F}_t, P)$, where $b(\cdot,\cdot,\cdot)$, $\sigma(\cdot,\cdot,\cdot)$, $g(\cdot)$ are given deterministic functions, $(B_t,\ t \geq 0)$ is a d-dimensional Brownian motion (in fact B need not be d-dimensional), x is the initial state at time s, and $u:\ [0,T] \mapsto U$, $v:\ [0,T] \mapsto I\!R^k$ with v nondecreasing componentwise, stand for the controls.

The expected cost has the form

$$J(\alpha) \equiv E^P \left\{ \int_s^T f(t, x_t, u_t)dt + \int_{[s,T)} c(t) \cdot dv_t \right\},$$

where $f(\cdot,\cdot,\cdot):\ [0,T] \times I\!R^d \times U \mapsto I\!R$, $c(\cdot):\ [0,T] \mapsto I\!R^k_+$ are given. We assume that the cost of applying the singular control is positive, i.e., $c^i(\cdot) > 0$, $i = 1, \ldots, k$.

Some special cases of the one-dimensional problem of this type (without the *classical control* u) have been studied by many authors including Beněs, Shepp, and Witsenhausen [2]; Chow, Menaldi, and Robin [6]; El Karoui and Karatzas [8]; Harrison and Taksar [11]; Karatzas [19], [20]; Karatzas and Shreve [21]; Lehoczky and Shreve [23]; Ma [24]; Menaldi and Robin [25], [26]; and Sun [32]. It is shown that the value function satisfies a variational inequality which gives rise to a free-boundary problem, and the optimal state process is a diffusion reflected at the free boundary. This approach encounters substantial difficulties for the problems in high dimension

* Received by the editors June 15, 1993; accepted for publication (in revised form) January 12, 1994. This work was supported by Natural Sciences and Engineering Research Council of Canada grant 88051.

[†] Department of Mathematics, University of British Columbia, Vancouver, British Columbia, Canada V6T 1Z2.

[‡] Faculty of Management, University of Toronto, Toronto, Ontario, Canada M5S 1V4.

due to the lack of information about the regularity of the associated free boundary. In Soner and Shreve [30], a special two-dimensional problem ($b = 0$, $\sigma = I$) was considered. It was shown there that the associated free boundary is smooth enough to construct a reflected diffusion in the continuation region. However, as the authors pointed out, the method depends heavily on the special features of the problem and cannot be extended to general problems. Another result about high-dimensional problems can be found in Menaldi and Taksar [27], who considered the n-dimensional case with $b =$ const, $\sigma =$ const. It was shown that the value function satisfies the associated Hamilton–Jacobi–Bellman equation, and the existence of the optimal control was proved without requiring any regularity about the free boundary.

In this paper, we apply the compactification method used in Haussmann [12], Haussmann and Lepeltier [13], and El Karoui, Nguyen, and Jeanblanc-Picqué [7] to show the existence of the optimal control. An advantage of this approach is that it does not require any regularity of the value function and thus needs only very mild hypothesis on the data. Unfortunately, it does not show any particular structure of the optimal control.

This paper is organized as follows: In §2 the concept of relaxed control (for the classical control u) is introduced and the problem is reformulated as a martingale problem on a canonical space. In §2.3 a topology is given to the canonical space to make it a metrizable separable space, which is necessary to apply Prohorov's theorem. A relative compactness criterion for the subsets of the canonical space is given. In §3 it is proved that the optimal control exists, and the value function is shown to be Borel measurable. We also make some comments about possible generalizations of the model and constraints on the state in §4.

This paper is one of a series that applies the direct method to study the singular control problems. In [14] we prove the abstract form of the dynamic programming principle under the same conditions on the data assumed in this paper. When assuming Lipschitz continuity of the data, it is shown that the value function is continuous and is the unique viscosity solution of the corresponding Hamilton–Jacobi–Bellman equation. This method can also be used to show (cf. [33]) that there is a region A such that the optimal control yields no jump when the state is in A, and there exists an optimal control to keep the state in the closure of A after a possible initial jump. In [15] the adaptedness of the optimal control to the state is considered.

Note that in the literature *singular control* usually means that there is no classical control u involved. But our model obviously includes this case by letting U be a singleton (or by letting σ, b, f be independent of u). It also includes the monotone follower as well as the bounded variation problem (cf. §4).

Finally, we list some notation that will be used throughout this paper:

- $I\!R^d, I\!R$ denote the d-dimensional Euclidean space and the real line, respectively. $I\!R_+ = \{x \in I\!R, x \geq 0\}$, and $I\!R^d_+$ is defined similarly. For $x = (x^i), y = (y^i) \in I\!R^d$, $x \cdot y = \sum_{i=1}^d x^i y^i$.
- $T > 0$ is the fixed horizon, and $\Sigma = [0,T] \times I\!R^d$.
- $C[0,T]$ denotes the collection of real-valued continuous functions defined on $[0,T]$, and $C^d[0,T]$ denotes the collection of $I\!R^d$-valued continuous functions defined on $[0,T]$.
- $\mathcal{D}^d[0,T]$ denotes the collection of $I\!R^d$-valued functions defined on $[0,T]$ that are left continuous and have right limits (i.e., *lcrl* functions).
- $\mathcal{A}^k[0,T]$ denotes the collection of functions $a : [0,T] \mapsto I\!R^k$ such that $a = (a^i) \in \mathcal{D}^k[0,T]$ and a^i is nondecreasing with $a^i(0) = 0$, $i = 1, \ldots, k$.

- $\mathcal{S}^{l\times k}$ is the space of $l\times k$ matrices with the $l\times k$-dimensional Euclidean norm.
- If Y is a metric space, $\mathcal{B}(Y)$ denotes the corresponding Borel σ-field, and $f\in\mathcal{B}(Y)$ means that f is a $\mathcal{B}(Y)$-measurable real-valued function. We denote by $I\!M_+(Y)$, $I\!M_1(Y)$ the space of nonnegative Radon measures and the space of probabilities on Y, respectively.
- If X is a random variable on a probability space $(\Omega,\mathcal{F},P)$, the expectation of X will be denoted by $E^P(X)$. $\mathcal{M}_2^c$ ($\mathcal{M}_2^{c,\mathrm{loc}}$) is the family of continuous square integrable martingales (local martingales, respectively) on some given probability space $(\Omega,\mathcal{F},P)$ with a filtration $\{\mathcal{F}_t\}$.

2. Formulation of the problem. We consider the following optimal control problem in which we allow both classical control and singular control to act at the same time. The dynamics are in the form

$$x_t=x+\int_s^t b(\theta,x_\theta,u_\theta)d\theta+\int_s^t\sigma(\theta,x_\theta,u_\theta)dB_\theta+\int_s^t g(\theta)dv_\theta \quad \text{a.s.} \tag{2.1}$$

for $(s,x)\in\Sigma$, $s\le t\le T$, where

- $u_\theta\in U$, $s\le\theta\le T$, and U, called the control set, is a compact metric space;
- $(\sigma,b):\Sigma\times U\mapsto\mathcal{S}^{d\times d}\times I\!R^d$, $g:\ [0,T]\mapsto\mathcal{S}^{d\times k}$; $\sigma(t,x,u)$, $b(t,x,u)$ are bounded, measurable, and continuous with respect to (x,u); $g(t)$ is continuous on $[0,T]$;
- $(B_t, 0\le t\le T)$ is a d-dimensional Brownian motion on some probability space;
- $v\in\mathcal{A}^k[0,T]$.

We introduce the concept of controls to the stochastic differential equation (2.1).

DEFINITION 2.1. *A control is a term* $\alpha=(\Omega,\mathcal{F},\mathcal{F}_t,P,B_t,x_t,u_t,v_t,s,x)$ *such that*

(1) $(s,x)\in\Sigma$;
(2) $(\Omega,\mathcal{F},P)$ *is a probability space with the filtration* $\{\mathcal{F}_t\}_{t\ge0}$;
(3) u_t *is a* U*-valued process, progressively measurable with respect to* $\{\mathcal{F}_t\}_{t\ge0}$;
(4) v *is* $I\!R_+^k$*-valued processes, progressively measurable with respect to* $\mathcal{F}_t$; *the sample paths of* v *are in* $\mathcal{A}^k[0,T]$, *i.e., for each* $\omega\in\Omega$, $v_\cdot(\omega)\in\mathcal{A}^k[0,T]$;
(5) B_t *is a standard d-dimensional Brownian motion on* $(\Omega,\mathcal{F},\mathcal{F}_t,P)$ *and* x_t, *the state process, is* $\mathcal{F}_t$*-adapted with sample paths in* $\mathcal{D}^d[0,T]$, *and such that* (2.1) *is satisfied. We assume that* $x_r=x$ *for* $0\le r\le s$.

We call (s,x) *the initial condition of the control* α.

The collection of controls with initial condition (s,x) is denoted by $\Lambda_{s,x}$. It is well known from the theory of stochastic differential equations that, under the above conditions, the set $\Lambda_{s,x}$ is nonempty for each fixed (s,x) (e.g., take u and v to be constants). The *cost* corresponding to the control α is defined in the form

$$J(\alpha)\equiv E^P\left\{\int_s^T f(t,x_t,u_t)dt+\int_{[s,T)}c(t)\cdot dv_t\right\}, \tag{2.2}$$

where

- $f:\Sigma\times U\mapsto I\!R$ is a measurable function and is lower semicontinuous in (x,u), satisfying

$$-K\le f(t,x,u)\le C(1+\|x\|^m),\quad (t,x,u)\in\Sigma\times U$$

for some constants $m\ge0$, $K\ge0$, and $C\ge0$;

- $c = (c^i) : [0,T] \mapsto I\!R^k$ is lower semicontinuous and $c^i > 0$, $1 \le i \le k$.

Throughout this paper, we write

$$\int_s^t k(\theta) \cdot da(\theta) = \sum_{i=1}^d \int_{[s,t)} k_i(\theta) da_i(\theta)$$

for any $I\!R^k$-valued Borel measurable functions $k = (k_i)$ and $a = (a_i) \in \mathcal{A}^k[0,T]$. For $v \in \mathcal{A}^k[0,T]$, define

$$G_t(v) = \int_{[s,t)} g(\theta) dv(\theta), \quad s < t \le T, \tag{2.3}$$

and $G_t(v) = 0$ if $0 \le t \le s$. It can be verified easily that $G_.(v) \in \mathcal{D}^d[0,T]$.

The value function of the problem is defined, for $(s,x) \in \Sigma$, as

$$W(s,x) = \inf_{\alpha \in \Lambda_{s,x}} J(\alpha). \tag{2.4}$$

A control $\alpha^* \in \Lambda_{s,x}$ is called an *optimal control* if $W(s,x) = J(\alpha^*)$.

Remark 2.2. Note that there is no terminal cost in our model. Because of the way we choose the topology on the canonical space, this method cannot treat the case with terminal cost. We will return to this point at the end of the paper.

2.1. Relaxed controls. In order to apply the compactification method, we now reformulate the problem. Since the Brownian motion in the definition of controls is unknown in advance, we can reformulate the above control problem as an equivalent martingale problem. This simplifies taking limits. In fact, let

$$\mathcal{L} \equiv \frac{1}{2} \sum_{i,j} a_{ij} \frac{\partial^2}{\partial x_i \partial x_j} + \sum_i b_i \frac{\partial}{\partial x_i},$$

where $a(t,x,u) = \sigma(t,x,u)\sigma(t,x,u)^*$. Then we can show that $\alpha \in \Lambda_{s,x}$ if and only if α satisfies (1)–(4) in Definition 2.1, and

(5′) x_t is an $\mathcal{F}_t$-adapted process with sample paths in $\mathcal{D}^d[0,T]$ such that

- $P(x_r = x,\ u_r = u^0,\ v_r = 0,\ 0 \le r \le s) = 1$, where $u^0 \in U$ is arbitrary but fixed,
- $\forall \phi \in C_b^2(I\!R^d)$, $\mathcal{M}\phi \in \mathcal{M}_2^c$, i.e., $M_t\phi$ $(0 \le t \le T)$ is a continuous square integrable martingale on the filtered probability space $(\Omega, \mathcal{F}, \mathcal{F}_t, P)$, where

$$\begin{aligned} M_t\phi(\omega) \equiv \phi(x_t) - \int_s^t \mathcal{L}\phi(\theta, x_\theta, u_\theta) d\theta - \int_s^t \nabla_x \phi(x_\theta) \cdot g(\theta) dv_\theta \\ - \sum_{s \le \theta < t} \left[\phi(x_{\theta+}) - \phi(x_\theta) - \nabla_x \phi(x_\theta) \cdot \triangle x_\theta \right]. \end{aligned}$$

Therefore, we can delete the term B_t from the notation of a control. The proof of the equivalence of the existence of weak solutions to a stochastic differential equation and the existence of solutions to the corresponding martingale problem, given by Ikeda and Watanabe in Proposition IV-2.1 of [16], also works here despite the extra term $G_.(v)$.

Next we introduce the concept of relaxed controls, which gives a more suitable topological structure when applying the compactification method. In a relaxed control problem, the U-valued process $\{u_t\}$ is replaced by an $\mathbb{M}_1(U)$-valued process $\{\mu_t\}$, where $\mathbb{M}_1(U)$ is the space of probability measures on U endowed with the topology of weak convergence. $\mathbb{M}_1(U)$ is also a compact metrizable space. If $\phi : U \mapsto \mathbb{R}$ is a bounded measurable function, then we extend ϕ to $\mathbb{M}_1(U)$ by letting

$$\phi(\mu) \equiv \int_U \phi(u)\mu(du).$$

DEFINITION 2.3. *$\alpha = (\Omega, \mathcal{F}, \mathcal{F}_t, P, x_t, \mu_t, v_t, s, x)$ is called a relaxed control if it satisfies conditions* (1), (2), *and* (4) *in Definition* 2.1, *and*

(3′) *μ_t is $\mathbb{M}_1(U)$-valued, progressively measurable with respect to $\{\mathcal{F}_t\}_{t\geq 0}$;*

(5″) *x_t is an $\mathcal{F}_t$-adapted process with sample paths in $\mathcal{D}^d[0,T]$ such that*

- *$P(x_r = x,\ \mu_r = \delta^0,\ v_r = 0,\ 0 \leq r \leq s) = 1$, where δ^0 is the Dirac measure at some arbitrary but fixed $u^0 \in U$;*
- *$\forall \phi \in C_b^2(\mathbb{R}^d)$, $\mathcal{M}\phi \in \mathcal{M}_2^c$, i.e., $M_t\phi$ $(0 \leq t \leq T)$ is a continuous square integrable martingale on the filtered probability space $(\Omega, \mathcal{F}, \mathcal{F}_t, P)$, where*

$$M_t\phi(\omega) \equiv \phi(x_t) - \int_s^t \mathcal{L}\phi(\theta, x_\theta, \mu_\theta)d\theta - \int_s^t \nabla_x\phi(x_\theta)\cdot g(\theta)dv_\theta$$
$$- \sum_{s\leq\theta<t} \left[\phi(x_{\theta+}) - \phi(x_\theta) - \nabla_x\phi(x_\theta)\cdot\triangle x_\theta\right].$$

Note that we never work with $\sigma(t,x,u)$. Instead, we will be working on $a(t,x,u)$ through the martingale formulation in Definition 2.3. It is not true in general that $\sigma(t,x,\mu)\sigma(t,x,\mu)^* = a(t,x,\mu)$ (cf. El Karoui, Nguyen, and Jeanblanc-Picqué [7]).

The collection of relaxed controls starting from time s with the initial state x is denoted by $\tilde{\Lambda}_{s,x}$. Note that $\Lambda_{s,x}$ can be imbedded into $\tilde{\Lambda}_{s,x}$ by letting $\mu_t(du) \equiv \delta_{u_t}(du)$. Here, δ_u denotes the Dirac measure at the point u. Hence, $\forall\alpha \in \Lambda_{s,x}$, there exists an $\tilde{\alpha} \in \tilde{\Lambda}_{s,x}$ such that

$$J(\tilde{\alpha}) = J(\alpha),$$

and, therefore,

$$\inf_{\tilde{\alpha}\in\tilde{\Lambda}_{s,x}} J(\tilde{\alpha}) \leq \inf_{\alpha\in\Lambda_{s,x}} J(\alpha).$$

In order to get the reverse inequality, we define for each $(t,x) \in \Sigma$ a subset of $\mathcal{S}^{d\times d} \times \mathbb{R}^d \times \mathbb{R}$:

$$K(t,x) \equiv \{(a(t,x,u), b(t,x,u), z) :\ z \geq f(t,x,u), u \in U\}. \tag{2.5}$$

PROPOSITION 2.4. *Assume that $K(t,x)$ is convex $\forall(t,x) \in \Sigma$. Then $\forall\tilde{\alpha} \in \tilde{\Lambda}_{s,x}$, there exists an $\alpha \in \Lambda_{s,x}$ such that*

$$J(\alpha) \leq J(\tilde{\alpha}).$$

The proof in Haussmann and Lepeltier [13, Thm. 3.6.], can be carried over here without any change. From this result we know that when $K(t,x)$ is convex,

$$\inf_{\tilde{\alpha}\in\tilde{\Lambda}_{s,x}} J(\tilde{\alpha}) = \inf_{\alpha\in\Lambda_{s,x}} J(\alpha).$$

Hence we can restrict our attention to the case of relaxed controls. Moreover, if the infimum over relaxed controls is attained, then so is the infimum over ordinary controls.

If the convexity condition fails, our existence results pertain only to relaxed controls.

2.2. Control rules. Now we choose a canonical space to simplify the arguments in the compactification method. Define

$$\mathcal{U} \equiv \{\mu : \ [0,T] \mapsto I\!\!M_1(U) \text{ is Borel measurable}\}.$$

Consider the canonical space

$$\mathcal{X} \equiv \mathcal{D}^d[0,T] \times \mathcal{U} \times \mathcal{A}^k[0,T].$$

All the above spaces are equipped with appropriate topologies, which we will discuss later. Let $\tilde{\mathcal{D}}$, $\tilde{\mathcal{U}}$, $\tilde{\mathcal{A}}$ denote their Borel σ-fields, and $\tilde{\mathcal{D}}_t$, $\tilde{\mathcal{U}}_t$, $\tilde{\mathcal{A}}_t$ denote the σ-fields up to time t, e.g.,

$$\tilde{\mathcal{A}}_t = \sigma\{v(\theta) : \ 0 \leq \theta \leq t\}, \ \ 0 \leq t \leq T.$$

Let

$$\tilde{\mathcal{X}} \equiv \tilde{\mathcal{D}} \times \tilde{\mathcal{U}} \times \tilde{\mathcal{A}},$$
$$\tilde{\mathcal{X}}_t \equiv \tilde{\mathcal{D}}_t \times \tilde{\mathcal{U}}_t \times \tilde{\mathcal{A}}_t.$$

DEFINITION 2.5. *A control rule is a probability R on the measurable space $(\mathcal{X}, \tilde{\mathcal{X}})$ such that*

$$\tilde{\alpha} = (\mathcal{X}, \tilde{\mathcal{X}}, \tilde{\mathcal{X}}_t, R, x_t, \mu_t, v_t, s, x)$$

is a relaxed control, where

$$x_t(\omega) = x_t, \ \mu_t(\omega) = \mu_t, \ v_t = v_t(\omega)$$

for $\omega = (x_\cdot, \mu_\cdot, v_\cdot) \in \mathcal{X}$, i.e.,

(1) *$(\mathcal{X}, \tilde{\mathcal{X}}, R)$ is a probability space with the filtration $\{\tilde{\mathcal{X}}_t\}$, $(s,x) \in \Sigma$, and*

$$R(x_r = x, \ \mu_r = \delta^0, \ v_r = 0, \ 0 \leq r \leq s) = 1;$$

(2) *$\forall \phi \in C_b^2(I\!\!R^d)$, $M_t\phi$ $(s \leq t \leq T)$ is in $\mathcal{M}_2^c$ on the filtered probability space $(\mathcal{X}, \tilde{\mathcal{X}}, \tilde{\mathcal{X}}_t, R)$ $(s \leq t \leq T)$, where*

$$M_t\phi(\omega) \equiv \phi(x_t) - \int_s^t \mathcal{L}\phi(\theta, x_\theta, \mu_\theta) d\theta - \int_s^t \nabla_x \phi(x_\theta) \cdot g(\theta) dv_\theta$$
$$- \sum_{s \leq \theta < t} \left[\phi(x_{\theta+}) - \phi(x_\theta) - \nabla_x \phi(x_\theta) \cdot \triangle x_\theta\right].$$

Let $J(s,R) \equiv J(\tilde{\alpha})$.

We denote by $\mathcal{R}_{s,x}$ the space of control rules such that the above relaxed control starts from time s with initial state x. We will suppress s in $J(s,R)$ when it is clear that $R \in \mathcal{R}_{s,x}$. Now the control problem can be described completely in terms of control rules.

PROPOSITION 2.6. *Let $\alpha = (\Omega, \mathcal{F}, \mathcal{F}_t, P, x_t, \mu_t, v_t, s, x)$ be a relaxed control*; *then there exists a control rule $R \in \mathcal{R}_{s,x}$ such that*

$$J(R) = J(\alpha).$$

Proof. The proof of this result is standard. Define a map $\Phi : \Omega \mapsto \mathcal{X}$ by

$$\Phi(\omega) \equiv (x.(\omega), \mu.(\omega), v.(\omega)).$$

This map is measurable, and $\mathcal{F}_t^{x,\mu,v} \subset \Phi^{-1}(\tilde{\mathcal{X}}_t) \subset \mathcal{F}_t$. We can show that

$$\tilde{\alpha} = (\Omega, \mathcal{F}, \Phi^{-1}(\tilde{\mathcal{X}}_t), P, x_t, \mu_t, v_t, s, x)$$

is also a relaxed control such that

$$J(\tilde{\alpha}) = J(\alpha).$$

Let $R = P \circ \Phi^{-1}$, which is a probability on $(\mathcal{X}, \tilde{\mathcal{X}})$. It is easily seen that

$$(\mathcal{X}, \tilde{\mathcal{X}}, \tilde{\mathcal{X}}_t, R, x_t, \mu_t, v_t, s, x)$$

is a relaxed control satisfying the requirements in the Proposition. □

2.3. The topology on the canonical space. In this section we define the topologies on the spaces $\mathcal{D}^d[0,T]$, $\mathcal{U}$, and $\mathcal{A}^k[0,T]$.

The space $\mathcal{D}^d[0,T]$. We first give a topology on $\mathcal{D}^d[0,\infty)$, the collection of *lcrl* functions on $\mathbb{R}_+$ taking values in $\mathbb{R}^d$. Then we can take $\mathcal{D}^d[0,T]$ as a subset of $\mathcal{D}^d[0,\infty)$ by extending each $x \in \mathcal{D}^d[0,T]$ to an $x' \in \mathcal{D}^d[0,\infty)$ through

$$x'(t) = \begin{cases} x(t), & 0 \le t < T, \\ x(T), & t \ge T, \end{cases}$$

and consider the induced topology on $\mathcal{D}^d[0,T]$.

Define a measure $\lambda(\cdot)$ on the Borel subsets of $\mathbb{R}_+$ by $\lambda(dt) = \mathrm{e}^{-t}dt$. For a Borel-measurable function $f : \mathbb{R}_+ \mapsto \mathbb{R}^d$, the image of the measure $\lambda(\cdot)$ under the mapping $t \mapsto (t, f(t))$ is a called the *pseudopath* of the function f. It is a probability law on $[0,\infty] \times \bar{\mathbb{R}}^d$ and is denoted by $\Psi(f)$. It is clear that Ψ identifies two functions if and only if they are equal almost everywhere in the Lebesgue sense and, in particular, Ψ is one-to-one on $\mathcal{D}^d[0,\infty)$. Thus it provides us with an imbedding of $\mathcal{D}^d[0,\infty)$ into the compact Polish space $\bar{\mathcal{P}}$ of all the probabilities on $[0,\infty] \times \bar{\mathbb{R}}^d$ (with the topology of weak convergence). The topology that $\bar{\mathcal{P}}$ induces on $\mathcal{D}^d[0,\infty)$ via the mapping Ψ is called the *pseudopath topology*, and makes $\mathcal{D}^d[0,\infty)$ a separable metrizable space. The associated Borel σ-algebra on $\mathcal{D}^d[0,\infty)$ is the same one that we get from the Skorohod topology. In fact, Lemma 1 in [28] tells us that convergence in the pseudopath topology is just convergence in measure. Let $\tilde{\mathcal{D}}$ denote the Borel σ-field on $\mathcal{D}^d[0,T]$, and $\tilde{\mathcal{D}}_t$ be the Borel σ-field up to time t, i.e.,

$$\tilde{\mathcal{D}}_t = \sigma\{x(\theta) : 0 \le \theta \le t\}, \quad 0 \le t \le T.$$

Now we state a relative compactness criterion for subsets of $\mathcal{D}^d[0,\infty)$. For $x \in \mathcal{D}^d[0,\infty)$, define $x^* = \sup_t \|x(t)\|$. For $u = (u^i), v = (v^i) \in \mathbb{R}^d$, $u < v$ means $u^i < v^i$ $(1 \le i \le d)$. Let $N^{uv}(x) = \sum_1^d N^{u^i v^i}(x^i)$, where $N^{u^i v^i}(x^i)$ denotes the number of upcrossings of $x^i(\cdot)$ on $[0,\infty)$ between the levels u^i and v^i. Then a subset $A \subset \mathcal{D}^d[0,\infty)$ such that

$$\sup_{x \in A} x^* < \infty, \quad \sup_{x \in A} N^{uv}(x) < \infty \tag{2.6}$$

for any $u < v$ is relatively compact in $\mathcal{D}^d[0,\infty)$ with the pseudopath topology. For details, see Meyer and Zheng [28].

Remark 2.7. As pointed out by Meyer and Zheng [28], $\mathcal{D}^d[0,\infty)$ with the pseudopath topology is not a Polish space. But from the definition we know that it is homeomorphic to a subspace of the Polish space $\bar{\mathcal{P}}$, and hence is a separable metric space. □

The space $\mathcal{U}$. $\mathcal{U}$ is the space of measurable transformations $\mu : [0,T] \mapsto \mathbb{M}_1(U)$ endowed with the stable topology, which is defined as follows: for $A \in \mathcal{B}([0,T])$, $B \in \mathcal{B}(U)$, define

$$\bar{\mu}(A \times B) \equiv \int_A \mu_t(B)dt.$$

Then $\bar{\mu}$ can be extended uniquely to an element in $\mathbb{M}_+([0,T] \times U)$, the space of nonnegative Radon measures on $[0,T] \times U$. The stable topology on $\mathcal{U}$ is the weakest topology that renders continuous the mappings

$$\bar{\mu} \mapsto \int_0^T \int_U \phi(t,u)\bar{\mu}(dt,du)$$

for all bounded measurable functions ϕ that are continuous in u.

Under this topology, we know that $\mathbb{M}_+([0,T] \times U)$ is a compact separable metrizable space. $\mathcal{U}$ is also endowed with its Borel σ-field $\tilde{\mathcal{U}}$, which is the smallest σ-field such that the mappings

$$\mu \mapsto \int_0^T \int_U \mu_t(du)f(t,u)dt$$

are measurable, where f is a bounded measurable function continuous with respect to the variable u. The filtration $\tilde{\mathcal{U}}_t$ is the σ-field generated by $\{1_{[0,t]}\mu, \quad \mu \in \mathcal{U}\}$. From the definition of the stable topology, we know that $\tilde{\mathcal{U}}_t$ is generated by the sets of the form

$$\left\{\mu : \int_0^s \mu_\theta d\theta \in B\right\}$$

with $s \leq t$ and B a Borel set in $\mathbb{M}_+(U)$.

For more details, see Haussmann and Lepeltier [13].

The space $\mathcal{A}^k[0,T]$. Let $\mathcal{V}$ be the collection of functions $a : [0,T] \mapsto \mathbb{R}$ such that each $a(\cdot)$ is of bounded variation and left continuous. We assume $a(0) = 0$. We first consider a topology on $\mathcal{V}$.

Let $\mathbb{M}[0,T]$ be the collection of signed Radon measures on $[0,T]$. Then there is a one-to-one correspondence between $\mathcal{V}$ and $\mathbb{M}[0,T])$: for $a \in \mathcal{V}$, let

$$\nu_a([s,t)) \equiv a(t) - a(s) \quad \forall\, 0 \leq s \leq t \leq T.$$

So we need only consider a topology on $\mathbb{M}[0,T]$, then we can get the induced topology on $\mathcal{V}$.

Denote by $C[0,T]$ the collection of real-valued continuous functions on $[0,T]$. It is well known that $\mathbb{M}[0,T]$, and, therefore, $\mathcal{V}$ is the dual space of $C[0,T]$ with the supremum norm: for $f \in C[0,T]$,

$$\|f\| = \sup_{0 \leq t \leq T} |f(t)|,$$

and the corresponding weak* topology on $\mathbb{M}[0,T]$ is the topology induced by the weak convergence of measures. It can be easily seen that the measures of the form $\sum_1^N a_i\delta_{x_i}$ with N finite and a_i, x_i rational comprise a countable dense subset of $\mathbb{M}[0,T]$; therefore, $\mathbb{M}[0,T]$, with the weak* topology, is separable.

Let $\mathcal{A}^0 = \{a \in \mathcal{V} : a$ nondecreasing$\}$. Then it is a closed subset of $\mathcal{V}$ under the weak convergence topology, and the corresponding closed subset in $\mathbb{M}[0,T]$ is $\mathbb{M}_+[0,T]$. By Stroock and Varadhan [31] we know that $\mathbb{M}_+[0,T]$ is a metrizable space. Therefore, we can conclude that under the weak convergence topology $\mathcal{A}^0$ is a separable metric space.

We state the following theorem which will be used later.

THEOREM 2.8. *For any constant $C > 0$,*

$$\{\lambda \in \mathbb{M}_+[0,T] : \lambda([0,T]) \leq C\} \tag{2.7}$$

is a compact subset of $\mathbb{M}_+[0,T]$.

Proof. Note that the set (2.7) is a closed subset in $\mathbb{M}[0,T]$ in both the (variation) norm topology and the weak* topology. Also notice that if $\lambda \in \mathbb{M}_+[0,T] \subset \mathbb{M}[0,T]$, then the norm of λ will be

$$\|\lambda\| = \sup_{\|f\|\leq 1}\left|\int_0^T f d\lambda\right| = \lambda([0,T]),$$

i.e., the set (2.7) is a norm-bounded subset of $\mathbb{M}[0,T]$. Therefore by the Banach–Alaoglu theorem (see Larsen [22, Thm. 9.4.1]) we can conclude that (2.7) is a compact subset of $\mathbb{M}[0,T]$, and, therefore, a compact subset of $\mathbb{M}_+[0,T]$. □

Finally, observe that $\mathcal{A}^k[0,T] = (\mathcal{A}^0)^k$ and consider the product topology on $\mathcal{A}^k[0,T]$ inherited from the weak topology of $\mathcal{A}^0$. We can state the following result.

COROLLARY 2.9. *$\mathcal{A}^k[0,T]$ is metrizable and separable. For $v_n, v \in \mathcal{A}^k[0,T]$, $v_n \to v$ if and only if*

$$\int_0^T f(t)\cdot dv_n(t) \to \int_0^T f(t)\cdot dv(t)$$

for any $f \in C([0,T],\mathbb{R}^k)$. Moreover, the set

$$V_M = \{v \in \mathcal{A}^k[0,T] :\ \|v(T)\| \leq M\} \tag{2.8}$$

is compact for any constant $M > 0$.

We write down the following observation for later use. Its proof is obvious from the relative compactness criterion for pseudopath topology on $\mathcal{D}^d[0,T]$. Recall that the map $G : \mathcal{A}^k[0,T] \mapsto \mathcal{D}^d[0,T]$ is defined by (2.3).

LEMMA 2.10. *For any constant $M > 0$*

$$G(V_M) = \{G.(v) :\ v \in V_M\}$$

is a relatively compact subset of $\mathcal{D}^d[0,T]$, where V_M is defined by (2.8).

From now on we will always use the notation $\Omega = \mathcal{X}$, $\mathcal{F} = \tilde{\mathcal{X}}$, and $\mathcal{F}_t = \tilde{\mathcal{X}}_t$. It is well known that $\mathbb{M}_1(\Omega)$, endowed with the Prohorov weak convergence topology, is then a separable metrizable space. Denote the collection of all the control rules with initial condition (s,x) by $\mathcal{R}_{s,x}$, which is a subset of $\mathbb{M}_1(\Omega)$. For any real number λ, define

$$\mathcal{R}_{s,x}^{\lambda} \equiv \{P \in \mathcal{R}_{s,x}, J(s,P) \leq \lambda\}.$$

PROPOSITION 2.11. *There exists a constant $C \geq 0$ such that for*

$$\lambda(x) = C(1 + \|x\|^m), \quad x \in \mathbb{R}^d, \tag{2.9}$$

we have $\mathcal{R}_{s,x}^{\lambda(x)} \neq \emptyset$ for each $(s, x) \in \Sigma$. Recall that m is given in the definition of f.

Proof. Under our assumptions, it is known from the theory of stochastic differential equations (cf. Stroock and Varadhan [31]) that there exists a $P_0 \in \mathcal{R}_{s,x}$ with

$$P_0(v_t = 0,\ \mu_t = \delta_{u^0},\ 0 \leq t \leq T) = 1.$$

Then from the definition of control rule we know that under P_0,

$$x_t = x + \int_s^t b(\theta, x_\theta, u^0)d\theta + \int_s^t \sigma(\theta, x_\theta, u^0)dB_\theta;$$

therefore, from the boundedness of b, σ, and the Burkholder–Davis–Gundy inequality we have

$$\begin{aligned} E^{P^0}\|x_t\|^m &\leq C\left\{\|x\|^m + E^{P_0}\left\|\int_s^t b(\theta, x_\theta, u^0)d\theta\right\|^m \right. \\ &\qquad \left. + E^{P_0}\sup_{0\leq\theta\leq t}\left\|\int_s^\theta \sigma(\theta', x_{\theta'}, u^0)dB_{\theta'}\right\|^m\right\} \\ &\leq C\left\{1 + \|x\|^m + E^{P_0}\left(\int_s^t \operatorname{tr}\big(a(\theta, x_\theta, u^0)\big)d\theta\right)^{\frac{m}{2}}\right\} \\ &\leq C\left(1 + \|x\|^m\right), \end{aligned}$$

where C is a constant independent of x. Now we have by definition

$$\begin{aligned} J(s, P_0) &= E^{P_0}\left\{\int_s^T f(\theta, x_\theta, u^0)d\theta\right\} \leq C\left(1 + \int_s^T E^{P_0}\|x_\theta\|^m d\theta\right) \\ &\leq C\left(1 + \|x\|^m\right), \end{aligned}$$

where C is independent of s, x. The proposition is thus proved by letting $\lambda(x) = C(1 + \|x\|^m)$. □

3. Existence of optimal controls. In order to show the existence of optimal controls, let us reformulate the problem as follows. Consider the stochastic differential equation (2.1), and let

$$y = x - G(v);$$

then $x_t = y_t + G_t(v)$ $(0 \leq t \leq T)$, and (2.1) becomes

$$y_t = x + \int_s^t b(\theta, x_\theta, u_\theta)d\theta + \int_s^t \sigma(\theta, x_\theta, u_\theta)dB_\theta; \tag{3.1}$$

this is a continuous process. We use this idea, but in the martingale setting, to replace the definition of a control rule.

PROPOSITION 3.1. *$P \in \mathcal{R}_{s,x}$ if and only if there exists an $\mathcal{F}_t$-adapted process y such that*

(1) *$y.$ is continuous with probability one (w.p. 1) and $P(x. = y. + G.(v)) = 1$;*
(2) *$P(x_r = x,\ \mu = \delta^0,\ v_r = 0,\ 0 \le r \le s) = 1$;*
(3) *$\tilde{M}_t\phi \in \mathcal{M}_2^{\mathrm{c}}$ for every $\phi \in C_b^2(I\!R^d)$, where*

$$\tilde{M}_t\phi(\omega) = \phi(y_t(\omega)) - \int_s^t \tilde{\mathcal{L}}\phi(\theta, x_\theta, y_\theta, \mu_\theta)d\theta, \tag{3.2}$$

and

$$\tilde{\mathcal{L}}\phi(\theta, x, y, u) \equiv \frac{1}{2}\sum_{ij=1}^{d} a_{ij}(\theta, x, u)\frac{\partial^2\phi(y)}{\partial y^i \partial y^j} + \sum_{i=1}^{d} b_i(\theta, x, u)\frac{\partial\phi(y)}{\partial y^i}.$$

Proof. The proof is comparable to that in Suo [33]. □

Some routine calculations can lead to the following result.

LEMMA 3.2. *Assume $P \in \mathcal{R}_{s,x}$. For any ϕ, $\psi \in C_b^2(I\!R^d)$ we have*

$$\langle \tilde{M}\phi, \tilde{M}\psi\rangle_t = \sum_{i,j=1}^{d} \int_s^t a_{ij}(\theta, x_\theta, \mu_\theta)\frac{\partial\phi(y_\theta)}{\partial y^i}\frac{\partial\psi(y_\theta)}{\partial y^j}d\theta \tag{3.3}$$

under probability P. In particular, if we define, for $1 \le i \le d$,

$$\tilde{M}^i(t) = y^i(t) - x^i - \int_s^t b_i(\theta, x_\theta, \mu_\theta)d\theta, \tag{3.4}$$

i.e., $\tilde{M}_t\phi$ with $\phi(y) = y^i - x^i$, then $\tilde{M} \in \mathcal{M}_2^{\mathrm{c,loc}}$, and

$$\langle \tilde{M}^i, \tilde{M}^j\rangle_t = \int_s^t a_{ij}(\theta, x_\theta, \mu_\theta)d\theta,\ \ 1 \le i,\ j \le d.$$

Recall that $C^d[0,T]$ is the space of $I\!R^d$-valued continuous functions on $[0,T]$. We give $C^d[0,T]$ the uniform topology, i.e., for $x, y \in C^d[0,T]$, the distance between x and y is defined by

$$\rho(x,y) \equiv \max_{0 \le t \le T} \|x(t) - y(t)\|.$$

This makes $C^d[0,T]$ a Polish space (cf. Billingsley [3]).

For a sequence $P^n \in \mathcal{R}_{s_n,x_n}$ with $(s_n, x_n) \in \Sigma$, the probability law of the process y, defined in Proposition 3.1, under P^n is defined by

$$\bar{P}^n(C) \equiv P^n(\omega :\ y\,(\omega) \in C)$$

for $C \in \tilde{\mathcal{C}}$, where $\tilde{\mathcal{C}}$ is the Borel σ-field of $C^d[0,T]$.

PROPOSITION 3.3. *If the sequence (s_n, x_n) is bounded in Σ, then $\{\bar{P}^n\}$ is relatively compact. Moreover, for any $\varepsilon > 0$, there exists a compact subset $K \subset C^d[0,T]$ such that*

$$\bar{P}^n(K) \ge 1 - \varepsilon\ \ \forall n. \tag{3.5}$$

Proof. To show that $\{\bar{P}^n\}$ is relatively compact, we need only verify the following:

(a) $\lim_{A\to\infty}\inf_n P^n(\|y(0)\|\le A)=1$, and
(b) for any $\gamma>0$,

$$\lim_{\delta\downarrow 0}\limsup_n P^n\left[\sup_{\substack{0\le s<t\le T\\ t-s<\delta}}\|y(t)-y(s)\|\ge\gamma\right]=0. \tag{3.6}$$

Note that (a) is obvious from the fact $P^n(y(0)=x_n)=1$. By Billingsley [3, Thm. 12.3], (b) is implied by

$$E^{P^n}\|y(t_2)-y(t_1)\|^4\le C|t_2-t_1|^2 \tag{3.7}$$

for any $n\ge 1$, $0\le t_1,t_2\le T$, where C is a constant.

Recall from the definition of y that $P^n(y(t)=x_n,\ 0\le t\le s_n)=1$, and for $t\ge s_n$,

$$y(t)=x_n+\int_{s_n}^t b(\theta,x_\theta,\mu_\theta)d\theta+\tilde{M}_n(t)$$

with $\tilde{M}_n\in\mathcal{M}_2^{c,\mathrm{loc}}$ under P_n, and

$$\langle\tilde{M}_n\rangle_t=\int_{s_n}^t \mathrm{tr}\big(a(\theta,x_\theta,\mu_\theta)\big)d\theta,\ \ t\ge s_n.$$

It can be easily verified that (3.7) follows from the boundedness of the coefficients $\sigma(\cdot,\cdot,\cdot)$ and $b(\cdot,\cdot,\cdot)$ and the Burkholder–Davis–Gundy inequality.

We have therefore shown that the probability laws of y under P^n are relatively compact. The existence of a compact subset K such that (3.5) holds is a consequence of Prohorov's theorem. □

PROPOSITION 3.4. *If A is a bounded subset of $\mathbb{R}^d$, then*

$$\lim_{M\to\infty}\inf_{\substack{P\in\mathcal{R}^\lambda_{s,x}\\(s,x)\in[0,T]\times A}} P\{\omega:\|v_T\|\le M\}=1, \tag{3.8}$$

where $\|\cdot\|$ denotes the Euclidean norm in $\mathbb{R}^k$.

Proof. If $P\in\mathcal{R}^\lambda_{s,x}$, then $J(P)\le\lambda$. Since the function f is assumed to be bounded from below, there exists a constant $K>0$ such that $f\ge -K$. From the assumption that $c(t)$ is strictly positive and lower semicontinuous, there exists a constant $c_0>0$ such that $c^i(t)\ge c_0$ $(1\le i\le k,\ 0\le t\le T)$. Thus

$$\begin{aligned}\lambda\ge J(P)&=E^P\left\{\int_0^T f(\theta,x_\theta,\mu_\theta)d\theta+\int_s^T c(\theta)\cdot dv_\theta\right\}\\&\ge E^P\{-KT+c_0\|v_T\|\}.\end{aligned}$$

Therefore, $E^P\|v_T\|\le\bar{K}\equiv(\lambda+KT)/c_0$ and

$$P(\|v_T\|\ge M)\le\frac{\bar{K}}{M}.$$

The proposition is now obvious. □

Define a function on Ω by

$$\Gamma_s(\omega) \equiv \int_s^T f(\theta, x_\theta, \mu_\theta)d\theta + \int_s^T c(\theta) \cdot dv_\theta \tag{3.9}$$

for $\omega = (x., \mu., v.)$. Then we have

LEMMA 3.5. *$\Gamma_s(\cdot)$ is lower semicontinuous on Ω, i.e.,*

$$\liminf_n \Gamma_s(\omega_n) \geq \Gamma_s(\omega) \tag{3.10}$$

if $\omega_n \to \omega$ in Ω.

Proof. We show the case when $s = 0$; for general s the proof is similar. We can also assume that $f(t, \cdot, \cdot)$, $c(\cdot)$ are continuous. In fact, since they are lower semicontinuous, we can find sequences of continuous functions $\{f_m(t, \cdot, \cdot)\}$, $\{c_m(\cdot)\}$ such that

$$f_m(t, \cdot, \cdot) \uparrow f(t, \cdot, \cdot), \quad c_m^i(\cdot) \uparrow c^i(\cdot), \quad 1 \leq i \leq k,$$

and thus if the lemma is true for continuous functions, then $\forall m \geq 1$,

$$\begin{aligned}\liminf_n \Gamma_0(\omega_n) &\geq \liminf_n \left\{ \int_0^T f_m(\theta, x_n(\theta), \mu_n(\theta))d\theta + \int_0^T c_m(\theta) \cdot dv_n(\theta) \right\} \\ &\geq \int_0^T f_m(\theta, x(\theta), \mu(\theta))d\theta + \int_0^T c_m(\theta) \cdot dv(\theta).\end{aligned}$$

Let $m \to \infty$, and use the monotone convergence theorem to conclude the result.

We assume that $f(t, \cdot, \cdot)$ and $c(\cdot)$ are continuous. From $v_n \to v$ in $\mathcal{A}^k[0, T]$ we have

$$\int_0^T c(\theta) \cdot dv_n(\theta) \to \int_0^T c(\theta) \cdot dv(\theta).$$

Now we show that as $n \to \infty$,

$$\int_0^T \int_U f(\theta, x_n(\theta), u)\mu_\theta^n(du)d\theta \to \int_0^T \int_U f(\theta, x(\theta), u)\mu_\theta(du)d\theta. \tag{3.11}$$

Let $dQ^n \equiv \mu_\theta^n(du)d\theta$, $dQ \equiv \mu_\theta(du)d\theta$; then the right-hand side of (3.11) can be rewritten as

$$\begin{aligned}\int_{[0,T]\times U} f(\theta, x_n(\theta), u)dQ^n &= \int_{[0,T]\times U} f(\theta, x(\theta), u)dQ^n \\ &\quad + \int_{[0,T]\times U} [f(\theta, x_n(\theta), u) - f(\theta, x(\theta), u)]dQ^n.\end{aligned} \tag{3.12}$$

From the definition of stable topology we know as $n \to \infty$,

$$\int_{[0,T]\times U} f(\theta, x(\theta), u)dQ^n \to \int_{[0,T]\times U} f(\theta, x(\theta), u)dQ.$$

Now we show that the limit of the second integral in (3.12) is zero. For any positive integer m and constant $\gamma > 0$, let

$$\begin{aligned} g(t, x, u) &= f(t, x, u) - f(t, x(t), u), \\ A_m &= \left\{ (t, u) : \sup_{\|x - x(t)\| \leq 1/m} |g(t, x, u)| \geq \gamma \right\}. \end{aligned}$$

Then each t-section of the set A_m is a closed subset of U from the continuity of the function $f(t,\cdot,\cdot)$. Also $A_1 \supset A_2 \supset \cdots \supset A_m \supset A_{m+1} \supset \cdots$, and

$$\bigcap_m A_m = \emptyset.$$

Applying the results in Jacod and Mémin [17, Prop. 2.11] we can get

$$\limsup_n Q_n(A_m) \le Q(A_m), \quad \lim_m Q(A_m) = 0. \tag{3.13}$$

Let $B_n = \{(t,u) : |g(t,x_n(t),u)| \ge \gamma\}$. In order to show that the limit of the last integral in (3.12) is zero, we need only show

$$\lim_{n\to\infty} Q_n(B_n) = 0 \tag{3.14}$$

by Jacod and Mémin [17, Cor. 2.18]. For a given $\varepsilon > 0$, from (3.13) there exists an $M > 0$ such that $Q(A_M) \le \varepsilon$. Recall that convergence in pseudopath topology is equivalent to the convergence in Lebesgue measure, therefore, $x_n(\cdot) \to x(\cdot)$ in the Lebesgue measure l, and there exists N such that when $n \ge N$,

$$l\left\{t : \|x_n(t) - x(t)\| > \frac{1}{M}\right\} < \varepsilon.$$

Let $C_n^m = \{t : \|x_n(t) - x(t)\| > 1/m\}$; then it is obvious that

$$B_n \setminus (C_n^M \times U) \subset A_M,$$

and, therefore, we have

$$Q_n(B_n) \le Q_n(A_M) + Q_n(C_n^M \times U).$$

But $Q_n(C_n^M \times U) = l(C_n^M) < \varepsilon$, hence

$$\begin{aligned}\limsup_n Q_n(B_n) &\le \limsup_n Q_n(A_M) + \varepsilon \\ &\le Q(A_M) + \varepsilon < 2\varepsilon.\end{aligned}$$

Since ε is arbitrary, we have shown (3.14), and the lemma is thus proved. □

For $P \in \mathcal{R}_{s,x}$, we have by definition that

$$J(s,P) = E^P\Gamma_s,$$

i.e., the cost corresponding to the control rule P. We can now state the following result.

THEOREM 3.6. *The mapping $(s,x,P) \to J(s,P)$ is lower semicontinuous on $\{(s,x,P) : (s,x) \in \Sigma,\ P \in \mathcal{R}_{s,x}\}$ with the induced topology of $[0,T] \times I\!M_+(\Omega)$, i.e., if $(s_n,x_n) \in \Sigma$, $P_n \in \mathcal{R}_{s_n,x_n}$, $s_n \to s$, $x_n \to x$, and $P_n \to P \in \mathcal{R}_{s,x}$ weakly, then*

$$J(s,P) \le \liminf_n J(s_n,P_n). \tag{3.15}$$

Proof. Assume $(s_n, x_n, P_n) \to (s, x, P)$ with $P_n \in \mathcal{R}_{s_n,x_n}$, $P \in \mathcal{R}_{s,x}$. It suffices to consider two cases: $s_n \uparrow s$ or $s_n \downarrow s$. When $s_n \uparrow s$,

$$\begin{aligned} E^{P_n}(\Gamma_{s_n} - \Gamma_s) &= E^{P_n}\left\{\int_{s_n}^{s} f(\theta, x_\theta, \mu_\theta)d\theta + \int_{s_n}^{s} c(\theta)\cdot dv_\theta\right\} \\ &\geq E^{P_n}\left(\int_{s_n}^{s} f(\theta, x_\theta, \mu_\theta)d\theta\right) \geq -K(s-s_n), \end{aligned}$$

because we have assumed that $c \geq 0$. It follows that

$$\liminf_n E^{P_n}(\Gamma_{s_n} - \Gamma_s) \geq 0. \tag{3.16}$$

When $s_n \downarrow s$, (3.16) follows from

$$E^{P_n}(\Gamma_{s_n} - \Gamma_s) = \int_s^{s_n} f(\theta, x_n, u^0)d\theta,$$

where we have used the fact that $P_n(v_\theta = 0,\ 0 \leq \theta \leq s_n) = 1$.

In either case, from (3.16) and the lower semicontinuity of $\Gamma_s(\cdot)$, we have

$$\begin{aligned} \liminf_n J(s_n, P_n) &\geq \liminf_n E^{P_n}\Gamma_s + \liminf_n E^{P_n}(\Gamma_{s_n} - \Gamma_s) \\ &\geq \liminf_n E^{P_n}\Gamma_s \geq E^P\Gamma_s = J(s, P), \end{aligned}$$

i.e., $J(\cdot,\cdot)$ is lower semicontinuous. □

THEOREM 3.7. *For any $\lambda > 0$, if A is a compact set in $\mathbb{R}^d$, then $\cup\{\mathcal{R}^\lambda_{s,x} : s \in [0,T],\ x \in A\}$ is compact.*

Proof. Since $\mathbb{M}_1(\Omega)$ is metrizable, we need only to show that each sequence $\{P^n\} \subset \mathcal{R}^\lambda_{s_n,x_n}$ has a subsequence $\{P^{n_k}\}$ such that $P^{n_k} \to P \in \mathcal{R}^\lambda_{s,x}$ for some $s \in [0,T]$, $x \in A$. Because A is compact and $T < \infty$, we may assume $s_n \to s$, $x_n \to x$ for some $s \in [0,T]$, $x \in A$.

By Proposition 3.1, the process y, defined by

$$y_\cdot(\omega) = x_\cdot(\omega) - G_\cdot(v(\omega))$$

for $\omega = (x, \mu, v)$, has continuous sample paths under the probability P^n and is such that $\tilde{M}_t\phi \in \mathcal{M}_2^c$ ($\tilde{M}_t\phi$ is defined in Proposition 3.1) for each $\phi \in C_b^2(\mathbb{R}^d)$. Now we introduce the following auxiliary space:

$$\begin{aligned} \mathcal{Z} &\equiv \Omega \times C^d[0,T], \\ \tilde{\mathcal{Z}} &\equiv \mathcal{F} \times \tilde{\mathcal{C}}. \end{aligned}$$

Define a probability $\tilde{P}^n$ on $(\mathcal{Z}, \tilde{\mathcal{Z}})$ by the probability law of (x, μ, v, y) with respect to P^n, i.e., for $Z \in \tilde{\mathcal{Z}}$,

$$\tilde{P}^n(Z) \equiv P^n\left(\omega :\ (x_\cdot(\omega), \mu_\cdot(\omega), v_\cdot(\omega), y_\cdot(\omega)) \in Z\right).$$

In other words, for $F \in \mathcal{F}$, $C \in \tilde{\mathcal{C}}$,

$$\tilde{P}^n(F \times C) = \int_F \delta_C(y(\omega))dP^n(\omega).$$

We will show that the sequence $\tilde{P}^n$ is tight.

For a positive constant M, let $V_M = \{v \in \mathcal{A}^k[0,T] : \|v_T\| \leq M\}$. From Corollary 2.9 we know that V_M is a compact Borel subset of $\mathcal{A}^k[0,T]$. Proposition 3.4 implies that for any given $\varepsilon > 0$, there exists an M such that

$$P(\mathcal{D}^d[0,T] \times \mathcal{U} \times V_M) \geq 1 - \varepsilon \tag{3.17}$$

for every $P \in \mathcal{R}^\lambda_{s,x}$, $s \in [0,T]$, $x \in A$. Therefore, for the corresponding $\tilde{P}$, we have

$$\tilde{P}(\mathcal{D}^d[0,T] \times \mathcal{U} \times V_M \times C^d[0,T]) \geq 1 - \varepsilon. \tag{3.18}$$

By Proposition 3.3 we know that the probability laws of y with respect to P^n, denoted by $\bar{P}^n$, are relatively compact, and there exists a compact subset $K \subset C^d[0,T]$ with

$$\bar{P}^n(K) \geq 1 - \varepsilon,$$

or equivalently,

$$P^n(\omega : y.(\omega) \in K) \geq 1 - \varepsilon \ \forall n. \tag{3.19}$$

From Proposition 3.1, (3.19) may be written as

$$\tilde{P}^n(\Omega \times K) \geq 1 - \varepsilon \ \forall n. \tag{3.20}$$

We now consider the coordinate process x. Let

$$D = K + G(V_M) = \{y + G(v),\ y \in K,\ v \in V_M\}.$$

By Lemma 2.10 we know that $G(V_M)$ is a relatively compact subset of $\mathcal{D}^d[0,T]$ under the pseudopath topology. Since the uniform topology is stronger than the pseudopath topology, K is also a compact subset in $\mathcal{D}^d[0,T]$, hence so is D. From Proposition 3.1 we have

$$\tilde{P}^n(D \times \mathcal{U} \times V_M \times K) = \tilde{P}^n(\mathcal{D}^d[0,T] \times \mathcal{U} \times V_M \times K). \tag{3.21}$$

Let $S = D \times \mathcal{U} \times V_M \times K$. Since $\mathcal{U}$ is a compact space, we know that S is a relatively compact subset in $\mathcal{Z}$. Moreover, from (3.18), (3.20), and (3.21) we have

$$\tilde{P}^n(S) \geq 1 - 2\varepsilon$$

for every n. Thus $\{\tilde{P}^n\}$ is a tight sequence of probability measures on $\mathcal{Z}$. By the Prohorov theorem there is a subsequence $\{\tilde{P}^{n_k}\}$ and a probability $\tilde{P}$ on $(\mathcal{Z}, \tilde{\mathcal{Z}})$ such that $\tilde{P}^{n_k} \to \tilde{P}$ weakly. Define

$$P = \tilde{P}|_\Omega,$$

i.e., in the terminology of Jacod and Mémin [17], P is the *Ω-marginal* of $\tilde{P}$, then it is easy to see that $P^{n_k} \to P$ weakly. The proof of the theorem will be completed if $P \in \mathcal{R}^\lambda_{s,x}$. By Proposition 3.1 we need only to show that there exists a continuous process $Y.$ on $(\Omega, \mathcal{F}, \mathcal{F}_t, P)$ such that

(1) $P(Y. = x. - G.(v)) = 1$;
(2) $P(x_r = x,\ v_r = 0,\ 0 \leq r \leq s) = 1$;
(3) $\tilde{M}_t\phi \in \mathcal{M}_2^c \ \forall \phi \in C_b^2(\mathbb{R}^d)$; and

(4) $J(s,P) \leq \lambda$.

Note that (4) is obvious from Theorem 3.6. To show (1), note that the set $\{(\omega, y): x.(\omega) = y. + G.(v(\omega))\}$ is a closed subset of $\mathcal{Z} = \Omega \times C^d[0,T]$, and, therefore,

$$\tilde{P}(x. = y. + G.(v)) \geq \limsup \tilde{P}^n(x. = y. + G.(v)) = 1. \tag{3.22}$$

If we define $Y.(\omega) = x.(\omega) - G.(v(\omega))$, then $\tilde{P}((\omega, y): Y.(\omega) = y.) = 1$. Thus Y is a continuous process on $(\Omega, \mathcal{F}, \mathcal{F}_t, P)$, and (1) follows from (3.22). Moreover, $\{y(0) = x\}$ is closed in $\mathcal{Z}$, so

$$\tilde{P}\{y(0) = x\} \geq \limsup_k \tilde{P}^{n_k}\{y(0) = x\} = 1$$

or $P\{\omega: Y(0) = x\} = 1$. It follows that $x(0) = x$ almost surely (P).

For (2), let

$$B_m = \left\{\omega = (x., \mu., v.): \|x(t) - x\| \leq \frac{1}{m},\ v_t = 0,\ 0 < t \leq s - \frac{1}{m}\right\}.$$

It is easy to see that B_m is closed in Ω and $B_1 \supset B_2 \supset \cdots$, and

$$\bigcap_m B_m = \{\omega = (x, \mu, v): x_t = x,\ v_t = 0,\ 0 < t \leq s\}.$$

Since $P^{n_k} \in \mathcal{R}_{s_{n_k}, x_{n_k}}$, we know for each m, $P^{n_k}(B_m) = 1$ for large k since $x_{n_k} \to x$. But $P^{n_k} \to P$ weakly, so

$$P(B_m) \geq \limsup_k P^{n_k}(B_m) = 1,$$

and, therefore,

$$\begin{aligned} &P\{\omega: x_t = x, v_t = 0,\ 0 \leq t \leq s\} \\ &\quad = P\left(\bigcap_m B_m \bigcap \{x(0) = x\}\right) = \lim_m P(B_m) = 1. \end{aligned}$$

Finally, we prove (3). For any bounded continuous function $H(\cdot)$ on Ω, if we define

$$\tilde{H}(\tilde{\omega}) = H(\omega)\ \ \forall \tilde{\omega} = (\omega, y.) \in \Omega \times C^d[0,T],$$

then $\tilde{H}$ is a continuous function on $\mathcal{Z}$. For any fixed $t \geq s$, the function

$$z = (x, \mu, v, y) \in \mathcal{Z} \mapsto \bar{M}_t\phi(z) = \phi(y_t) - \int_s^t \tilde{\mathcal{L}}\phi(\theta, x_\theta, y_\theta, \mu_\theta)d\theta$$

is continuous on $\mathcal{Z}$. In fact, it can be shown as in the proof of Lemma 3.5 that the integral part of the function $\bar{M}_t\phi$ is continuous on $\mathcal{Z}$, and the continuity of the function

$$z = (x, \mu, v, y) \to \phi(y_t)$$

on $\mathcal{Z}$ follows from the fact that $C^d[0,T]$ is endowed with the uniform topology. Thus, for $0 \leq u < t \leq T$, the function

$$z = (x, \mu, v, y) \to \tilde{H}(z)\left[\bar{M}_t\phi(z) - \bar{M}_u\phi(z)\right]$$

is a bounded and continuous function on $\mathcal{Z}$, and since $\tilde{P}^{n_k} \to \tilde{P}$ weakly, we have

$$\lim_k E^{\tilde{P}^{n_k}} \left\{\tilde{H}[\bar{M}_t\phi - \bar{M}_u\phi]\right\} = E^{\tilde{P}} \left\{\tilde{H}[\bar{M}_t\phi - \bar{M}_u\phi]\right\}. \tag{3.23}$$

Again, by (3.22) and the definition of Y, we have $\tilde{P}(\tilde{M}.\phi = \bar{M}.\phi) = 1$, where $\tilde{M}_t\phi$ is defined by (3.2) with y replaced by Y. Thus (3.23) can be rewritten as

$$\lim_k E^{P^{n_k}} \left\{H[\tilde{M}_t\phi - \tilde{M}_u\phi]\right\} = E^{P} \left\{H[\tilde{M}_t\phi - \tilde{M}_u\phi]\right\}. \tag{3.24}$$

For every bounded continuous function H on Ω that is $\mathcal{F}_u$-measurable, the left-hand side of (3.24) is zero since $\tilde{M}\phi \in \mathcal{M}_2^c$ on the filtered probability space $(\Omega, \mathcal{F}, \mathcal{F}_t, P)$. By a routine limit procedure, we have

$$E^P[H(\tilde{M}_t\phi - \tilde{M}_u\phi)] = 0$$

for each bounded $\mathcal{F}_u$-measurable function H. Thus $(\tilde{M}_t\phi, \mathcal{F}_t)$ is a martingale under P. The continuity of this martingale follows from that of Y. The proof is therefore complete. □

We can now prove the main theorem of this paper.

THEOREM 3.8. *The control problem has an optimal solution, i.e., there exists a* $P^* \in \mathcal{R}_{s,x}$ *such that*

$$J(s, P^*) = \inf_{P \in \mathcal{R}_{s,x}} J(s, P).$$

Proof. By Proposition 2.11 and Theorem 3.7 we know that $\mathcal{R}_{s,x}^{\lambda(x)}$ is nonempty and compact. Moreover, it is obvious that

$$\inf_{P \in \mathcal{R}_{s,x}} J(s, P) = \inf_{P \in \mathcal{R}_{s,x}^{\lambda(x)}} J(s, P).$$

Now $J(s, \cdot)$ is a lower-semicontinuous function on $\mathcal{R}_{s,x}$, so it attains its minimum on the compact set $\mathcal{R}_{s,x}^{\lambda(x)} \subset I\!M_1(\Omega)$, i.e., there exists a $P^* \in \mathcal{R}_{s,x}^{\lambda(x)} \subset \mathcal{R}_{s,x}$ such that

$$J(s, P^*) = \inf_{P \in \mathcal{R}_{s,x}^{\lambda(x)}} J(s, P) = \inf_{P \in \mathcal{R}_{s,x}} J(s, P),$$

which completes the proof. □

Recall that the value function is defined by

$$W(s, x) = \inf_{P \in \mathcal{R}_{s,x}} J(s, P).$$

Let us define

$$\mathcal{R}_{s,x}^o = \{P \in \mathcal{R}_{s,x} : W(s, x) = J(s, P)\}.$$

By Theorem 3.8, $\mathcal{R}_{s,x}^o \neq \emptyset$ for any $(s, x) \in \Sigma$. It can be easily verified that it is a compact subset of $I\!M_1(\Omega)$.

Before we prove the measurability of the value function W, we give a result which is used in Haussmann and Suo [14]. We adopt the notations of Stroock and Varadhan [31, Chap. 12].

LEMMA 3.9. *The map $\mathcal{R}^o : \Sigma \mapsto \text{comp}(\mathbb{M}_1(\Omega))$ is Borel measurable. Moreover, there exists a measurable selector H of $\mathcal{R}^o$, i.e., $H(s,x) \in \mathcal{R}^o_{s,x}$ $\forall (s,x) \in \Sigma$ and $H : \Sigma \mapsto \mathbb{M}_1(\Omega)$ is Borel measurable.*

Proof. By [31, Lem. 12.1.8], we need only to show the following: for $(s_n, x_n) \in \Sigma$, $s_n \to s$, $x_n \to x$, $P^n \in \mathcal{R}^o_{s_n,x_n}$, there exists a subsequence P^{n_k} and $P \in \mathcal{R}^o_{s,x}$ such that $P^{n_k} \to P$.

Since $x_n \to x$, we may assume $\lambda(x_n) \leq \lambda$ for some constant λ. Therefore $\{P^n\} \subset \{\mathcal{R}^\lambda_{s,x},\ (s,x) \in [0,T] \times A\}$ with $A = \{x, x_1, x_2, \ldots\}$ a compact set. By Theorem 3.7, there exists a subsequence P^{n_k} and $P \in \mathcal{R}_{s,x}$ such that $P^{n_k} \to P$. From Theorem 3.6, it can be seen easily that $P \in \mathcal{R}^o_{s,x}$. The measurability of $\mathcal{R}^o$ is thus proved.

The existence of a measurable selector H is a consequence of [31, Thm. 12.1.10]. □

COROLLARY 3.10. *$W(\cdot,\cdot)$ is a Borel-measurable function.*

Proof. From Theorem 3.6 we know the map $(s,x,P) \mapsto J(s,P)$ is lower semicontinuous and thus Borel measurable. The corollary follows from the fact that $W(s,x) = J(s, H(s,x))$ is the composition of two Borel-measurable mappings. □

4. Some comments. (a) The model studied in this paper includes the case of the *monotone follower* problem as formulated in Karatzas [18] and Karatzas and Shreve [21] by letting $k = d$ and $g(\theta) = I$, $0 \leq \theta \leq T$, where I denotes the $d \times d$ unit matrix. Moreover, if we take $k = 2d$ and $g(\theta) = (I, -I)$, $0 \leq \theta \leq T$, then the model reduces the *bounded variation* problem as discussed in Chow, Menaldi, and Robin [6], among others.

We have assumed that $c^i(\cdot) > 0$. This condition is necessary for the existence of optimal controls to the general problem formulated in this paper (see the proof of Proposition 3.4). Thus it excludes the case of the so-called *cheap control* problems, i.e., $c(\cdot) = 0$. This type of problem is discussed in Chiarolla and Haussmann [4], [5] and Menaldi and Robin [25]. However, our method works for problems with *finite fuel* constraints, because for any $y \in \mathcal{D}^k[0,T]$, the set

$$\{v \in \mathcal{A}^k[0,T] : v^i(t) \leq y^i(t),\ \forall t,\ 1 \leq i \leq k\}$$

is closed in $\mathcal{A}^k[0,T]$. This follows from the fact that $v_n \to v$ in $\mathcal{A}^k[0,T]$, then $v^i_n(t) \to v^i(t)$ $(1 \leq i \leq k)$ at all the continuity points of v^i. For the problems with finite fuel constraints, see Karatzas [20], Karatzas and El Karoui [8], and a more recent work by Bridge and Shreve [1] among others.

(b) Now we explain why the method used in this paper is not suitable to the problem where terminal costs are allowed. If we define $\Gamma_s(\omega) = \phi(x_T(\omega))$ for $\omega = (x, \mu, v)$ with ϕ a continuous function defined in $\mathbb{R}^d$, we cannot get the lower semicontinuity for Γ_s. The reason is that $x^n \to x$ in the pseudopath topology only ensures that $x^n(\cdot) \to x(\cdot)$ in Lebesgue measure, and, therefore, $x^n(T) \to x(T)$ may not hold.

But we can modify the formulation of the problem to allow a terminal cost, i.e., let

$$J(\alpha) \equiv E^P \left\{ \int_s^T f(t, x_t, u_t)dt + \int_{[s,T]} c(t) \cdot dv_t + \phi(x_{T+}) \right\}.$$

Note the additional term $c(T) \cdot \Delta v_T$ in the second integral and the relation $x_{T+} = x_T + g(T)\Delta v_T$. We can replace $\mathcal{D}^d[0,T]$ by $\mathcal{D}^d[0,T] \times \mathbb{R}^d$ and $\mathcal{A}^d[0,T]$ by $\mathcal{A}^d[0,T] \times \mathbb{R}^d$ in the canonical space to obtain the results if ϕ is lower semicontinuous.

(c) For the same reason as explained in (b) we cannot allow pointwise constraints of the type $\Phi(x(t_0)) = 0$ almost surely for some lower-semicontinuous function Φ on $I\!R^d$ (which may take the values $\pm\infty$) and fixed $t_0 \in [0,T]$. But it is seen easily that the following kind of integral constraint may be added to the problem:

$$\int_s^T f_0(t, x_t, u_t)dt + \int_s^T c_0(t) \cdot dv_t \leq 0 \text{ a.s.},$$

where f_0, c_0 satisfy the same conditions as f and c, except that the positivity of $c_0(\cdot)$ is relaxed to *bounded below*. Moreover, from the proof of Lemma 3.5 we can conclude that the following kind of constraint may also be added to the model:

$$\int_s^T f_1(t, x_t, u_t)dt + \int_s^T c_1(t) \cdot dv_t = 0$$

with $f_1(t, \cdot, \cdot)$, $c_1(\cdot)$ continuous on Σ (for each $0 \leq t \leq T$), $[0,T]$, respectively. Of course, we must now assume the existence of an admissible control. See Haussmann and Lepeltier [13] for constraints of these types in the classical control problems.

REFERENCES

[1] D. S. Bridge and S. E. Shreve, *Multi-dimensional finite-fuel singular stochastic control*, preprint.

[2] V. E. Beněs, L. A. Shepp, and H. S. Witsenhausen, *Some solvable stochastic control problems*, Stochastics Stochastics Rep., 4 (1980), pp. 39–83.

[3] P. Billingsley, *Convergence of Probability Measures*, John Wiley, New York, 1968.

[4] M. Chiarolla and U. G. Haussmann, *The free boundary of the monotone follower*, SIAM J. Control Optim., 32 (1994), pp. 690–727.

[5] ———, *The optimal control of the cheap monotone follower*, Stochastic Stochastic Rep., 49 (1994), pp. 99–128.

[6] P. L. Chow, J. L. Menaldi, and M. Robin, *Additive control of stochastic linear system with finite time horizon*, SIAM J. Control Optim., 23 (1985), pp. 858–899.

[7] N. El Karoui, Huu Nguyen, and M. Jeanblanc-Picqué, *Compactification methods in the control of degenerate diffusions: Existence of an optimal control*, Stochastics Stochastics Rep., 20 (1987), pp. 169–219.

[8] N. El Karoui and I. Karatzas, *Probabilistic aspects of finite-fuel, reflected follower problem*, Acta Appl. Math., 11 (1988), pp. 223–258.

[9] W. H. Fleming and R. W. Rishel, *Deterministic and Stochastic Optimal Control*, Springer-Verlag, New York, 1975.

[10] W. H. Fleming and H. M. Soner, *Controlled Markov Processes and Viscosity Solutions*, Springer-Verlag, New York, 1993.

[11] J. M. Harrison and M. I. Taksar, *Instantaneous control of Brownian motion*, Math. Oper. Res., 8 (1983), pp. 439–453.

[12] U. G. Haussmann, *Existence of optimal Markovian controls for degenerate diffusions*, Lecture Notes in Control and Inform. Sci., 78 (1986), pp. 171–186.

[13] U. G. Haussmann and J. P. Lepeltier, *On the existence of optimal control*, SIAM J. Control Optim., 28 (1990), pp. 851–902.

[14] U. G. Haussmann and W. Suo, *Singular optimal stochastic controls II: Dynamic programming*, SIAM J. Control Optim., 33 (1995), pp. 937–959.

[15] ———, *Existence of singular optimal control laws for stochastic differential equations*, Stochastics Stochastics Rep., 48 (1994), pp. 249–272.

[16] N. Ikeda and S. Watanabe, *Stochastic Differential Equations and Diffusion Processes*, North-Holland, Amsterdam, 1981.

[17] J. Jacod and J. Mémin, *Sur un type de convergence intermédiaire entre la convergence en loi et la convergence en probabilité*, in Séminaire de Probabilité XV, Lecture Notes in Mathematics 850, Spinger-Verlag, 1981, pp. 529–540.

[18] I. Karatzas, *The monotone follower problem in stochastic decision theory*, Appl. Math. Optim., 7(1981), pp. 175–189.

[19] I. KARATZAS, *A class of singular control problems*, Adv. in Appl. Probab., 15 (1983), pp. 225–254.

[20] ———, *Stochastic control under finite fuel constraints*, in The IMA Volumes in Math. and Its Appl., Vol. 10, W. H. Fleming and P. L. Lions, eds., Springer-Verlag, New York, 1988, pp. 225–240.

[21] I. KARATZAS AND S. E. SHREVE, *Connections between optimal stopping and stochastic control I: Monotone follower problems*, SIAM J. Control Optim., 22 (1984), pp. 856–877; *Connections between optimal stopping and stochastic control II: Reflected follower problems*, SIAM J. Contol Optim., 22 (1985), pp. 433–451.

[22] R. LARSEN, *Functional Analysis, An Introduction*, Marcel Dekker, New York, 1973.

[23] J. P. LEHOCZKY AND S. E. SHREVE, *Absolutely continuous and singular stochastic control*, Stochastics Stochastics Rep., 17 (1986), pp. 91–109.

[24] J. MA, *On the principle of smooth fit for a class of singular stochastic control problems for diffusion*, SIAM J. Control Optim., 30 (1992), pp. 975–999.

[25] J. L. MENALDI AND M. ROBIN, *On some cheap control problems for diffusion processes*, Trans. Amer. Math. Soc., 278 (1983), pp. 771–802.

[26] ———, *On singular stochastic control problems for diffusions with jumps*, IEEE Trans. Automat. Control, AC-29(1984), pp. 991–1004.

[27] J. L. MENALDI AND M. I. TAKSAR, *Optimal correction problem of a multidimensional stochastic system*, Automatica J. IFAC, 25 (1989), pp. 223–232.

[28] P. A. MEYER AND W. A. ZHENG, *Tightness criteria for laws of semimartingales*, Ann. Inst. Henri Poincaré Probab. Statist., 20 (1984), pp. 353–372.

[29] S. E. SHREVE, *An introduction to singular stochastic control*, The IMA Volumes in Math. and Its Appl., Vol. 10, W. H. Fleming and P. L. Lions, eds., Springer-Verlag, New York, 1988, pp. 513–528.

[30] H. M. SONER AND S. E. SHREVE, *Regularity of the value function for a two-dimensional singular stochastic control problem*, SIAM J. Control Optim., 27 (1989), pp. 876–907.

[31] D. W. STROOCK AND S. R. S. VARADHAN, *Multidimensional Diffusion Processes*, Springer-Verlag, New York, 1979.

[32] M. SUN, *Singular control problems in bounded intervals*, Stochastics Stochastics Rep., 10 (1983), pp. 103–113.

[33] W. SUO, *Existence of Singular Optimal Stochastic Controls*, Ph.D. thesis, Univ. of British Columbia, Vancouver, British Columbia, 1994.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 33, No. 3, pp. 937–959, May 1995

© 1995 Society for Industrial and Applied Mathematics
013

SINGULAR OPTIMAL STOCHASTIC CONTROLS II: DYNAMIC PROGRAMMING*

ULRICH G. HAUSSMANN† AND WULIN SUO‡

Abstract. The dynamic programming principle for a multidimensional singular stochastic control problem is established in this paper. When assuming Lipschitz continuity on the data, it is shown that the value function is continuous and is the unique viscosity solution of the corresponding Hamilton–Jacobi–Bellman equation.

Key words. singular controls, control rules, value function, dynamic programming principle, Hamilton–Jacobi–Bellman equation, viscosity solution

AMS subject classifications. 49J30, 49A55, 60G44, 93E20

1. Introduction.

In [8] we applied a direct method to study the existence of optimal controls for the stochastic control problem in which the state is governed by the stochastic differential equation

$$x_t = x + \int_s^t b(\theta, x_\theta, u_\theta)d\theta + \int_s^t \sigma(\theta, x_\theta, u_\theta)dB_\theta + \int_s^t g(\theta)dv_\theta$$

on some filtered probability space $(\Omega, \mathcal{F}, \mathcal{F}_t, P)$, where $b(\cdot,\cdot,\cdot)$, $\sigma(\cdot,\cdot,\cdot)$, $g(\cdot)$ are given deterministic functions, $(B_t,\ t \geq 0)$ is a d-dimensional Brownian motion (in fact, $B.$ need not be d-dimensional), x is the initial state at time s, and $u:\ [0,T] \mapsto U$, $v:\ [0,T] \mapsto I\!R^k$, with v nondecreasing componentwise, stand for the controls.

The expected cost has the form

$$J(\alpha) \equiv E^P \left\{ \int_s^T f(t, x_t, u_t)dt + \int_{[s,T)} c(t) \cdot dv_t \right\},$$

where $f(\cdot,\cdot,\cdot):\ [0,T] \times I\!R^d \times U \mapsto I\!R$, $c(\cdot):\ [0,T] \mapsto I\!R^k_+$ are given. We assume that the cost of applying the singular control is positive, i.e., $c^i(\cdot) > 0$, $i = 1, \ldots, k$. For this type of problem, the reader may consult the paper by Haussmann and Suo [8] and the list of references therein.

This paper is a continuation of Haussmann and Suo [8]. As is well known for the classical stochastic control problem, the dynamic programming principle is satisfied and, if the value function has appropriate regularity, it satisfies a second-order nonlinear partial differential equation (the Hamilton–Jacobi–Bellman equation) (cf. Fleming and Rishel [4] and Lions [10], among others). This is still the case for singular stochastic control where the Hamilton–Jacobi–Bellman equation is a second-order variational inequality (see Fleming and Soner [5] and the list of references in Haussmann and Suo [8]). In this paper, in §3 we adopt a probabilistic approach used in Haussmann [6], Haussmann and Lepeltier [7], and El Karoui, Nguyen, and Jeanblanc-Picqué [3] to establish the dynamic programming principle under very mild conditions of the data. Then, in §4, assuming Lipschitz continuity of the coefficients, we prove

* Received by the editors June 15, 1993; accepted for publication (in revised form) January 12, 1993. This work was supported by Natural Sciences and Engineering Research Council of Canada grant 88051.

† Department of Mathematics, University of British Columbia, Vancouver, British Columbia, Canada V6T 1Z2.

‡ Faculty of Management, University of Toronto, Toronto, Ontario, Canada M5S 1V4.

that the value function is continuous. In §5 the Hamilton–Jacobi–Bellman equation is derived heuristically, and the value function is shown to be the unique viscosity solution of this equation. For the reader's convenience, the main results of Haussmann and Suo [8] are recalled in §2, along with the formulation of the problem.

We list some notation that will be used throughout this paper:

- $I\!R^d, I\!R$ denote the d-dimensional Euclidean space and the real line, respectively. $I\!R_+ = \{x \in I\!R, x \geq 0\}$ and $I\!R^d_+$ is defined similarly. For $x = (x^i), y = (y^i) \in I\!R^d$, $x \cdot y = \sum_{i=1}^d x^i y^i$.
- $T > 0$ is the fixed horizon, and $\Sigma = [0, T] \times I\!R^d$.
- $\mathcal{D}^d[0, T]$ denotes the collection of $I\!R^d$-valued functions defined on $[0, T]$ which are left continuous and have right limits.
- $\mathcal{A}^k[0, T]$ denotes the collection of functions $a : [0, T] \mapsto I\!R^k_+$ such that $a = (a^i) \in \mathcal{D}^k[0, T]$ and a^i is nondecreasing with $a^i(0) = 0$, $i = 1, \ldots, k$.
- $\mathcal{S}^{l\times k}$ is the space of $l \times k$ matrices with the $l \times k$-dimensional Euclidean norm.
- If Y is a metric space, $\mathcal{B}(Y)$ denotes the corresponding Borel σ-field, and $f \in \mathcal{B}(Y)$ means that f is a $\mathcal{B}(Y)$-measurable real-valued function. We denote by $I\!M_1(Y)$ ($I\!M_+(Y)$, respectively) the space of probabilities (nonnegative Radon measures, respectively) on Y with the weak convergence topology.
- U, called the control set, is a compact metric space. It is well known that $I\!M_1(U)$ is also a compact metrizable space. If $\phi : U \mapsto I\!R$ is a bounded measurable function, we can extend ϕ to $I\!M_1(U)$ by letting
$$\phi(\mu) \equiv \int_U \phi(u)\mu(du).$$
Define
$$\mathcal{U} \equiv \{\mu : [0, T] \mapsto I\!M_1(U) \text{ is Borel measurable}\}.$$
- If X is a random variable on a probability space $(\Omega, \mathcal{F}, P)$, the expectations of X will be denoted by $E^P(X)$. $\mathcal{M}^c_2$ is the family of continuous square integrable martingales on some given probability space $(\Omega, \mathcal{F}, P)$ with a filtration $\{\mathcal{F}_t\}$.
- C stands for a constant, but not necessarily the same one from line to line.

2. Formulation of the problem. We consider the following optimal control problem in which we allow both classical control and singular control to act at the same time, i.e., the dynamics are in the form,

$$x_t = x + \int_s^t b(\theta, x_\theta, \mu_\theta)d\theta + \int_s^t \sigma(\theta, x_\theta, \mu_\theta)dB_\theta + \int_s^t g(\theta)dv_\theta \text{ a.s.} \tag{2.1}$$

for $(t, x) \in \Sigma$, $s \leq t \leq T$, where

- $(\sigma, b) : \Sigma \times U \mapsto \mathcal{S}^{d\times d} \times I\!R^d$, $g : [0, T] \mapsto \mathcal{S}^{d\times k}$,
- $(B_t, 0 \leq t \leq T)$ is a d-dimensional Brownian motion on some probability space, and
- $v \in \mathcal{A}^k[0, T]$.

We assume that σ, b are bounded measurable functions, continuous with respect to (x, u), and g is continuous on $[0, T]$.

DEFINITION 2.1. *A relaxed control is a term* $\alpha = (\Omega, \mathcal{F}, P, \mathcal{F}_t, B_t, x_t, \mu_t, v_t, s, x)$ *such that*

(1) *$(s,x)\in\Sigma$;*
(2) *$(\Omega,\mathcal{F},P)$ is a probability space with the filtration $\{\mathcal{F}_t\}_{t\ge 0}$;*
(3) *$\mu.\in\mathcal{U}$ is progressively measurable with respect to $\mathcal{F}_t$;*
(4) *v is an $\mathbb{R}^k_+$-valued process progressively measurable with respect to $\mathcal{F}_t$; the sample paths of v are in $\mathcal{A}^k[0,T]$, i.e., for each $\omega\in\Omega$, $v.(\omega)\in\mathcal{A}^k[0,T]$;*
(5) *B_t is a standard d-dimensional Brownian motion on $(\Omega,\mathcal{F},P,\mathcal{F}_t)$ and x_t, the state process, is $\mathcal{F}_t$-adapted with sample paths in $\mathcal{D}^d[0,T]$ such that (2.1) is satisfied. We assume that $x_r=x$ for $0\le r\le s$.*

We call (s,x) the initial condition of the relaxed control α.

The collection of relaxed controls with initial condition (s,x) is denoted by $\tilde{\Lambda}_{s,x}$. For $\alpha\in\tilde{\Lambda}_{s,x}$, the associated cost is defined as follows:

$$J(\alpha)\equiv E^P\left\{\int_s^T f(t,x_t,\mu_t)dt+\int_{[s,T)}c(t)\cdot dv_t\right\},\tag{2.2}$$

where

- $f:\Sigma\times U\mapsto\mathbb{R}$ is a measurable function and is lower semicontinuous in (x,u), satisfying
$$-K\le f(t,x,u)\le C(1+\|x\|^m),\quad (t,x,u)\in\Sigma\times U$$
for some constants $m\ge 0$, $K\ge 0$ and $C\ge 0$;
- $c=(c^i):[0,T]\mapsto\mathbb{R}^k$ is lower semicontinuous and $c^i>0$, $1\le i\le k$.

The value function is defined by

$$W(s,x)=\inf_{\alpha\in\tilde{\Lambda}_{s,x}}J(\alpha).\tag{2.3}$$

A relaxed control $\alpha^*\in\tilde{\Lambda}_{s,x}$ is called an optimal relaxed control if $W(s,x)=J(\alpha^*)$.

Throughout this paper we write

$$\int_s^t k(\theta)\cdot da(\theta)=\sum_{i=1}^d\int_{[s,t)}k_i(\theta)da_i(\theta)$$

for any $\mathbb{R}^k$-valued Borel-measurable function $k=(k_i)$ and $a=(a_i)\in\mathcal{A}^k[0,T]$.

Now we introduce the canonical space for our problem. Define

$$\Omega\equiv\mathcal{D}^d[0,T]\times\mathcal{U}\times\mathcal{A}^k[0,T],$$

where $\mathcal{D}^d[0,T]$, $\mathcal{U}$ and $\mathcal{A}^k[0,T]$ are endowed with the pseudopath topology, stable topology, and weak convergence topology, respectively. Ω is metrizable and separable under the product topology; see Haussmann and Suo [8] for details. Let $\tilde{\mathcal{D}},\tilde{\mathcal{U}},\tilde{\mathcal{A}}$ denote their Borel σ-fields, $\tilde{\mathcal{D}}_t,\tilde{\mathcal{U}}_t,\tilde{\mathcal{A}}_t$, the σ-fields up to time t, and define

$$\mathcal{F}\equiv\tilde{\mathcal{D}}\times\tilde{\mathcal{U}}\times\tilde{\mathcal{A}},$$
$$\mathcal{F}_t\equiv\tilde{\mathcal{D}}_t\times\tilde{\mathcal{U}}_t\times\tilde{\mathcal{A}}_t.$$

Let $a=\sigma\sigma^*$, where $*$ denotes the transpose of a matrix, and define

$$\mathcal{L}\equiv\frac{1}{2}\sum_{i,j}a_{ij}\frac{\partial^2}{\partial x_i\partial x_j}+\sum_i b_i\frac{\partial}{\partial x_i}.$$

We recall the following definition from Haussmann and Suo [8]

DEFINITION 2.2. *A control rule is a probability R on the measurable space $(\Omega, \mathcal{F})$ such that*

$$\tilde{\alpha} = (\Omega, \mathcal{F}, \mathcal{F}_t, R, x_t, \mu_t, v_t, s, x) \in \tilde{\Lambda}_{s,x},$$

where

$$x_t(\omega) = x_t, \ \mu_t(\omega) = \mu_t, \ v_t = v_t(\omega)$$

for $\omega = (x_., \mu_., v_.) \in \Omega$, i.e.,

(1) *$(\Omega, \mathcal{F}, R)$ is a probability space with the filtration $\mathcal{F}_t$, $(s, x) \in \Sigma$, and*

$$R(x_r = x, \ \mu_r = \delta^0, \ v_r = 0, \ 0 \le r \le s) = 1,$$

where δ^0 denotes the Dirac measure at an arbitray but fixed point $u^0 \in U$;

(2) *$\forall \phi \in C_b^2(I\!R^d)$, $M_t\phi$ $(s \le t \le T)$ is in $\mathcal{M}_2^c$ on the filtered probability space $(\Omega, \mathcal{F}, \mathcal{F}_t, R)$ $(s \le t \le T)$, where*

$$M_t\phi(\omega) \equiv \phi(x_t) - \int_s^t \mathcal{L}\phi(\theta, x_\theta, \mu_\theta)d\theta - \int_s^t \nabla_x\phi(x_\theta) \cdot g(\theta)dv_\theta \\ - \sum_{s \le \theta < t} \left[\phi(x_{\theta+}) - \phi(x_\theta) - \nabla_x\phi(x_\theta) \cdot \triangle x_\theta\right]. \tag{2.4}$$

We write $J(\alpha) = J(s, R)$.

Define

$$\Gamma_s(\omega) \equiv \int_s^T f(\theta, x_\theta, \mu_\theta)d\theta + \int_s^T c(\theta) \cdot dv_\theta \tag{2.5}$$

for $\omega = (x_., \mu_., v_.)$. Then, by definition, $J(\alpha) = E^R\Gamma_s$. We denote by $\mathcal{R}_{s,x}$ the space of control rules with initial condition (s, x). We will suppress s in $J(s, R)$ when it is clear that $R \in \mathcal{R}_{s,x}$. It is shown in Haussmann and Suo [8] that for any relaxed control $\alpha \in \tilde{\Lambda}_{s.x}$, there exists a control rule $R \in \mathcal{R}_{s,x}$ such that $J(R) = J(\alpha)$. Thus the control problem can be described in terms of control rules.

We recall the main results from Haussmann and Suo [8].

THEOREM 2.3 (Haussmann and Suo [8]). (a) *For each $(s, x) \in \Sigma$ there exists an optimal control rule $P^* \in \mathcal{R}_{s,x}$, i.e., $W(s, x) = J(s, P^*)$.*

(b) *The value function $W(\cdot, \cdot)$ is Borel measurable on Σ.*

(c) *There exists a Borel-measurable map $H(\cdot, \cdot) : \Sigma \mapsto I\!M_1(\Omega)$ such that for each $(s, x) \in \Sigma$, $H(s, x) \in \mathcal{R}_{s,x}^o$, where*

$$\mathcal{R}_{s,x}^o = \{P \in \mathcal{R}_{s,x}, \ J(s, P) = W(s, x)\}.$$

We call H a *measurable selector* of $\mathcal{R}^o$.

3. The dynamic programming principle. In this section we will apply the method used in Haussmann [6], Haussmann and Lepeltier [7], and El Karoui, Nguyen, and Jeanblanc-Picqué [3] to establish the dynamic programming principle. Note that this method does not require any regularity of the value function W.

3.1. Some preparations. We denote by $\mathcal{F}^s$ the σ-algebra of events that occur after time s, i.e., $\mathcal{F}^s = \sigma\{\omega_t, s \le t \le T\}$.

Define $\Theta_{s,\bar{\omega}} : \Omega \mapsto \Omega$ by

$$\Theta_{s,\bar{\omega}}(\omega)_t = \begin{cases} \bar{\omega}_t, & 0 \le t \le s, \\ (x_t, \mu_t, \bar{v}_s + v_t - v_s), & s < t \le T. \end{cases} \tag{3.1}$$

Note that if $\tilde{\omega} = \Theta_{s,\bar{\omega}}(\omega)$, then $\tilde{\omega} = \bar{\omega}$ on $[0,s]$, $(x.(\tilde{\omega}), \mu.(\tilde{\omega})) = (x.(\omega), \mu.(\omega))$ on $(s,T]$, and $v.(\tilde{\omega}) - v_s(\tilde{\omega}) = v.(\omega) - v_s(\omega)$ on $[s,T]$.

LEMMA 3.1. *If P is a probability on $(\Omega, \mathcal{F}^s)$ $(0 \le s \le T)$ and $\bar{\omega} \in \Omega$, then there exists a unique probability measure, denoted by $\delta_{\bar{\omega}} \otimes_s P$ on $(\Omega, \mathcal{F})$ such that*

$$\delta_{\bar{\omega}} \otimes_s P(\omega : \omega_t = \bar{\omega}_t, 0 \le t \le s) = 1, \tag{3.2}$$

$$\delta_{\bar{\omega}} \otimes_s P(A) = P(\Theta^{-1}_{s,\bar{\omega}}(A)) \quad \forall\, A \in \mathcal{F}^s, \tag{3.3}$$

where by $\omega_t = \bar{\omega}_t$, $0 \le t \le s$, we mean $x_t = \bar{x}_t$, $v_t = \bar{v}_t$, $0 \le t \le s$, and $\mu_t = \bar{\mu}_t$ a.e. on $[0,s]$.

Proof. The uniqueness of such a probability measure is obvious, so we only need to show its existence.

If I is a subinterval of $[0,T]$ we write $V(I)$ for the set of measurable functions

$$I \mapsto \mathbb{M}_1(U).$$

Let us recall an equivalent definition of the stable topology on $\mathcal{U} = V([0,T])$. For $\mu \in \mathcal{U}$ define a mapping $i : \mathcal{U} \mapsto C([0,T], \mathbb{M}_+(U))$ by

$$i(\mu)(\cdot) = \int_0^{\cdot} \mu_\theta \, d\theta.$$

The topology on $\mathcal{U}$ induced by i is exactly the stable topology we introduced in Haussmann and Suo [8]. For a discussion of this see Haussmann and Lepeltier [7, §3.10]. Similarly, we can consider the topology on $V(I)$ induced by the mapping

$$i_I(\mu)(\cdot) = \int_0^{\cdot} \mathbf{1}_I(\theta)\mu_\theta d\theta \in C([0,T], \mathbb{M}_+(U)), \quad \mu \in V(I).$$

Write $\mathcal{A}^k_I$ for the $\mathbb{R}^k_+$-valued nondecreasing functions on I with the inherited topology from $\mathcal{A}^k[0,T]$. Let

$$\begin{aligned} X_0 &\equiv \mathcal{D}^d[0,s] \times i_{[0,s]}V([0,s]) \times \mathcal{A}^k_{[0,s]}, \\ X &\equiv \mathcal{D}^d[s,T] \times i_{[s,T]}V([s,T]) \times \mathcal{A}^k_{[s,T]}, \\ \tilde{X} &\equiv X_0 \times X, \end{aligned}$$

and define $\Phi_0 : \Omega \mapsto X_0$, $\Phi : \Omega \mapsto X$ by

$$\Phi_0(\omega)_t = \left(x_t, \int_0^t \mu_\theta d\theta, v_t\right), \quad 0 \le t \le s,$$

$$\Phi(\omega)_{t'} = \left(x_{t'}, \int_s^{t'} \mu_\theta d\theta, v_{t'}\right), \quad s \le t' \le T,$$

where $\omega = (x, \mu, v) \in \Omega$. Define $\Psi : \tilde{X} \mapsto \Omega$ by

$$\Psi(\omega_1, \omega_2) = \Theta_{s,\tau_0(\omega_1)}(\tau(\omega_2)),$$

where $\tau_0 :\ X_0 \mapsto \Omega,\ \tau :\ X \mapsto \Omega$:

$$\begin{aligned}\tau_0(\tilde{\omega})_t &= (\tilde{x}_{t\wedge s}, i^{-1}_{[0,s]}(\tilde{\mu})(t \wedge s), \tilde{v}_{t\wedge s}),\\ \tau(\omega)_t &= (x_{t\vee s}, i^{-1}_{[s,t]}(\mu)(t \vee s), v_{t\vee s})\end{aligned}$$

with $\tilde{\omega} = (\tilde{x}, \tilde{\mu}, \tilde{v}) \in X_0$, $\omega = (x, \mu, v) \in X$.

We define a probability $\tilde{P}$ on $\tilde{X}$ by $\tilde{P} = \delta_{\bar{\omega}} \circ \Phi_0^{-1} \times P \circ \Phi^{-1}$, and let

$$\bar{P} = \tilde{P} \circ \Psi^{-1}.$$

Now we verify that $\bar{P}$ satisfies conditions (3.2) and (3.3). Note that $\tau_0 \circ \Phi_0(\bar{\omega})_t = \bar{\omega}_t,\ 0 \le t \le s$, so we have

$$\begin{aligned}\bar{P}(\omega :\ \omega_t = \bar{\omega}_t,\ 0 \le t \le s) &= \tilde{P}((\omega_1, \omega_2) :\ \Theta_{s,\tau_0(\omega_1)}(\tau(\omega_2))_t = \bar{\omega}_t,\ \ 0 \le t \le s)\\ &= \tilde{P}((\omega_1, \omega_2) :\ \tau_0(\omega_1)_t = \bar{\omega}_t,\ \ 0 \le t \le s)\\ &= \delta_{\bar{\omega}}(\omega_1 :\ \tau_0 \circ \Phi_0(\omega_1)_t = \bar{\omega}_t,\ 0 \le t \le t)\\ &= 1.\end{aligned}$$

For $A \in \mathcal{F}^s$,

$$\begin{aligned}\bar{P}(A) &= \tilde{P}((\omega_1, \omega_2) :\ \Psi(\omega_1, \omega_2) \in A)\\ &= \tilde{P}((\omega_1, \omega_2) :\ \Theta_{s,\tau_0(\omega_1)}(\tau(\omega_2)) \in A)\\ &= \tilde{P}((\omega_1, \omega_2) :\ \tau(\omega_2) \in \Theta^{-1}_{s,\tau_0(\omega_1)}(A))\\ &= \int_{X_0} P \circ \Phi^{-1}(\omega_2 :\ \tau(\omega_2) \in \Theta^{-1}_{s,\tau_0(\omega_1)}(A))\, \delta_{\bar{\omega}} \circ \Phi_0^{-1}(d\omega_1)\\ &= \int_{X_0} P(\omega :\ \tau \circ \Phi(\omega) \in \Theta^{-1}_{s,\tau_0(\omega_1)}(A))\, \delta_{\Phi_0(\bar{\omega})}(d\omega_1)\\ &= P(\omega :\ \tau \circ \Phi(\omega) \in \Theta^{-1}_{s,\tau_0(\Phi_0(\bar{\omega}))}(A))\\ &= P(\Theta^{-1}_{s,\bar{\omega}}(A)),\end{aligned}$$

since $\tau \circ \Phi(\omega)_t = \omega_t$ holds on $[s, T]$. The lemma is proved by letting $\delta_{\bar{\omega}} \otimes_s P = \bar{P}$. □

Remark 3.2. Note that for $P \in \mathcal{R}_{s,x}$, there exists a P-null set N_0 such that if $\bar{\omega} \notin N_0$, $A \in \mathcal{F}^s$, then $P(A) = P \circ \Theta^{-1}_{s,\bar{\omega}}(A)$ and

$$E^P\left\{\int_s^t h(\theta) \cdot dv_\theta\right\} = E^{P\circ\Theta^{-1}_{s,\bar{\omega}}}\left\{\int_s^t h(\theta) \cdot dv_\theta\right\},\quad t \ge s$$

for any bounded $\mathbb{R}^k$-valued Borel-measurable function $h(\cdot)$. These properties will be used repeatedly in the rest of this section.

Assume that τ is an $\mathcal{F}_t$-stopping time, $0 \le \tau \le T$. A τ-transition probability (τ-t.p.) is a family $\{Q_\omega :\ \omega \in \Omega\}$ of probability measures on $(\Omega, \mathcal{F})$ such that

$$\omega \mapsto Q_\omega(A) \text{ is } \mathcal{F}_\tau \text{ measurable } \forall\, A \in \mathcal{F}.$$

Note that $\mathcal{F}_\tau$ is the collection of sets A such that $A\bigcap\{\tau \le t\} \in \mathcal{F}_t\ \forall\, t \le T$. For a fixed τ and ω such that $\tau(\omega) \le T$, we denote by $\mathcal{F}^{\tau(\omega)}$ the σ-field generated by the sets of the form

$$\left\{\tilde{\omega} \in \Omega :\ \tilde{x}_t \in A,\ \int_{\tau(\omega)}^t \tilde{\mu}_\theta d\theta \in B,\ \tilde{v}_t \in C\right\}, \tag{3.4}$$

where $A \in \mathcal{B}(I\!\!R^d)$, $B \in \mathcal{B}(I\!\!M_+(U))$, $C \in \mathcal{B}(I\!\!R^k)$, $\tau(\omega) \le t \le T$, so $\mathcal{F}^{\tau(\omega)}$ is the collection of events that occur after the time $\tau(\omega)$. Note that the topology on Ω is separable, so $\mathcal{F}_t$ is countably generated, and for a given probability P on $(\Omega, \mathcal{F})$, the regular conditional probability distribution (r.c.p.d.) of P for given $\mathcal{F}_\tau$ exists and will be denoted by $P_{\tau,\omega}$.

Given a stopping time τ $(0 \le \tau \le T)$ and a τ-t.p. Q_ω, if $\bar{\omega} \in \Omega$, then from Lemma 3.1 we know that there exists a unique $\delta_{\bar{\omega}} \otimes_\tau Q_\omega \in I\!\!M_1(\Omega)$ such that

$$\begin{aligned}&\delta_{\bar{\omega}} \otimes_\tau Q_\omega\{\tilde{\omega} : \tilde{\omega}_t = \bar{\omega}_t, 0 \le t \le \tau(\omega)\} = 1,\\ &\delta_{\bar{\omega}} \otimes_\tau Q_\omega(A) = Q_\omega(\Theta^{-1}_{\tau_\omega,\bar{\omega}}(A)) \quad \forall\, A \in \mathcal{F}^{\tau(\omega)}.\end{aligned}$$

We write $\delta_\omega \otimes_\tau Q_\omega = Q^\tau_\omega$. When $\bar{\omega} = (x(\tau(\omega)), \delta^0, 0)$, we write

$$\delta_{\bar{\omega}} \otimes_\tau Q_\omega = \overset{\circ}{Q}{}^\tau_\omega\ .$$

It can be seen easily that for $s \le T$ and a stopping time $\tau : s \le \tau \le T$,

$$Q_\omega(\Gamma_{\tau(\omega)}) = Q^\tau_\omega(\Gamma_{\tau(\omega)}) = \overset{\circ}{Q}{}^\tau_\omega\ (\Gamma_{\tau(\omega)}), \tag{3.5}$$

where Γ is defined by (2.5).

If $P \in I\!\!M_1(\Omega)$, τ is a stopping time, and $\{Q_\omega\}$ is a τ-t.p., then we have the following result, which is analogous to Theorem 6.1.2 in Stroock and Varadhan [11].

LEMMA 3.3. *There exists a unique probability, denoted by $P \otimes_\tau Q$, such that*

(1) $P \otimes_\tau Q(A) = P(A)$ if $A \in \mathcal{F}_\tau$,

(2) the r.c.p.d. of $P \otimes_\tau Q$ with respect to $\mathcal{F}_\tau$ is Q^τ_ω.

Proof. The proof of [11, Thm. 6.1.2] can be adopted here with minor modifications (cf. Suo [12]). □

LEMMA 3.4. *Assume $P \in \mathcal{R}_{s,x}$, τ is an $\mathcal{F}_t$-stopping time: $s \le \tau \le T$, and $\bar{\omega} \in \Omega$. Then $(M_t\phi, \mathcal{F}_t, P \circ \Theta^{-1}_{\tau_{\bar{\omega}},\bar{\omega}})$ is a martingale after $\tau_{\bar{\omega}}$ for $\phi \in C^2_b(I\!\!R^d)$.*

Proof. Since $P \in \mathcal{R}_{s,x}$, we know that $(M_t\phi, \mathcal{F}_t, P)$ is a martingale, i.e.,

$$\int_A M_t\phi dP = \int_A M_u\phi dP \ \ \forall\, A \in \mathcal{F}_u, \ \ s \le u < t \le T, \tag{3.6}$$

or

$$E^P\left\{1_A(\cdot)\left[M_t\phi(\cdot) - M_u\phi(\cdot)\right]\right\} = 0. \tag{3.7}$$

When $t \ge \tau_{\bar{\omega}}$, we have

$$\begin{aligned} M_t\phi(\Theta_{\tau_{\bar{\omega}},\bar{\omega}}(\omega)) = \phi(x_t) &- \int_{\tau_{\bar{\omega}}}^t \mathcal{L}\phi(\theta, x_\theta, \mu_\theta)d\theta - \int_{\tau_{\bar{\omega}}}^t \nabla_x\phi(x_\theta)\cdot g(\theta)dv^c_\theta\\ &- \sum_{\tau_{\bar{\omega}}<\theta<t} [\phi(x_{\theta+}) - \phi(x_\theta)] - [\phi(x_{\tau+}) - \phi(\bar{x}_\tau)]\\ &- \int_s^{\tau_{\bar{\omega}}} \mathcal{L}\phi(\theta, \bar{x}_\theta, \bar{\mu}_\theta)d\theta - \int_s^{\tau_{\bar{\omega}}} \nabla_x\phi(\bar{x}_\theta)\cdot g(\theta)d\bar{v}^c_\theta\\ &- \sum_{s\le\theta<\tau_{\bar{\omega}}} [\phi(\bar{x}_{\theta+}) - \phi(\bar{x}_\theta)]\,. \end{aligned}$$

Therefore we can get for $\tau_{\bar{\omega}} \le u \le t \le T$, $A \in \mathcal{F}_u$,

$$\begin{aligned}
&E^{P\circ\Theta^{-1}_{\tau_{\bar{\omega}},\bar{\omega}}} \{1_A(\cdot)\,[M_t\phi(\cdot) - M_u\phi(\cdot)]\} \\
&= E^P\{1_A(\Theta_{\tau_{\bar{\omega}},\bar{\omega}}(\cdot))\,[M_t\phi(\Theta_{\tau_{\bar{\omega}},\bar{\omega}}(\cdot)) - M_u\phi(\Theta_{\tau_{\bar{\omega}},\bar{\omega}}(\cdot))]\} \\
&= E^P\{1_A(\Theta_{\tau_{\bar{\omega}},\bar{\omega}}(\cdot))\,[M_t\phi(\cdot) - M_u\phi(\cdot)]\} \\
&= 0,
\end{aligned}$$

because it is obvious that $\Theta^{-1}_{\tau_{\bar{\omega}},\bar{\omega}}(A) \in \mathcal{F}_u$. The lemma is thus proved. □

Recall that for a stopping time τ, $P_{\tau,\omega}$ denotes the regular conditional probability distribution of P given $\mathcal{F}_\tau$. $P_{\tau,\omega}$ is obviously a τ-t.p., so $\overset{\circ}{P}{}^{\tau}_{\omega} \equiv \delta_{\bar{\omega}} \otimes_\tau P_{\tau,\omega}$ is defined, where $\bar{\omega} = (x(\tau(\omega)), \delta^0, 0)$.

LEMMA 3.5. *If $(M_t, \mathcal{F}_t, P)$ is a martingale and τ is an $\mathcal{F}_t$-stopping time then there is a P-null set $N \in \mathcal{F}_\tau$ such that for $\omega \notin N$, $(M_t, \mathcal{F}_t, P_{\tau,\omega})$ is a martingale for $t \ge \tau_\omega$.*

Proof. The proof is comparable to that of Stroock and Varadhan [11, Thm. 1.2.10.] □

COROLLARY 3.6. *If $P \in \mathcal{R}_{s,x}$ and τ is an $\mathcal{F}_t$-stopping time, then there is a P-null set $N \in \mathcal{F}_\tau$ such that for $\omega \notin N$, $(M_t\phi, \mathcal{F}_t, P_{\tau,\omega} \circ \Theta^{-1}_{\tau_\omega,\omega})$ is a martingale for $t \ge \tau_\omega$ when $\phi \in C_b^2(I\!R^d)$.*

Proof. The proof is obvious from Lemmas 3.4 and 3.5. □

The next two results are important for the rest of the paper. The first one states that a control rule remains a control rule for the problem starting at a later time from the point reached at that time. The second one says that if we take a control rule and at some later time switch to another control rule, then this concatenated object is still a control rule.

PROPOSITION 3.7 (closure under conditioning). *If $P \in \mathcal{R}_{s,x}$ and τ is a stopping time, $s \le \tau \le T$, then there exists a P-null set $N \in \mathcal{F}_\tau$ such that $\overset{\circ}{P}{}^{\tau}_{\omega} \in \mathcal{R}_{\tau_\omega, x_{\tau_\omega}}$ for $\omega \notin N$.*

Proof. Let $\{\phi_m\}$ be a dense subset of $C_0^\infty(I\!R^d)$. Then for each m, $(M_t\phi_m - M_{t\wedge\tau}\phi_m, \mathcal{F}_t, P)$ $(t \ge s)$ is a martingale. For $\bar{\omega} \in \Omega$ define

$$\begin{aligned}
\bar{M}_t^{\bar{\omega}}\phi_m = \phi(x_t) - \phi(x_{t\wedge\tau_{\bar{\omega}}}) &- \int_{t\wedge\tau_{\bar{\omega}}}^t \mathcal{L}\phi_m(\theta, x_\theta, \mu_\theta)d\theta \\
&- \int_{t\wedge\tau_{\bar{\omega}}}^t \nabla_x\phi_m(x_\theta)\cdot g(\theta)dv_\theta^c - \sum_{t\wedge\tau_{\bar{\omega}} \le \theta < t} [\phi_m(x_{\theta+}) - \phi(x_\theta)].
\end{aligned}$$

Then, by Lemma 3.5 there exists a P-null set $N_m \in \mathcal{F}_\tau$ such that $(\bar{M}_t^{\bar{\omega}}\phi_m, \mathcal{F}_t, P_{\tau,\bar{\omega}})$ is a martingale for $\bar{\omega} \notin N_m$. By Lemma 3.4 we know that $(\bar{M}_t^{\bar{\omega}}\phi_m, \mathcal{F}_t, P_{\tau,\bar{\omega}} \circ \Theta^{-1}_{\tau,\bar{\omega}})$ is a martingale for $t \ge \tau_{\bar{\omega}}$, $\bar{\omega} \notin N_m$. Certainly, $(\bar{M}^{\bar{\omega}}_{t\wedge\tau_{\bar{\omega}}}\phi_m, \mathcal{F}_t, \delta_{\bar{\omega}'})$ is a martingale, where $\bar{\omega}' = (\bar{x}_{\tau_{\bar{\omega}}}, \delta^0, 0)$. Therefore $(\bar{M}_{t\wedge\tau_{\bar{\omega}}}\phi_m, \mathcal{F}_t, \overset{\circ}{P}{}^{\tau}_{\omega})$ is a martingale. It is obvious from the definition that

$$\overset{\circ}{P}{}^{\tau}_{\bar{\omega}}\,(\omega :\ x_r = \bar{x}_{\tau_{\bar{\omega}}},\ \mu_r = \delta^0,\ v_r = 0,\ 0 \le r \le \tau_{\bar{\omega}}) = 1.$$

Let $N = \bigcup_m N_m$; then $P(N) = 0$. Through a limit procedure we can show that $\overset{\circ}{P}{}^{\tau}_{\omega} \in \mathcal{R}_{\tau_\omega, x_{\tau_\omega}}$ for $\omega \notin N$. □

PROPOSITION 3.8 (closure under concatenation). *Let $P \in \mathcal{R}_{s,x}$ and τ be a stopping time such that $s \leq \tau \leq T$. If Q_ω is a transition probability such that $Q_\omega \in \mathcal{R}_{\tau(\omega),x(\tau(\omega))}$, then*

$$P \otimes_\tau Q \in \mathcal{R}_{s,x}.$$

Proof. It is obvious that we need only to show that for each $\phi \in C_b^2(\mathbb{R}^d)$, $(M_t\phi, \mathcal{F}_t, P \otimes_\tau Q)$ is a martingale after time s. From Remark 3.2 and the fact that $Q_\omega \in \mathcal{R}_{\tau(\omega),x(\tau(\omega))}$, it can be easily verified that $(M_t\phi, \mathcal{F}_t, \delta_\omega \otimes_\tau Q_\omega)$ is a martingale after time τ. By definition we know that $Q_\omega^\tau = \delta_\omega \otimes_\tau Q_\omega$ equals the regular conditional probability distribution of $P \otimes_\tau Q$ given $\mathcal{F}_\tau$. The proof of the proposition now follows Theorem 1.2.10 in Stroock and Varadhan [11]. □

Set $Q_\omega = H(\tau(\omega), x(\tau(\omega)))$, where H is a measurable selector of $\mathcal{R}^o$ and, by definition, it is a τ-t.p. We denote

$$P \otimes_\tau H = P \otimes_\tau Q_\omega.$$

COROLLARY 3.9. *For P in $\mathcal{R}_{s,x}$ we have $P \otimes_\tau H \in \mathcal{R}_{s,x}$.*

3.2. The dynamic programming principle. With the preparations of §3.1, we can now establish the dynamic programming principle.

For a given probability $P \in \mathcal{R}_{s,x}$, define $\mathcal{R}_{s,x}^\tau(P)$ to be the set of probabilities in $\mathcal{R}_{s,x}$ which coincide with P up to time τ, i.e.,

$$\mathcal{R}_{s,x}^\tau(P) = \{P \otimes_\tau Q : \quad Q_\omega \in \mathcal{R}_{\tau_\omega, x_{\tau_\omega}} \text{ such that } Q_\omega \text{ is a } \tau - \text{t.p.}\} \subset \mathcal{R}_{s,x}.$$

We introduce the following notation: for a measurable function $\phi \in \mathcal{B}(\Sigma)$, let

$$\Gamma_s(t, \phi)(\omega) = \int_s^t f(\theta, x_\theta, \mu_\theta) d\theta + \int_s^t c(\theta) \cdot dv(\theta) + \phi(t, x_t),$$

where $\omega = (x, \mu, v) \in \Omega$.

THEOREM 3.10 (dynamic programming principle). (a) *If τ is an $\mathcal{F}_t$-stopping time, $s \leq \tau \leq T$, and $P \in \mathcal{R}_{s,x}$, then*

$$\begin{aligned} E^P \left\{ \int_s^\tau f(\theta, x_\theta, \mu_\theta) d\theta + \int_s^\tau c(\theta) \cdot dv_\theta + W(\tau, x_\tau) \right\} \\ = \inf \left\{ \hat{P}(\Gamma_s) : \hat{P} \in \mathcal{R}_{s,x}^\tau(P) \right\}, \end{aligned} \tag{3.8}$$

where $\Gamma_s(\cdot)$ is defined by (2.5).

(b) *For $P \in \mathcal{R}_{s,x}$, $(\Gamma_s(t, W), \mathcal{F}_t, P)$ is a submartingale.*

(c) *If $s \leq \tau \leq T$, then*

$$W(s, x) = \inf \{P\Gamma_s(\tau, W), P \in \mathcal{R}_{s,x}\}. \tag{3.9}$$

(d) *$(\Gamma_s(t, W), \mathcal{F}_t, P)$ is a martingale under P if and only if $P \in \mathcal{R}_{s,x}$ is optimal.*

Proof. (a) Recall that H denotes a measurable selector of $\mathcal{R}^o$; therefore, by (3.3) and Remark 3.2 the left-hand side (LHS) of (3.8) is

$$\begin{aligned} &E^P \left\{ \int_s^\tau f(\theta, x_\theta, \mu_\theta) d\theta + \int_s^\tau c(\theta) \cdot dv_\theta + H(\tau(\cdot), x_\tau(\cdot))\Gamma_\tau \right\} \\ &= E^P \left\{ \int_s^\tau f(\theta, x_\theta, \mu_\theta) d\theta + \int_s^\tau c(\theta) \cdot dv_\theta + \delta_. \otimes_\tau H(\Gamma_\tau) \right\} \\ &= P \otimes_\tau H(\Gamma_s), \end{aligned}$$

and by Proposition 3.8 we know that $P \otimes_\tau H \in \mathcal{R}^\tau_{s,x}(P)$. Hence, in (3.8) LHS $\geq$ RHS. On the other hand, for $\hat{P} = P \otimes_\tau Q \in \mathcal{R}^\tau_{s,x}(P)$,

$$\begin{aligned}
\hat{P}(\Gamma_s) &= P \otimes_\tau Q(\Gamma_s) \\
&= E^P \left\{ \int_s^\tau f(\theta, x_\theta, \mu_\theta) d\theta + \int_s^\tau c(\theta) \cdot d\nu_\theta + Q_\cdot(\Gamma_\tau) \right\} \\
&\geq E^P \left\{ \int_s^\tau f(\theta, x_\theta, \mu_\theta) d\theta + \int_s^\tau c(\theta) \cdot d\nu_\theta + W(\tau(\cdot), x_\tau(\cdot)) \right\} \\
&= \text{LHS of (3.8)},
\end{aligned}$$

so the proof of (a) is completed.

(b) For any $s \leq t < t + h$, we have

$$\begin{aligned}
&E^P\{\Gamma_s(t+h, W) - \Gamma_s(t, W) | \mathcal{F}_t\} \\
&= E^P \left\{ \int_s^{t+h} f(\theta, x_\theta, \mu_\theta) d\theta + \int_s^{t+h} c(\theta) \cdot d\nu_\theta + W(t+h, x_{t+h}) \right. \\
&\qquad \left. - \left[\int_s^t f(\theta, x_\theta, \mu_\theta) d\theta + \int_s^t c(\theta) \cdot d\nu_\theta + W(t, x_t) \right] \middle| \mathcal{F}_t \right\} \\
&= E^P \left\{ \int_t^{t+h} f(\theta, x_\theta, \mu_\theta) d\theta + \int_t^{t+h} c(\theta) \cdot d\nu_\theta + W(t+h, x_{t+h}) \middle| \mathcal{F}_t \right\} - W(t, x_t) \\
&= P_{t,\omega} \Gamma_t(t+h, W) - W(t, x_t) \\
&= \overset{\circ}{P}{}^t_\omega \, \Gamma_t(t+h, W) - W(t, x_t)
\end{aligned}$$

by Lemma 3.3(2). Now we apply (3.8) and the fact that $\mathcal{R}^{t+h}_{t,x_t}(\overset{\circ}{P}{}^t_\omega) \subset \mathcal{R}_{t,x_t}$, i.e., Proposition 3.7, to get

$$\begin{aligned}
&E^P\{\Gamma_s(t+h, W) - \Gamma_s(t, W) | \mathcal{F}_t\} \\
&= \inf \left\{ \hat{P}\Gamma_t : \hat{P} \in \mathcal{R}^{t+h}_{t,x_t}(\overset{\circ}{P}{}^t_\omega) \right\} - W(t, x_t) \\
&\geq \inf \left\{ \hat{P}\Gamma_t : \hat{P} \in \mathcal{R}_{t,x_t} \right\} - W(t, x_t) \\
&= 0.
\end{aligned}$$

Therefore $(\Gamma_s(t, W), \mathcal{F}_t, P)$ is a submartingale.

(c) Let $\nu(d\theta dx)$ be the distribution of (τ, x_τ) under P. Then

$$\begin{aligned}
E^P W(\tau, x_\tau) &= \int W(\theta, x) \nu(d\theta dx) \\
&= \int J(\theta, H(\theta, x)) \nu(d\theta dx) \\
&= E^P J(\tau, H(\tau, x_\tau)) \\
&= J(\tau, P \otimes_\tau H).
\end{aligned}$$

Note that

$$E^P \Gamma_s(\tau, W) = E^P \left\{ \int_s^\tau f(\theta, x_\theta, \mu_\theta) d\theta + \int_s^\tau c(\theta) \cdot d\nu_\theta \right\} + E^P W(\tau, x_\tau)$$

$$\begin{aligned}
&= E^P\left\{\int_s^\tau f(\theta,x_\theta,\mu_\theta)d\theta + \int_s^\tau c(\theta)\cdot dv_\theta\right\} + J(\tau, P\otimes_\tau H)\\
&= J(s, P\otimes_\tau H)\\
&\geq \inf\left\{J(s,Q), Q \in \mathcal{R}_{s,x}\right\}\\
&= W(s,x).
\end{aligned}$$

On the other hand, $E^P W(\tau, x_\tau) \leq E^P J(\tau, \overset{\circ}{P^\tau_\omega}) = E^P J(\tau, P^\tau_\omega) = J(\tau, P)$ and, therefore, by (b),

$$\begin{aligned}
(3.10)\qquad W(s,x) &\leq E^P\Gamma_s(\tau, W)\\
&= E^P\left\{\int_s^\tau f(\theta,x_\theta,\mu_\theta)d\theta + \int_s^\tau c(\theta)\cdot dv_\theta\right\} + E^P W(\tau, x_\tau)\\
&\leq E^P\left\{\int_s^\tau f(\theta,x_\theta,\mu_\theta)d\theta + \int_s^\tau c(\theta)\cdot dv_\theta\right\} + J(\tau, P)\\
&= J(s,P),
\end{aligned}$$

and thus (c) is proved if we take the infimum over $P \in \mathcal{R}_{s,x}$ on the right-hand side (RHS).

(d) If $\Gamma_s(t, W)$ is a P-martingale, then

$$W(s,x) = E^P\Gamma_s(s,W) = E^P\Gamma_s(T,W) = E^P\Gamma_s(T,0) = E^P\Gamma_s = J(s,P),$$

because from our assumptions, $W(T,\cdot) = 0$. So P is optimal.

If we assume that $P \in \mathcal{R}_{s,x}$ is optimal, then by (3.10), Proposition 3.7, and Corollary 3.6,

$$(3.11)\qquad W(s,x) \leq E^P\Gamma_s(t,W) = E^P\Gamma_s = W(s,x).$$

Therefore $(\Gamma_s(t,W), \mathcal{F}_t, P)$ is a submartingale with constant mean value, so it is indeed a martingale. □

4. Continuity of the value function.

In the rest of the paper we add the following assumptions:

- $c(\cdot)$ is Lipschitz continuous, $f(\cdot,\cdot,\cdot)$ is bounded, and $g = (g^{ij})$ is a constant $d \times k$-matrix;
- $f(\cdot,\cdot,\cdot)$, $b(\cdot,\cdot,\cdot)$, $\sigma(\cdot,\cdot,\cdot)$ satisfy the following conditions:

$$(4.1)\qquad \begin{aligned}
|f(t,x,u) - f(s,y,u)| &\leq C(|t-s| + \|x-y\|),\\
\|b(t,x,u) - b(s,y,u)\| &\leq C(|t-s| + \|x-y\|),\\
\|\sigma(t,x,u) - \sigma(s,y,u)\| &\leq C(|t-s| + \|x-y\|)
\end{aligned}$$

uniformly for $0 \leq s,t \leq T$, x, $y \in I\!R^d$, $u \in U$.

We will prove that under these conditions the value function $W(\cdot,\cdot)$ is uniformly continuous on Σ. In fact, there exists a constant $C \geq 0$ such that

$$|W(t,x) - W(s,y)| \leq C\left(|t-s|^{\frac{1}{2}} + \|x-y\|\right),\ 0 \leq s,t \leq T,\ x,y \in I\!R^d.$$

Note that the constancy of g is only required in the proof of Theorem 4.2.

THEOREM 4.1. *The value function $W(s,x)$ is uniformly Lipschitz continuous in the state variable x, i.e., there exists a constant $C > 0$ such that*

$$(4.2)\qquad |W(s,x') - W(s,x)| \leq C\|x' - x\|\ \forall 0 \leq t \leq T,\ x,\ x' \in I\!R^d.$$

Proof. In the following, we use the same notation C to denote the constants, which may change from time to time. For any $0 \le s \le T$, x, $x' \in I\!R^d$,

$$W(s,x') - W(s,x) \le \sup_{P \in \mathcal{R}_{s,x}} (E^Q \Gamma_s - E^P \Gamma_s)$$

for each $Q \in \mathcal{R}_{s,x'}$, where Γ is the cost function defined by (2.5).

Take an arbitrary $P \in \mathcal{R}_{s,x}$. By the definition of control rules, there exists a standard extension $(\tilde{\Omega}, \tilde{\mathcal{F}}, \tilde{\mathcal{F}}_t, \tilde{P})$ of $(\Omega, \mathcal{F}, \mathcal{F}_t, P)$, i.e., there exists another probability space $(\Omega', \mathcal{F}', \mathcal{F}'_t, P')$ such that $\tilde{\Omega} = \Omega \times \Omega'$, $\tilde{\mathcal{F}} = \mathcal{F} \times \mathcal{F}'$, $\tilde{\mathcal{F}}_t = \mathcal{F}_t \times \mathcal{F}'_t$, and $\tilde{P} = P \times P'$. We can extend the processes $x_\cdot, \mu_\cdot, v_\cdot$ to $\tilde{\Omega}$ by the following: for $\tilde{\omega} = (\omega, \omega') \in \tilde{\Omega}$,

$$x_\cdot(\tilde{\omega}) = x_\cdot(\omega),\ \mu_\cdot(\tilde{\omega}) = \mu_\cdot(\omega),\ v_\cdot(\tilde{\omega}) = v_\cdot(\omega).$$

On $(\tilde{\Omega}, \tilde{\mathcal{F}}, \tilde{\mathcal{F}}_t, \tilde{P})$ there exists a standard d-dimensional Brownian motion $B_\cdot$ such that for $s \le t \le T$,

$$x_t = x + \int_s^t b(\theta, x_\theta, \mu_\theta) d\theta + \int_s^t \sigma(\theta, x_\theta, \mu_\theta) dB_\theta + g v_t \quad \text{a.s.} \tag{4.3}$$

Consider the same equation (4.3) with the initial state x', i.e.,

$$y_t = x' + \int_s^t b(\theta, y_\theta, \mu_\theta) d\theta + \int_s^t \sigma(\theta, y_\theta, \mu_\theta) dB_\theta + g v_t \tag{4.4}$$

on the stochastic basis $(\tilde{\Omega}, \tilde{\mathcal{F}}, \tilde{\mathcal{F}}_t, \tilde{P})$. The strong solution for (4.4) exists from the assumptions on $b(\cdot,\cdot,\cdot)$ and $\sigma(\cdot,\cdot,\cdot)$, and so $\alpha \equiv (\tilde{\Omega}, \tilde{\mathcal{F}}, \tilde{\mathcal{F}}_t, \tilde{P}, y_t, \mu_t, v_t, s, x') \in \tilde{\Lambda}_{s,x'}$. Therefore, there exists a control rule $Q \in \mathcal{R}_{s,x'}$ such that

$$J(\alpha) = J(s,Q) \doteq E^Q \Gamma_s. \tag{4.5}$$

By definition,

$$\begin{aligned} E^P \Gamma_s &= E^P \left\{ \int_s^T f(\theta, x_\theta, \mu_\theta) d\theta + \int_s^T c(\theta) \cdot dv_\theta \right\} \\ &= E^{\tilde{P}} \left\{ \int_s^T f(\theta, x_\theta, \mu_\theta) d\theta + \int_s^T c(\theta) \cdot dv_\theta \right\}, \end{aligned}$$

and, by (4.5),

$$E^Q \Gamma_s = E^{\tilde{P}} \left\{ \int_s^T f(\theta, y_\theta, \mu_\theta) d\theta + \int_s^T c(\theta) \cdot dv_\theta \right\};$$

therefore, from the Lipschitz continuity of f we have

$$\begin{aligned} E^Q \Gamma_s - E^P \Gamma_s &\le E^{\tilde{P}} \left\{ \int_s^T |f(\theta, y_\theta, \mu_\theta) - f(\theta, x_\theta, \mu_\theta)| d\theta \right\} \\ &\le C E^{\tilde{P}} \int_s^T \|y_\theta - x_\theta\| d\theta \\ &\le C \left(\int_s^T E^{\tilde{P}} \|y_\theta - x_\theta\|^2 d\theta \right)^{\frac{1}{2}}. \end{aligned}$$

Now, from equations (4.3), (4.4), and the Lipschitz continuity conditions on b, σ, we have for $\theta \geq s$,

$$\begin{aligned} E^{\tilde{P}}\|y_\theta - x_\theta\|^2 &\leq C\|x' - x\|^2 + CE^{\tilde{P}}\left(\int_s^\theta \|b(h, y_h, \mu_h) - b(h, x_h, \mu_h)\|\,dh\right)^2 \\ &\quad + CE^{\tilde{P}}\sup_{\theta' \leq \theta}\left\|\int_s^{\theta'} (\sigma(h, y_h, \mu_h) - \sigma(h, x_h, \mu_h))\,dB_h\right\|^2 \\ &\leq C\|x' - x\|^2 + CE^{\tilde{P}}\int_s^\theta \|b(h, y_h, \mu_h) - b(h, x_h, \mu_h)\|^2 dh \\ &\quad + CE^{\tilde{P}}\int_s^\theta \|\sigma(h, y_h, \mu_h) - \sigma(h, x_h, \mu_h)\|^2 dh \\ &\leq C\left(\|x' - x\|^2 + \int_s^\theta E^{\tilde{P}}\|y_h - x_h\|^2 dh\right). \end{aligned}$$

We have used the Burkholder–Gundy inequality to get the second inequality. By Gronwall's inequality,

$$E^{\tilde{P}}\|y_\theta - x_\theta\|^2 \leq C\|x' - x\|^2 e^{C(\theta - s)} \leq C\|x' - x\|^2.$$

Hence we have

$$E^Q\Gamma_s - E^P\Gamma_s \leq C\|x' - x\|$$

and, therefore, $W(s, x,) - W(s, x) \leq C\|x' - x\|$.

The proof of the theorem is thus complete since x, $x' \in I\!R^d$ are arbitrary. □

Next we consider the continuity of the value function in the time variable t.

THEOREM 4.2. *The value function $W(t, x)$ is uniformly continuous in the time variable t. In fact, there exists a constant $C > 0$ such that*

$$|W(s, x) - W(s', x)| \leq C\,|s - s'|^{\frac{1}{2}} \tag{4.6}$$

for all $0 \leq s$, $s' \leq T$, $x \in I\!R^d$.

Proof. As in the proof of Theorem 4.1, we use the same notation C to denote the constants. First we assume $s \leq s'$, so that

$$W(s', x) - W(s, x) \leq \sup_{P \in \mathcal{R}_{s,x}} (E^Q\Gamma_{s'} - E^P\Gamma_s) \tag{4.7}$$

for each $Q \in \mathcal{R}_{s',x}$. From the strict positivity of $c(\cdot)$, we may actually take the supremum in (4.7) over a subset $\mathcal{R}_{s,x}(\lambda)$ for some $\lambda > 0$, where

$$\mathcal{R}_{s,x}(\lambda) = \{P \in \mathcal{R}_{s,x} :\ E^P\|v(T)\| \leq \lambda\}. \tag{4.8}$$

Now, for $P \in \mathcal{R}_{s,x}(\lambda)$, as in the proof Theorem 4.1, there exists a standard extension $(\tilde{\Omega}, \tilde{\mathcal{F}}, \tilde{\mathcal{F}}_t, \tilde{P})$ of $(\Omega, \mathcal{F}, \mathcal{F}_t, P)$ such that $x(\cdot)$ is a solution of

$$x_t = x + \int_s^t b(\theta, x_\theta, \mu_\theta)d\theta + \int_s^t \sigma(\theta, x_\theta, \mu_\theta)dB_\theta + gv_t \tag{4.9}$$

on the stochastic basis $(\tilde{\Omega}, \tilde{\mathcal{F}}, \tilde{\mathcal{F}}_t, \tilde{P})$, where $B.$ is a d-dimensional Brownian motion. Define $\hat{\mathcal{F}}_t = \tilde{\mathcal{F}}_{t-s'+s}$, $\hat{\mu}_t = \mu_{t-s'+s}$, $\hat{v}_t = v_{t-s'+s}$, and $\hat{B}_t = B_{t-s'+s}$ for $t \geq s'$. Consider the following stochastic differential equation:

$$y_t = x + \int_{s'}^{t} b(\theta, y_\theta, \hat{\mu}_\theta) d\theta + \int_{s'}^{t} \sigma(\theta, y_\theta, \hat{\mu}_\theta) d\hat{B}_\theta + g\hat{v}_t, \quad t \geq s' \tag{4.10}$$

on the stochastic basis $(\tilde{\Omega}, \tilde{\mathcal{F}}, \hat{\mathcal{F}}_t, \tilde{P})$. We know that under assumption (4.1) there exists a unique strong solution $y.$, and by definition, $\alpha = (\tilde{\Omega}, \tilde{\mathcal{F}}, \hat{\mathcal{F}}_t, \tilde{P}, y_t, \hat{\mu}_t, \hat{v}_t, s', x) \in \tilde{\Lambda}_{s'x}$. Therefore, there exists a control rule $Q \in \mathcal{R}_{s',x}$ such that

$$J(\alpha) = J(s', Q) = E^Q \Gamma_{s'}. \tag{4.11}$$

Recall that Γ_s is defined by (2.5). Thus, by definition,

$$E^P \Gamma_s = E^{\tilde{P}} \left\{ \int_s^T f(\theta, x_\theta, \mu_\theta) d\theta + \int_s^T c(\theta) \cdot dv_\theta \right\},$$

and by (4.11),

$$\begin{aligned} E^Q \Gamma_{s'} &= E^{\tilde{P}} \left\{ \int_{s'}^T f(\theta, y_\theta, \hat{\mu}_\theta) d\theta + \int_{s'}^T c(\theta) \cdot d\hat{v}_\theta \right\} \\ &= E^{\tilde{P}} \left\{ \int_s^{T-s'+s} f(\theta + s' - s, y_{\theta+s'-s}, \mu_\theta) d\theta + \int_s^{T-s'+s} c(\theta + s' - s) \cdot dv_\theta \right\}. \end{aligned}$$

Therefore, noticing that f is bounded below by a constant $-K$ ($K \geq 0$) and $P \in \mathcal{R}_{s,x}(\lambda)$, we get

$$\begin{aligned} E^Q \Gamma_{s'} - E^P \Gamma_s \leq E^{\tilde{P}} \Bigg\{ & \int_s^{T-(s'-s)} |f(\theta, x_\theta, \mu_\theta) - f(\theta + s' - s, y_{\theta+s'-s}, \mu_\theta)| d\theta \\ & + \int_s^{T-(s'-s)} \|c(\theta) - c(\theta + s' - s)\| \, d\|v_\theta\| + K|s' - s| \Bigg\} \\ \leq E^{\tilde{P}} \Bigg\{ & \int_s^{T-(s'-s)} |f(\theta, x_\theta, \mu_\theta) - f(\theta + s' - s, y_{\theta+s'-s}, \mu_\theta)| d\theta \Bigg\} \\ & + C|s' - s| \left(E^P \|v(T)\| + 1 \right). \end{aligned}$$

From the Lipschitz continuity of the function f, we have

$$|f(\theta, x_\theta, \mu_\theta) - f(\theta + s' - s, y_{\theta+s'-s}, \mu_\theta)| \leq C(|s' - s| + \|x_\theta - y_{\theta+s'-s}\|).$$

Therefore

$$E^Q \Gamma_{s'} - E^P \Gamma_s \leq C \left(|s' - s| + E^{\tilde{P}} \int_s^{T-(s'-s)} \|x_\theta - y_{\theta+s'-s}\| d\theta \right). \tag{4.12}$$

By (4.10) we have for $\theta \geq s$,

$$y_{\theta+s'-s} = x + \int_{s'}^{\theta+s'-s} b(h, y_h, \hat{\mu}_h) dh \tag{4.13}$$

$$
\begin{aligned}
&+\int_{s'}^{\theta+s'-s} \sigma(h, y_h, \hat{\mu}_h) d\hat{B}_u + g\hat{v}_{\theta+s'-s} \\
&= x + \int_s^\theta b(h+s'-s, y_{h+s'-s}, \mu_h) dh \\
&\qquad + \int_s^\theta \sigma(h+s'-s, y_{h+s'-s}, \mu_h) dB_h + gv_\theta.
\end{aligned}
$$

So from (4.9), (4.13), and the Burkholder–Davis–Gundy inequality, we have for $\theta \geq s$,

$$
\begin{aligned}
&E^{\tilde{P}} \|x_\theta - y_{\theta+s'-s}\|^2 \\
&\leq 2E\left(\int_s^\theta \|b(h, x_h, \mu_h) - b(h+s'-s, y_{h+s'-s}, \mu_h)\| dh\right)^2 \\
&\qquad + 2E^{\tilde{P}} \sup_{\theta' \leq \theta} \left\|\int_s^{\theta'} (\sigma(h, x_h, \mu_h) - \sigma(h+s'-s, y_{h+s'-s}, \mu_h)) dB_h\right\|^2 \\
&\leq 2TE^{\tilde{P}} \int_s^\theta \|b(h, x_h, \mu_h) - b(h+s'-s, y_{h+s'-s}, \mu_h)\|^2 dh \\
&\qquad + 2E^{\tilde{P}} \int_s^\theta \|\sigma(h, x_h, \mu_h) - \sigma(h+s'-s, y_{h+s'-s}, \mu_h)\|^2 dh \\
&\leq CE^{\tilde{P}} \int_s^\theta (|s'-s|^2 + \|x_h - y_{h+s'-s}\|^2) dh \\
&\leq C\left(|s'-s|^2 + \int_s^\theta E^{\tilde{P}} \|x_h - y_{h+s'-s}\|^2 dh\right).
\end{aligned}
$$

Gronwall's inequality implies

$$
E^{\tilde{P}} \|x_\theta - y_{\theta+s'-s}\|^2 \leq C|s'-s|^2 e^{C(\theta-s)} \leq C|s'-s|^2.
$$

Hence, from (4.12) we have

$$
E^Q \Gamma_{s'} - E^P \Gamma_s \leq C|s'-s|
$$

and, therefore, $W(s', x) - W(s, x) \leq C\,|s'-s|$ for $s' > s$.

Now we assume $s' < s$. By the dynamic programming principle (cf. Theorem 3.10),

$$
W(s', x) = \inf_{P \in \mathcal{R}_{s',x}} E^P \left\{\int_{s'}^s f(\theta, x_\theta, \mu_\theta) d\theta + \int_{s'}^s c(\theta) \cdot dv_\theta + W(s, x_s)\right\}.
$$

Take $P^0 \in \mathcal{R}_{s',x}$ such that $P^0(\mu_\theta = \delta_{\{u^0\}}, v_\theta = 0,\ 0 \leq \theta \leq T) = 1$ for some arbitrary but fixed $u^0 \in U$. Then

$$
\begin{aligned}
W(s', x) &\leq E^{P^0} \left\{\int_{s'}^s f(\theta, x_\theta, u^0) d\theta + W(s, x_s)\right\} \\
&\leq C|s'-s| + E^{P^0} W(s, x_s)
\end{aligned}
$$

by the boundedness of f, and by Theorem 4.1 there exists a constant C such that

$$
W(s, x_s) \leq W(s, x) + C\|x_s - x\|.
$$

Hence

$$W(s',x) - W(s,x) \le C(|s'-s| + E^{P^0}\|x_s - x\|) \le C\left(|s'-s| + (E^{P^0}\|x_s - x\|^2)^{\frac{1}{2}}\right). \tag{4.14}$$

From the definition of control rules, we know that under P^0,

$$x_s - x = \int_{s'}^{s} b(\theta, x_\theta, u^0)d\theta + M_s, \tag{4.15}$$

where M is a continuous square integrable martingale with

$$\langle M\rangle_s = \int_{s'}^{s} \operatorname{tr}(a(\theta, x_\theta, u^0))d\theta.$$

Therefore, by (4.15) and the Burkholder–Davis–Gundy inequality, we have

$$E^{P^0}\|x_s - x\|^2 \le C\left(|s'-s|^2 + |s'-s|\right). \tag{4.16}$$

Combining (4.14) and (4.16) we have

$$W(s',x) - W(s,x) \le C\,|s'-s|^{\frac{1}{2}}.$$

The theorem is thus proved. □

Remark 4.3. We have assumed that the function f is bounded. It is easy to see from the proof of Theorem 4.2 that without this condition the constant C will depend on x.

Combining Theorems 4.1 and 4.2 we can state the main result of this section.

THEOREM 4.4. *The value function W is uniformly continuous on Σ. Moreover, there exists a constant $C \ge 0$ such that*

$$|W(t,x) - W(s,y)| \le C\left(|t-s|^{\frac{1}{2}} + \|x-y\|\right), \quad 0 \le s,t \le T, \; x,y \in I\!R^d.$$

5. The dynamic programming equation. Before we derive the dynamic programming equation heuristically, we prove a result which shows that there exists a set such that the optimal state process is continuous when it is in this set.

THEOREM 5.1. (a) *Assume $(t,x) \in \Sigma$; then*

$$W(t,x) \le W(t, x+gh) + c(t)\cdot h \tag{5.1}$$

for each $h \in I\!R^k_+$. Moreover, if equality holds for some $h = (h^i) \in I\!R^k$, then the same equality holds when we replace h by $\bar h$ with $\bar h = (\bar h^i) \in I\!R^k_+$, $\bar h^i \le h^i$ $(1 \le i \le k)$.

(b) *Define, for $0 \le t \le T$,*

$$A_t \equiv \{x : \; W(t,x) < W(t,x+gh) + c(t)\cdot h, \; \forall\, h \in I\!R^k_+, \; h \ne 0\}.$$

Then the optimal state process x_t is continuous when it is in A_t. To be precise, we have

$$P(\Delta x_t \ne 0, x_t \in A_t) = 0, \quad s \le t \le T \tag{5.2}$$

for every $P \in \mathcal{R}^o_{s,x}$, $(s,x) \in \Sigma$.

Proof. (a) If (5.1) fails for some $h \in \mathbb{R}^k_+$, then

$$W(t,x) > W(t,x+gh) + c(t)\cdot h. \tag{5.3}$$

Take $P \in \mathcal{R}^o_{t,x+gh}$. We define $\Theta : \Omega \to \Omega$ by

$$\Theta(\omega)_s = \begin{cases} (x_s - gh, \mu_s, 0), & 0 \le s \le t, \\ (x_s, \mu_s, v_s + h), & t < s \le T \end{cases}$$

for $\omega = (x_\cdot, \mu_\cdot, v_\cdot)$, and let $\tilde{P} = P \circ \Theta^{-1}(\cdot)$. As in the proof of Lemma 3.4, we can show that $\tilde{P} \in \mathcal{R}_{t,x}$. From the definition of $\tilde{P}$ we have

$$\begin{aligned} J(t,\tilde{P}) &= E^{\tilde{P}}\left\{\int_t^T f(\theta, x_\theta, \mu_\theta)d\theta + \int_t^T c(\theta)\cdot dv_\theta\right\} \\ &= E^P\left\{\int_t^T f(\theta, x_\theta, \mu_\theta)d\theta + \int_t^T c(\theta)\cdot dv_\theta + c(t)\cdot h\right\} \\ &= J(t,P) + c(t)\cdot h \\ &= W(t,x+gh) + c(t)\cdot h. \end{aligned}$$

Therefore, $J(t,\tilde{P}) < W(t,x)$ from (5.3), a contradiction.

Next, if (5.1) holds as an equality for h, then for $\bar{h}$, $\bar{h}^i \le h^i$, $1 \le i \le k$,

$$\begin{aligned} W(t,x) - c(t)\cdot h &= W(t,x+gh) \\ &\ge W(t,x+g\bar{h}) - c(t)\cdot(h-\bar{h}) \\ &\ge W(t,x) - c(t)\cdot\bar{h} - c(t)\cdot(h-\bar{h}) \\ &= W(t,x) - c(t)\cdot h. \end{aligned}$$

Therefore

$$W(t,x) = W(t,x+g\bar{h}) + c(t)\cdot\bar{h}.$$

(b) For $P \in \mathcal{R}_{s,x}$, we know that P-a.s.,

$$\begin{aligned} x_t = x + \int_s^t b(\theta, x_\theta, \mu_\theta)d\theta + gv_t \\ + \text{ a continuous local martingale.} \end{aligned}$$

Thus $\triangle x_t = g\triangle v_t$, and $x_{t+} = x_t + \triangle x_t = x_t + g\triangle v_t$. Since $W(\cdot,\cdot)$ is continuous on Σ, for $t \le s \le T$,

$$W(t,x_{t+}) = \lim_{t'\downarrow\downarrow t} W(t',x_{t'}). \tag{5.4}$$

Assume $P \in \mathcal{R}^o_{s,x}$, then by the dynamic programming principle (cf. Theorem 3.10) we know that $(\Gamma_s(t,W), \mathcal{F}_t, P)$ is a martingale. Hence, for $t' > t$,

$$W(s,x) = E^P\left\{\int_s^{t'} f(\theta, x_\theta, \mu_\theta)d\theta + \int_s^{t'} c(\theta)\cdot dv_\theta + W(t',x_{t'})\right\}. \tag{5.5}$$

Let $t' \downarrow\downarrow t$; note that from our assumption we know that $c(\cdot)$ is continuous on $[0,T]$. Therefore

$$\int_s^{t'} f(\theta, x_\theta, \mu_\theta)d\theta \to \int_s^t f(\theta, x_\theta, \mu_\theta)d\theta,$$
$$\int_s^{t'} c(\theta)\cdot dv_\theta \to \int_s^t c(\theta)\cdot dv_\theta + c(t)\cdot \Delta v_t.$$

Hence if (5.2) fails, then (5.4) and (5.5) imply

$$\begin{aligned} W(s,x) &= E^P\left\{\int_s^t f(\theta, x_\theta, \mu_\theta)d\theta + \int_s^t c(\theta)\cdot dv_\theta + c(t)\cdot\Delta v_t + W(t, x_t + g\Delta v_t)\right\} \\ &> E^P\left\{\int_s^t f(\theta, x_\theta, \mu_\theta)d\theta + \int_s^t c(\theta)\cdot dv_\theta + W(t,x_t)\right\} \\ &= E^P\Gamma_s(t,W), \end{aligned}$$

which contradicts the fact that $P \in \mathcal{R}^o_{s,x}$. The inequality follows from the facts that

$$W(t,x_t) \le W(t, x_t + g\Delta x_t) + c(t)\cdot\Delta v_t$$

and the strict inequality holds if $x_t \in A_t$ and $\Delta x_t \neq 0$. □

Remark 5.2. From Theorem 5.1 (a), we can see that if the value function $W \in C^{1,2}(\Sigma)$, then

$$(g^*\nabla_x W(t,x))^i + c^i(t) \ge 0, \quad i = 1, 2, \ldots, k,$$

where $*$ means transpose and $(\cdot)^i$ denotes the ith coordinate of the point in $\mathbb{R}^k$. For $x \notin A_t$, there exists $h = (h^i) \in \mathbb{R}^k_+$ such that for $\bar{h} = (\bar{h}^i) \in \mathbb{R}^k_+$, $\bar{h}^i \le h^i$, $1 \le i \le k$,

$$W(t,x) = W(t, x + g\bar{h}) + c(t)\cdot\bar{h}.$$

Therefore we have

$$(g^*\nabla_x W(t,x))^i + c^i(t) = 0$$

for those i such that $h^i > 0$.

5.1. Heuristic derivation of the dynamic programming equation. Recall Ito's formula. For $\phi \in C^{1,2}(\Sigma)$, $(s,x) \in \Sigma$, $t \ge s$,

$$\begin{aligned} \phi(t,x_t) = \phi(s,x) &+ \int_s^t \left(\frac{\partial}{\partial t} + \mathcal{L}\right)\phi(\theta, x_\theta, \mu_\theta)d\theta \\ &+ \int_s^t \nabla_x\phi(\theta,x_\theta)\cdot a(\theta, x_\theta, \mu_\theta)dB_\theta + \int_s^t \nabla_x\phi(\theta, x_\theta)\cdot g dv_\theta \\ &+ \sum_{s\le\theta<t}\left[\phi(\theta, x_{\theta+}) - \phi(\theta, x_\theta) - \nabla_x\phi(\theta,x_\theta)\cdot\Delta x_\theta\right]. \end{aligned} \tag{5.6}$$

Let

$$\tilde{\mathcal{L}} \equiv \frac{\partial}{\partial t} + \mathcal{L} = \frac{\partial}{\partial t} + \frac{1}{2}\sum_{i,j} a_{ij}\frac{\partial^2}{\partial x^i \partial x^j} + \sum_i b_i \frac{\partial}{\partial x^i};$$

then for $P \in \mathcal{R}_{s,x}$, (5.6) may be written as

$$E^P\phi(t,x_t) = \phi(s,x) + E^P\left\{\int_s^t \tilde{\mathcal{L}}\phi(\theta,x_\theta,\mu_\theta)d\theta + \int_s^t \nabla_x\phi(\theta,x_\theta)\cdot g dv_\theta \right.$$
$$\left. + \sum_{s\le\theta<t} [\phi(\theta,x_{\theta+}) - \phi(\theta,x_\theta) - \nabla_x\phi(\theta,x_\theta)\cdot\triangle x_\theta]\right\}.$$

By the dynamic programming principle (cf. Theorem 3.10),

$$W(s,x) = \inf_{P\in\mathcal{R}_{s,x}} E^P\left\{\int_s^t f(\theta,x_\theta,\mu_\theta)d\theta + \int_s^t c(\theta)\cdot dv_\theta + W(t,x_t)\right\},$$

so if we assume $W \in C^{1,2}(\Sigma)$, then

$$\begin{aligned}(5.7)\quad 0 = \inf_{P\in\mathcal{R}_{s,x}} E^P\Bigg\{&\int_s^t (\tilde{\mathcal{L}}W + f)(\theta,x_\theta,\mu_\theta)d\theta \\ &+ \int_s^t \big(c(\theta) + g^*\nabla_x W(\theta,x_\theta)\big)\cdot dv_\theta \\ &+ \sum_{s\le\theta<t}\big[W(\theta,x_{\theta+}) - W(\theta,x_\theta) - \nabla_x W(\theta,x_\theta)\cdot\triangle x_\theta\big]\Bigg\}.\end{aligned}$$

If we take the infimum over all those $P \in \mathcal{R}_{s,x}$ such that $P(v_t = 0,\ 0 \le t \le T) = 1$, then from (5.7) we get

$$\inf_{P\in\mathcal{R}_{s,x}} E^P\left\{\int_s^t (\tilde{\mathcal{L}}W + f)(\theta,x_\theta,\mu_\theta)d\theta\right\} \ge 0.$$

Let $t \downarrow s$; we have

$$\inf_{P\in\mathcal{R}_{s,x}} E^P(\tilde{\mathcal{L}}W + f)(s,x_s,\mu_s) \ge 0.$$

Moreover, from Remark 5.2 we can conclude that on Σ,

$$(5.8)\qquad (g^*\nabla_x W)^i + c^i \ge 0,\quad i = 1,2,\dots,k.$$

Therefore we can expect that W formally satisfies the following *variational inequality*, or *Hamilton–Jacobi–Bellman equation*,

$$(5.9)\quad \min\left\{\inf_{u\in U}(\tilde{\mathcal{L}}W + f)(t,x,u), (g^*\nabla_x W(t,x))^i + c^i(t),\ i = 1,2,\dots,k\right\} = 0$$

on Σ. For simplicity of notation, we write (5.9) as

$$\min\{\inf_{u\in U}(\tilde{\mathcal{L}}W + f),\ g^*\nabla_x W + c\} = 0.$$

5.2. Viscosity solution. As is well known in classical control problems, the value function is a solution to the corresponding Hamilton–Jacobi–Bellman equation when it has sufficient regularity (cf. Fleming and Rishel [4], Krylov [9]). If it is only known that the value function is continuous, then, as observed by Lions [10], the value function is a solution to the Hamilton–Jacobi–Bellman equation in the *viscosity* sense, which we will define as follows.

DEFINITION 5.3. *A function ψ is a* viscosity solution *of* (5.9) *if $\psi \in C(\Sigma)$ and, for every $\phi \in C^{1,2}(\Sigma)$,*

(1) *for each local maximum point* (t_0, x_0) *of* $\psi - \phi$ *in the interior of* Σ, *we have*

$$\min\left\{\inf_{u\in U}(\tilde{\mathcal{L}}\phi + f),\ g^*\nabla_x\phi + c\right\} \geq 0 \tag{5.10}$$

at (t_0, x_0), *i.e.,* ψ *is a subsolution*;

(2) *for each local minimum point* (t_0, x_0) *of* $\psi - \phi$ *in the interior of* Σ, *we have*

$$\min\left\{\inf_{u\in U}(\tilde{\mathcal{L}}\phi + f),\ g^*\nabla_x\phi + c\right\} \leq 0 \tag{5.11}$$

at (t_0, x_0), *i.e.,* ψ *is a supersolution.*

For an introduction to the viscosity solution and its applications to stochastic optimal control problems, see Fleming and Soner [5]. For an extensive bibliography on viscosity solutions, see Crandall, Ishii, and Lions [2].

THEOREM 5.4. *The value function* $W(\cdot,\cdot)$ *is a viscosity solution of* (5.9).

Proof. By Theorem 4.4 we know $W \in C(\Sigma)$. We first show that W is a subsolution. For $\phi \in C^{1,2}(\Sigma)$, if $(t_0, x_0) \in \text{int}(\Sigma)$ is a local maximum point of $W - \phi$, then there is a neighborhood $O^1(t_0, x_0)$ of (t_0, x_0) in Σ such that

$$W(t,x) - \phi(t,x) \leq W(t_0,x_0) - \phi(t_0,x_0),\ \ (t,x) \in \bar{O}^1(t_0,x_0),$$

or

$$W(t,x) - W(t_0,x_0) \leq \phi(t,x) - \phi(t_0,x_0),\ \ (t,x) \in \bar{O}^1(t_0,x_0), \tag{5.12}$$

where $\bar{O}^i(t_0,x_0)$ denotes the closure of $O^i(t_0,x_0)$. If (5.10) fails, then one of the following will be true:

$$\inf_{u\in U}(\tilde{\mathcal{L}}\phi + f)(t_0,x_0,u) < 0, \tag{5.13}$$

$$(g^*\nabla_x\phi(t_0,x_0))^i + c^i(t_0) < 0 \ \text{ for some } 1 \leq i \leq k. \tag{5.14}$$

If (5.13) is true, then from assumption (4.1) and $\phi \in C^{1,2}(\Sigma)$, we can find $u^0 \in U$ and a neighborhood $O^2(t_0,x_0)$ of (t_0,x_0) such that

$$(\tilde{\mathcal{L}}\phi + f)(t,x,u^0) < 0$$

for $(t,x) \in O^2(t_0,x_0)$. Take $P \in \mathcal{R}_{t_0,x_0}$ such that

$$P(\mu_r = \delta_{\{u^0\}},\ v_r = 0,\ 0 \leq r \leq T) = 1. \tag{5.15}$$

The existence of such a $P \in \mathcal{R}_{t_0,x_0}$ is obvious. Define

$$\tau = \inf\{t > t_0,\ (t,x_t) \notin O(t_0,x_0)\},$$

where $O(t_0,x_0) = O^1(t_0,x_0) \cap O^2(t_0,x_0)$. Since the state process $x.$ is continuous a.s.(P), we can see immediately that

$$P(\tau > t_0) = 1,$$

and for $t_0 \leq \theta < \tau$,

$$(\tilde{\mathcal{L}}\phi + f)(\theta, x_\theta, \mu_\theta) < 0 \ \text{ a.s.}(P).$$

Therefore

$$E^P \int_{t_0}^{\tau} (\tilde{\mathcal{L}}\phi + f)(\theta, x_\theta, \mu_\theta) d\theta < 0.$$

By the definition of control rules,

$$\phi(\tau, x_\tau) = \phi(t_0, x_0) + \int_{t_0}^{\tau} \tilde{\mathcal{L}}\phi(\theta, x_\theta, \mu_\theta) d\theta + M_\tau \phi \ \text{ a.s.}(P),$$

where $M\phi \in \mathcal{M}_2^c$, i.e., a continuous square integrable martingale with respect to P. Note that Theorem 4.2.1 of Stroock and Varadhan [11] allows us to replace $\mathcal{L}$ by $\tilde{\mathcal{L}}$ in (2.4), at least when $v = 0$. Hence

$$E^P \phi(\tau, x_\tau) - \phi(t_0, x_0) = E^P \int_{t_0}^{\tau} \tilde{\mathcal{L}}\phi(\theta, x_\theta, \mu_\theta) d\theta.$$

Noting that $(\tau, x_\tau) \in \bar{O}(t_0, x_0)$, we have

$$\begin{aligned} E^P W(\tau, x_\tau) - W(t_0, x_0) &\leq E^P \phi(\tau, x_\tau) - \phi(t_0, x_0) \\ &= E^P \int_{t_0}^{\tau} \tilde{\mathcal{L}}\phi(\theta, x_\theta, \mu_\theta) d\theta \\ &< -E^P \int_{t_0}^{\tau} f(\theta, x_\theta, \mu_\theta) d\theta, \end{aligned}$$

which, by (5.15) can be rewritten as

$$\begin{aligned} W(t_0, x_0) &> E^P \left\{ \int_{t_0}^{\tau} f(\theta, x_\theta, \mu_\theta) d\theta + W(\tau, x_\tau) \right\} \\ &= E^P \left\{ \int_{t_0}^{\tau} f(\theta, x_\theta, \mu_\theta) d\theta + \int_{t_0}^{\tau} c(\theta) \cdot dv_\theta + W(\tau, x_\tau) \right\}. \end{aligned}$$

This contradicts the dynamic programming principle (cf. (3.9)).

Next, if (5.14) holds at (t_0, x_0) for some i, then we can take $h^i > 0$ small enough such that

$$\phi(t_0, x_0 + g^i h^i) - \phi(t_0, x_0) < -c^i(t_0) h^i,$$

where g^i denotes the ith column of the $d \times k$ matrix g. Therefore, by (5.12) we have

$$W(t_0, x + g^i h^i) - W(t_0, x_0) < -c^i(t_0) h^i$$

and, therefore,

$$W(t_0, x_0) > W(t_0, x_0 + gh) + c(t_0) \cdot h,$$

where $h = (0, \ldots, h^i, \ldots, 0)$. This is a contradiction of Theorem 5.1, and thus we have shown that W is a subsolution of (5.9)

Now we show that W is also a supersolution of (5.9). If $\phi \in C^{1,2}(\Sigma)$ such that $W - \phi$ has a local minimum point at $(t_0, x_0) \in \text{int}(\Sigma)$, then there exists a neighborhood $O^1(t_0, x_0)$ of (t_0, x_0) satisfying

$$W(t, x) - W(t_0, x_0) \geq \phi(t, x) - \phi(t_0, x_0), \quad (t, x) \in \bar{O}^1(t_0, x_0). \tag{5.16}$$

If (5.11) fails, then

$$\inf_{u\in U}\left(\tilde{\mathcal{L}}\phi + f\right) > 0,\ \left(g^*\nabla\phi\right)^i + c^i > 0$$

at (t_0, x_0) for $i = 1, \dots, k$. From assumption (4.1) and the fact that $\phi \in C^{1,2}(\Sigma)$, we can find a neighborhood $O^2(t_0, x_0)$ of (t_0, x_0) such that for some $\varepsilon > 0$,

$$\inf_{u\in U}\left(\tilde{\mathcal{L}}\phi + f\right) > \varepsilon,\ \left(g^*\nabla\phi\right)^i + c^i > \varepsilon$$

on $\bar{O}^2(t_0, x_0)$ for $i = 1, \dots, k$. Let $O(t_0, x_0) = O^1(t_0, x_0) \cap O^2(t_0, x_0)$; then for $(t, x) \in O(t_0, x_0)$, we have for small $h \in \mathbb{R}_+^k$, $h \neq 0$,

$$\phi(t, x + gh) - \phi(t, x) > -c(t) \cdot h.$$

Therefore, by (5.16),

$$W(t, x + gh) - W(t, x) > -c(t) \cdot h$$

or $x \in A_t$. Hence for $P \in \mathcal{R}^o_{t_0,x_0}$,

$$P(x_{t_0+} = x_{t_0}) = 1 \tag{5.17}$$

by Theorem 5.1. Define

$$\tau = \inf\{t \geq t_0,\ \ (t, x_t) \notin O(t_0, x_0)\};$$

then from (5.17) we see that $P(\tau > t_0) = 1$ for $P \in \mathcal{R}^o_{t_0,x_0}$, and it can be seen that

$$(t, x_t) \in \bar{O}(t_0, x_0),\ \ x_t \in A_t,\ \ t_0 \leq t \leq \tau.$$

Therefore we have

$$E^P\left\{\int_{t_0}^{\tau} (\tilde{\mathcal{L}}\phi + f)(\theta, x_\theta, \mu_\theta)d\theta + \int_{t_0}^{\tau} \left(g^*\nabla_x\phi(\theta, x_\theta) + c_\theta\right) \cdot dv_\theta\right\} \geq \varepsilon E^P(\tau - t_0).$$

Applying Ito's formula and noting that the state process $x.$ is continuous a.s. (P) when $t_0 \leq t \leq \tau$, we have

$$E^P\phi(\tau, x_\tau) = \phi(t_0, x_0) + E^P\left\{\int_{t_0}^{\tau} \tilde{\mathcal{L}}\phi(\theta, x_\theta, \mu_\theta)d\theta + \int_{t_0}^{\tau} \nabla_x\phi(\theta, x_\theta) \cdot g dv_\theta\right\},$$

which may be rewritten as

$$\begin{aligned} E^P[\phi(\tau, x_\tau) - \phi(t_0, x_0)] \geq E^P &\left\{\int_{t_0}^{\tau} -f(\theta, x_\theta, \mu_\theta)d\theta\right. \\ &\left. - \int_{t_0}^{\tau} c(\theta) \cdot dv_\theta\right\} + \varepsilon E^P(\tau - t_0). \end{aligned}$$

By (5.16) and the fact that $P(\tau > t_0) > 0$, we have

$$E^P[W(\tau, x_\tau) - W(t_0, x_0)] > E^P\left\{\int_{t_0}^{\tau} -f(\theta, x_\theta, \mu_\theta)d\theta - \int_{t_0}^{\tau} c(\theta) \cdot dv_\theta\right\}$$

or

$$W(t_0, x_0) < E^P \left\{ \int_{t_0}^{\tau} f(\theta, x_\theta, \mu_\theta) d\theta + \int_{t_0}^{\tau} c(\theta) \cdot dv_\theta + W(\tau, x_\tau) \right\},$$

which contradicts the dynamic programming principle.

The proof of this theorem is therefore complete. □

Let us define the function space

$$\begin{aligned} \mathcal{C}(\Sigma) \equiv \{W(\cdot,\cdot) : \ & W \in C(\Sigma; I\!R) \text{ with } W \text{ bounded and} \\ & |W(t,x) - W(t,y)| \le C\|x-y\| \text{ for some } C \ge 0\}. \end{aligned}$$

By Theorem 4.4 we know that the value function $W \in \mathcal{C}(\Sigma)$. The proof of the next theorem is a modification of the methods used in Fleming and Soner [5]. For details see Suo [12].

THEOREM 5.5. *There exists a unique viscosity solution in $\mathcal{C}(\Sigma)$ to the dynamic programming equation* (5.9) *with the boundary condition $W(T, x) = 0$, $x \in I\!R^d$, which can be identified as the value function.*

REFERENCES

[1] P. Billingsley, *Convergence of Probability Measures*, John Wiley, New York, 1968.

[2] M. G. Crandall, H. Ishii, and P.-L. Lions, *A user's guide to viscosity solutions*, Bull. Amer. Math. Soc. (N.S.), 27 (1992), pp. 1–67.

[3] N. El Karoui, Huu Nguyen, and M. Jeanblanc-Picqué, *Compactification methods in the control of degenerate diffusions: Existence of an optimal control*, Stochastics Stochastics Rep., 20 (1987), pp. 169–219.

[4] W. H. Fleming and R. W. Rishel, *Deterministic and Stochastic Optimal Control*, Springer-Verlag, New York, 1975.

[5] W. H. Fleming and H. M. Soner, *Controlled Markov Processes and Viscosity Solutions*, Springer-Verlag, New York, 1993.

[6] U. G. Haussmann, *Existence of optimal Markovian controls for degenerate diffusions*, Lecture Notes in Control and Inform. Sci., 78 (1986), pp. 171–186.

[7] U. G. Haussmann and J. P. Lepeltier, *On the existence of optimal control*, SIAM J. Control Optim., 28 (1990), pp. 851–902.

[8] U. G. Haussmann and W. Suo, *Singular optimal stochastic controls I: Existence*, SIAM J. Control Optim., 33 (1995), pp. 916–936.

[9] N. Krylov, *Controlled Diffusion Processes*, Springer-Verlag, New York, 1980.

[10] P. L. Lions, *Optimal control of diffusion processes and Hamilton–Jacobi–Bellman equations. Part 1: The dynamic programming principle and applications, and Part 2: Viscosity solutions and uniqueness*, Comm. Partial Differential Equations, 8 (1983), pp. 1101–1174 and pp. 1229–1276.

[11] D. W. Stroock and S. R. S. Varadhan, *Multidimensional Diffusion Processes*, Springer-Verlag, New York, 1979.

[12] W. Suo, *The Existence of Singular Optimal Controls for Stochastic Differential Equations*, Ph.D. thesis, Univ. of British Columbia, Vancouver, British Columbia, 1994.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 33, No. 3, pp. 960–970, May 1995

© 1995 Society for Industrial and Applied Mathematics
014

AN EXISTENCE RESULT IN A PROBLEM OF THE VECTORIAL CASE OF THE CALCULUS OF VARIATIONS *

ARRIGO CELLINA† AND SANDRO ZAGATTI‡

Abstract. We prove that the problem

$$\text{Minimize} \int_\Omega g(\Phi(\nabla T(x)))dx, \quad T \in T_B + W_0^{1,\infty}(\Omega, \mathbb{R}^n)$$

admits at least one solution for any lower-semicontinuous extended valued function g, for any quasi-affine real-valued function Φ, and for any piecewise-affine boundary datum T_B such that $\Phi(\nabla T_B)$ is constant.

Key words. minimum problem, Jacobian determinant, quasi-affine function

AMS subject classification. 49A50

1. Introduction. In this paper we consider the problem of existence of solutions for the problem

$$\mathcal{P}: \quad \text{Minimize} \int_\Omega g(\Phi(\nabla T(x))dx, \quad T \in T_B + W_0^{1,\infty}(\Omega, \mathbb{R}^n),$$

where Φ is a real-valued quasi-affine function defined on the space $\mathcal{M}_n$ of $n \times n$ matrices, T a transformation from Ω to $\mathbb{R}^n$, $T_B : \bar{\Omega} \mapsto \mathbb{R}^n$ is a piecewise-affine boundary datum and Ω is an open and bounded subset of $\mathbb{R}^n$.

When $\Phi(A) = \det(A)$ (the determinant is the simplest example of nonaffine quasi-affine function), the problem, which arises in the study of equilibrium of gases and constitutes a typical nonconvex problem in the vectorial case of the calculus of variations, has been considered in [D2] and [MS].

In [D2] it is proved that the relaxed problem

$$\text{Minimize} \int_\Omega g^{**}(\det(\nabla u(x))dx, \quad u \in C^\infty(\Omega, \mathbb{R}^3), \quad u = u_0 \text{ on } \partial\Omega$$

admits at least one smooth solution provided that $g : \mathbb{R}^+ \to \mathbb{R}$ satisfies some growth conditions, the boundary datum u_0 is a homeomorphism with positive Jacobian determinant, and Ω is diffeomorphic to the unit sphere.

In [MS] the authors give a proof, based on Moser's Theorem on volume-preserving diffeomorphisms (see [M], [DM]), of existence of a solution for the problem

$$\text{Minimize} \int_\Omega g(\det(\nabla u(x))dx, \quad u \in u_0 + W^{1,\infty}(\Omega, \mathbb{R}^n)$$

*Received by the editors April 21, 1993; accepted for publication (in revised form) January 4, 1994.

†Scuola Internazionale Superiore di Studi Avanzati (SISSA), via Beirut 2–4, I–34014 Trieste, Italy.

‡Dipartimento di Matematica, Università di Bologna, Piazza di Porta S. Donato 5, 40127 Bologna, Italy.

for $g : \mathbb{R}^+ \to \mathbb{R}$ continuous, satisfying the growth condition at 0^+ and at $+\infty$, and for a C^2 homeomorphism u_0 with positive Jacobian determinant in $\bar{\Omega}$.

In this paper we consider an extended-valued lower-semicontinuous function g, defined on $\mathbb{R}$, with superlinear growth at infinity and show that $\mathcal{P}$ admits at least one solution for any quasi-affine function Φ and for any piecewise-affine boundary datum T_B such that $\Phi(\nabla T_B)$ is constant (possibly zero).

We wish to point out that the proof is easier whenever the datum T_B is such that ∇T_B is never a critical point for Φ (in the case of the Jacobian determinant this means that the rank of ∇T_B is larger than or equal to $n-1$); in the general case the result is obtained by proving the existence of a piecewise-affine transformation $T_r \in T_B + W_0^{1,\infty}(\Omega, \mathbb{R}^n)$ such that $\Phi(\nabla T_r) = \Phi(\nabla T_B)$ and $d\Phi(\nabla T_r) \neq 0$ almost everywhere in Ω, and then by solving the equivalent problem

$$\text{Minimize} \int_\Omega g(\Phi(\nabla T(x))dx, \quad T \in T_r + W_0^{1,\infty}(\Omega, \mathbb{R}^n).$$

2. Preliminaries and notation.

In this paper we use the following notation. Vectors (of $\mathbb{R}^n$) are columns. Given $b \in \mathbb{R}^n$, b^t is the transpose of b and $(b)^\perp$ is the orthogonal complement of span(b), $a \cdot b$ is the inner product of a and b vectors of $\mathbb{R}^n$, and $|\cdot|$ is the associated norm. The canonical base in $\mathbb{R}^n$ is denoted by $\{e_i, i = 1, \ldots, n\}$. A subset of $\mathbb{R}^n$ is called n-dimensional if its linear span is the whole space; for a convex polytope P, $V(P)$ is the set of its vertices.

An $n \times n$ matrix A is written as

$$A = (a_1, \ldots, a_n) = \begin{pmatrix} a^1 \\ \vdots \\ a^n \end{pmatrix} = \begin{pmatrix} a_1^1 & \cdots & a_n^1 \\ \vdots & & \vdots \\ a_1^n & \cdots & a_n^n \end{pmatrix},$$

where the a_i are its columns and a^i are its rows. We denote by $\mathcal{M}_n$ the space of $n \times n$ matrices endowed with the inner product

$$\langle\langle A, B\rangle\rangle_n = \sum_{i,j=1}^n a_i^j b_i^j = \sum_{i=1}^n a_i \cdot b_i = \sum_{j=1}^n a^j \cdot b^j.$$

Given two vectors $v, w \in \mathbb{R}^n$ we denote by $v \otimes w$ the matrix of rank one obtained by taking the usual row-times-column product of matrices of v and w^t, i.e., writing $v = (v_1, \ldots, v_n)$ and $w = (w_1, \ldots, w_n)$,

$$v \otimes w = \begin{pmatrix} v_1 w_1 & \cdots & v_1 w_n \\ \vdots & & \vdots \\ v_n w_1 & \cdots & v_n w_n \end{pmatrix} = (w_1 v, \ldots, w_n v) = \begin{pmatrix} v_1 w \\ \vdots \\ v_n w \end{pmatrix}.$$

For T, a regular transformation from an open subset of $\mathbb{R}^n$ to $\mathbb{R}^n$, ∇T is the Jacobian matrix; for v, a scalar valued function, ∇v is its gradient, seen as a row vector. By this way, given a vector b, $b \otimes \nabla v$ is an $n \times n$ matrix, while $\nabla v \cdot b$ is a scalar (inner product). The complement of a subset E of $\mathbb{R}^n$ is E^c; the Lebesgue measure is denoted by $\mu(\cdot)$.

We use the Sobolev spaces $W_0^{1,\infty}(\Omega, \mathbb{R}^n)$ and $W_0^{1,\infty}(\Omega, \mathbb{R})$, endowed with the usual norms, and adopt the convention that an element of $W_0^{1,\infty}(\Omega, \mathbb{R})$ or $W_0^{1,\infty}(\Omega, \mathbb{R}^n)$ is said to be continuous if it admits a continuous representative.

An open bounded subset E of $\mathbb{R}^n$ is called *regular* if the divergence theorem can be applied to E and to ∂E. We call a *regular partition* of E a finite collection $\{E_i, i = 1, \dots, m\}$ such that each E_i is open and regular, $E_i \cap E_j = \emptyset$ if $i \neq j$, and $E = (\bigcup_{i=1}^m E_i) \bigcup N$, where N is a null set.

An element T of $W^{1,\infty}(\Omega, \mathbb{R}^n)$ is called *piecewise affine* if it is continuous and there exists a regular partition of Ω such that T is affine on each element of the partition.

For a smooth function Φ defined on $\mathcal{M}_n$ we denote by $d^l\Phi(A)$ the lth differential of Φ at A and, abusing the notation, by

$$(d^l\Phi(A))_{(i_1,j_1),\dots,(i_l,j_l)}, \quad i_k, j_k = 1, \dots, n$$

the tensor representing the lth differential of Φ at A with respect to the canonical base in $\mathcal{M}_n$. In particular

$$(d\Phi(A))_{(i_1,j_1)}, \quad i_1, j_1 = 1, \dots, n$$

is the $n \times n$ matrix representing the first differential, in the sense that, for any $B \in \mathcal{M}_n$,

$$(d\Phi(A))(B) = \langle\langle d\Phi(A), B \rangle\rangle_n.$$

We recall, from [D1, p. 99], the following definition.

DEFINITION 2.1. *A Borel-measurable and locally integrable function $\Phi : \mathcal{M}_n \to \mathbb{R}$ is said to be* quasi-affine *if*

$$\Phi(A) = \frac{1}{\mu(D)} \int_D \Phi(A + \nabla u(x))dx$$

for every bounded domain $D \subset \mathbb{R}^n$, for every $A \in \mathcal{M}_n$, and for every $u \in W_0^{1,\infty}(D, \mathbb{R}^n)$.

We recall also that there exists a representation theorem for quasi-affine functions (see [B] or [D1, p. 117]) expressed in terms of the map

$$L : \mathcal{M}_n \to \mathcal{M}_{\nu(1)} \times \mathcal{M}_{\nu(2)} \times \cdots \times \mathcal{M}_{\nu(n-1)} \times \mathcal{M}_{\nu(n)},$$

where $\nu(s) = \binom{n}{s} = \frac{n!}{s!(n-s)!}$, given by

$$L(A) = \left(A, \mathrm{adj}_2(A), \dots, \mathrm{adj}_{n-1}(A), \det A\right),$$

where $\mathrm{adj}_s(A)$ stands for the $\nu(s) \times \nu(s)$ matrix of $s \times s$ minors of A. Roughly speaking, $L(A)$ is a "vector" whose "components" are square matrices of order $\nu(s)$.

Ball's theorem says that any quasi-affine function of A can be represented as a "scalar product" of $L(A)$ with a constant "vector." More precisely we have the following theorem.

THEOREM 2.1. *Let $\Phi : \mathcal{M}_n \to \mathbb{R}$. Then the following conditions are equivalent.*

(i) *Φ is quasi-affine.*

(ii) *There exists*

$$\beta = (\beta_1, \dots, \beta_n) \in \mathcal{M}_{\nu(1)} \times \mathcal{M}_{\nu(2)} \times \cdots \times \mathcal{M}_{\nu(n-1)} \times \mathcal{M}_{\nu(n)}$$

such that

$$\Phi(A) = \Phi(0) + \sum_{s=1}^{n} \langle\langle \beta_s, adj_s(A) \rangle\rangle_{\nu(s)}.$$

(iii) *For any* $a, b \in \mathbb{R}^n$

$$\Phi(A + a \otimes b) = \Phi(A) + \langle\langle d\Phi(A), a \otimes b\rangle\rangle_n.$$

Remarks. 1. When $\Phi(A) = \det A$, identifying as usual the differential with the representing matrix, it is $d\Phi(A) = \mathrm{adj}_{n-1}(A)$ (see [D1, p. 191]). Hence a matrix A has rank k if and only if $d^l\Phi(A) = 0$ for $l = 1, \dots, n-k-1$ and $d^l\Phi(A) \neq 0$ for $l = n-k, \dots, n$.

2. Point (ii) implies that Φ is a polynomial of degree less or equal than n; hence, in particular, if Φ is nonconstant, for every $A \in \mathcal{M}_n$ there exists $l \in \{1, \dots, n\}$ such that $d^l\Phi(A) \neq 0$.

3. Since the matrix representing the differential of the function determinant is the matrix of its maximal minors, point (ii) implies that each entry of the tensor representing a differential of some order of Φ at A is still a quasi-affine real-valued function of A.

LEMMA 2.1. *Let E be an open bounded subset of $\mathbb{R}^n$ and $V = \{v_i, i = 1, \dots, m\}$ a set of vectors of $\mathbb{R}^n$ such that $0 \in \mathrm{int}(\mathrm{co}(V))$. Then there exist a regular partition of E: $\{E_i, i = 1, \dots, m\}$ and a continuous function $w \in W_0^{1,\infty}(E, \mathbb{R})$ such that*

(i) $\nabla w = \sum_{i=1}^m v_i \chi_{E_i}$ *a.e. on Ω and*

(ii) $\sum_{i=1}^m \mu(E_i) v_i = 0$.

Proof. Let V^* be the polar set of $\mathrm{co}\{v_i, i = 1, \dots, m\}$. By Lemma 1 in [C] there exist a collection of m polytopes $V_1^*, \dots, V_m^*$ contained in V^* and a Lipschitz continuous function u, defined on $\mathbb{R}^n$ such that $V^* = \bigcup_{i=1}^m V_i^*$, $\mathrm{int}(V_i^*) \bigcap \mathrm{int}(V_j^*) = \emptyset$ for $i \neq j$, $u\big|_{(V^*)^c} = 0$, $\nabla u = \sum_{i=1}^m v_i \chi_{V_i^*}$, and

$$\sum_{i=1}^{m} \mu(V_i^*) v_i = 0. \tag{2.1}$$

Consider the following Vitali covering of E,

$$\{x + rV^*, \quad x \in E, \quad 0 < r < \mathrm{dist}(x, E^c)\},$$

and select a denumerable subcovering $\{S^j\}_{j \in \mathbb{N}}$,

$$S^j = \{x_j + r_j V^*, \quad x_j \in E, \quad r_j > 0\}$$

such that

(a) $\mathrm{int}(S^j) \bigcap \mathrm{int}(S^k) = \emptyset$ for $j \neq k$,

(b) $E = \left(\bigcup_{j=1}^\infty S^j\right) \bigcup N$, N null set, and

(c) $\mu(E) = \mu(V^*) \sum_{j=1}^\infty r_j^n$.

For any $j \in \mathbb{N}$ we define the subsets of S^j

$$S_i^j = \{x_j + r_j V_i^*\}, \quad i = 1, \dots, m;$$

for any $x \in \mathbb{R}^n$, we set

$$u_j(x) \equiv r_j u\left(\frac{x - x_j}{r_j}\right),$$

and, for $k \in \mathbb{N}$,

$$U_k = \left(\sum_{j=1}^k u_j\right)\bigg|_E.$$

Since u_j has the same regularity of u and $u_j\big|_{(S^j)^c} \equiv 0$, U_k belongs to $W_0^{1,\infty}(E,\mathbb{R})$ and moreover, for any $l \in \mathbb{N}$,

$$U_{k+l}(x) - U_k(x) = \begin{cases} 0 & \text{for } x \in \left(\bigcup_{j=1}^k S^j\right) \cup \left(\bigcup_{j=k+l}^\infty S^j\right) \\ u_j(x) & \text{for } x \in S^j, \quad j = k, \dots, k+l. \end{cases}$$

By (b),

$$\lim_{k\to\infty} \mu\left(\{x \in E : U_{k+l}(x) - U_k(x) \neq 0\}\right) = 0,$$

hence the sequence $\{U_k\}_{k\in\mathbb{N}}$ is fundamental in $W^{1,1}(E,\mathbb{R})$.

Now set

$$E_i = \bigcup_{j=1}^\infty \operatorname{int}(S_i^j)$$

and

$$w = \lim_{k\to\infty} U_k \quad \text{in} \quad W^{1,1}(E,\mathbb{R}).$$

We remark that, since each E_i is the union of a countable family of interiors of polytopes, it is regular in the sense specified above. Obviously w belongs to $W_0^{1,1}(E,\mathbb{R})$, and, given $j \in \mathbb{N}$,

$$U_k(x) = U_j(x) = u_j(x) \quad \text{for any } x \in S^j \text{ and for any } k \geq j;$$

hence, by pointwise convergence, $w(x) = u_j(x)$ for a.e. x in S^j. Thus w is continuous and belongs to $W_0^{1,\infty}(\Omega,\mathbb{R})$, and

$$\nabla w(x) = \nabla u\left(\frac{x - x_j}{r_j}\right) = v_i \quad \text{for a.e. } x \in S_i^j.$$

This implies (i). Statement (ii) is a trivial consequence of (2.1), (c), and the fact that

$$\mu(E_i) = \mu(V_i^*) \sum_{j=1}^\infty r_j^n. \qquad \square$$

3. Main result. We shall need the following lemmas.

LEMMA 3.1. *Let $D \in \mathcal{M}_n$ be a nonzero matrix. Let $\gamma_1, \gamma_2 \in \mathbb{R}$ with $\gamma_1 < 0 < \gamma_2$. Then there exist a vector $b \in \mathbb{R}^n$ and an n-dimensional polytope $P \subset \mathbb{R}^n$ with vertices $\{v_i^1, v_i^2, i = 1, \dots, n\}$, containing zero in its relative interior, such that*

$$\langle\langle D, b \otimes v_i^j \rangle\rangle_n = \gamma_j, \quad i = 1, \dots, n, \; j = 1, 2.$$

Proof. Write

$$D = \begin{pmatrix} d^1 \\ \vdots \\ d^n \end{pmatrix}$$

take a row d^j different from zero, and choose $b = e_j$. Let S be a symplex in $(d^j)^\perp$ with vertices $\{s_i, i = 1, \dots, n\}$ containing zero in its relative interior and define

$$v_i^1 = s_i + \frac{\gamma_1}{|d^j|^2} d^j, \quad v_i^2 = s_i + \frac{\gamma_2}{|d^j|^2} d^j, \quad i = 1, \dots, n.$$

Notice that for any vector v, $e_j \otimes v$ is the matrix whose rows are all zero except the jth one, which coincides with the row vector v^t. Hence

$$\langle\langle D, e_j \otimes v_i^j \rangle\rangle_n = d^j \cdot v_i^j$$

and, by the choice of v_i^j, we have the result. □

LEMMA 3.2. *Let Φ be a nonconstant real-valued quasi-affine function defined on $\mathcal{M}_n$, and let A be a critical point for Φ. Let $k \in \{0, \dots, n-2\}$ be such that $d^{n-k}\Phi(A) \neq 0$ and $d^l\Phi(A) = 0$ for any $l = 1, \dots, n-k-1$. Then there exist a vector $b \in \mathbb{R}^n$ and an n-dimensional polytope $P \subset \mathbb{R}^n$ with vertices $\{v_i, i = 1, \dots, 2n\} = V(P)$, containing zero in its interior, such that*

$$d^{n-k-1}\Phi(A + b \otimes v_i) \neq 0, \quad i = 1, \dots, 2n,$$

$$d^l\Phi(A + b \otimes v_i) = 0, \quad i = 1, \dots, 2n, \ l = 1, \dots, n-k-2.$$

Proof. Consider the tensor representing the $(n-k-1)$th differential of Φ at A:

$$(d^{n-k-1}\Phi(A))_{(i_1, j_1), \dots, (i_{n-k-1}, j_{n-k-1})}.$$

By assumption this tensor is zero and there exists a multiindex

$$\bar{J}_{n-k-1} = (\bar{i}_1, \bar{j}_1), \dots, (\bar{i}_{n-k-1}, \bar{j}_{n-k-1})$$

such that

$$d\,(d^{n-k-1}\Phi(A))_{\bar{J}_{n-k-1}}$$

is a nonzero matrix. By Lemma 3.1 there exist a vector b and a polytope P such that

$$\langle\langle d\,(d^{n-k-1}\Phi(A))_{\bar{J}_{n-k-1}}, b \otimes v_i \rangle\rangle_n \neq 0$$

for every $v_i \in V(P)$. Now we recall that the map $A \to (d^{n-k-1}\Phi(A))_{\bar{J}_{n-k-1}}$ is real valued and quasi-affine; hence, by point (iii) of Theorem 2.1,

$$(d^{n-k-1}\Phi(A + b \otimes v_i))_{\bar{J}_{n-k-1}} = (d^{n-k-1}\Phi(A))_{\bar{J}_{n-k-1}}$$

$$+ \langle\langle d\,(d^{n-k-1}\Phi(A))_{\bar{J}_{n-k-1}}, b \otimes v_i \rangle\rangle_n \neq 0$$

for every $v_i \in V(P)$.

Moreover, for any $l \in \{1, \dots, n-k-2\}$, if we denote a generic entry of the lth differential of Φ in A by

$$d\,(d^l\Phi(A))_{J_l},$$

where $J_l = (i_1, j_1), \dots, (i_l, j_l)$, we have, as before,

$$(d^l\,(\Phi(A + b \otimes v_i))_{J_l} = (d^l\,(\Phi(A))_{J_l} + \langle\langle d\,(d^l\Phi(A))_{J_l}, b \otimes v_i \rangle\rangle_n = 0$$

for any $v_i \in V(P)$. Hence all of the differentials of Φ at A are zero up to order $(n-k-2)$, while the $(n-k-1)$st differential is different from zero. □

The following lemma defines the auxiliary transformation $T_r \in T_B + W_0^{1,\infty}(\Omega, \mathbb{R}^n)$.

LEMMA 3.3. *Let $T_B : \mathbb{R}^n \to \mathbb{R}^n$ be a piecewise-affine transformation with $\Phi(\nabla T_B)$ constant. Then there exists a piecewise-affine transformation $T_r \in T_B + W_0^{1,\infty}(\Omega, \mathbb{R}^n)$ such that*

(i) $\Phi(\nabla T_r) \equiv \Phi(\nabla T_B)$ *and*

(ii) $d\Phi(\nabla T_r) \neq 0$ *a.e. in* Ω.

Proof. By assumption

$$T_B = \sum_{h=1}^{m} T^h \chi_{\Omega_h},$$

where $\{\Omega_i, i = 1, \dots, m\}$ is a regular partition of Ω and each T^i is affine. If $d\Phi(\nabla T^i) \neq 0$ for every i, we set $T_r \equiv T_B$.

Consider a generic T^j such that $d\Phi(\nabla T^j) = 0$. We claim that there exists a piecewise-affine T_r^j in $T^j + W_0^{1,\infty}(\Omega_j, \mathbb{R}^n)$ such that $\Phi(\nabla T_r^j) \equiv \Phi(\nabla T^j)$ and $d\Phi(\nabla T_r^j) \neq 0$ a.e. in Ω_j.

To prove this claim we set, for notational convenience, $\Omega_j = E$ and $T^j = T$.

Let k ($k \geq n-2$) be such that $d^{n-k}\Phi(\nabla T) \neq 0$ and $d^l\Phi(\nabla T) = 0$ for $l = 1, \dots, n-k-1$. By Lemma 3.2 there exist $b \in \mathbb{R}^n$ and an n-dimensional polytope P containing zero in its interior such that

$$d^{n-k-1}\Phi(\nabla T + b \otimes v_i) \neq 0,$$
$$d^l\Phi(\nabla T + b \otimes v_i) = 0, \quad l = 1, \dots, n-k-2.$$

for any $v_i \in V(P)$. Let $u \in W_0^{1,\infty}(E, \mathbb{R}^n)$ be defined as in Lemma 2.1 with $V = V(P)$. Consider the transformation

$$T_1(x) = T(x) + u(x)b$$

so that

$$\nabla T_1(x) = \nabla T + b \otimes \nabla u(x), \text{ a.e. in } \Omega.$$

T_1 is continuous and belongs to $T + W_0^{1,\infty}(E, \mathbb{R}^n)$ and there exists a regular partition of E: $\{E_i^1, i = 1, \dots, 2n\}$ such that

$$T_1 = \sum_{i=1}^{2n} T_1^i \chi_{E_i^1},$$

where each T_1^i is affine and, more precisely,

$$\nabla T_1^i(x) = \nabla T + b \otimes v_i, \quad i = 1, \dots, 2n.$$

Hence, for any $i \in \{1, \dots, 2n\}$, $d^{n-k-1}\Phi(\nabla T_1^i) \neq 0$, $d^l\Phi(\nabla T_1^i) = 0$ for $l = 1, \dots, n-k-2$, and $\Phi(\nabla T_1^i) = \Phi(\nabla T)$. If $k = n-2$ we set $T_r(E) = T_1$. Otherwise, repeating the previous procedure, for each $i \in \{1, \dots, 2n\}$ we can define a piecewise-affine transformation $T_2^i \in T_1^i + W_0^{1,\infty}(E_i^1, \mathbb{R}^n)$ such that

$$d^{n-k-2}\Phi(\nabla T_2^i) \neq 0,$$

$$d^l\Phi(\nabla T_2^i) = 0, \quad l = 1, \dots, n-k-3,$$

$$\Phi(\nabla T_2^i) = \Phi(\nabla T_1^i) = \Phi(\nabla T).$$

Extend T_2^i on E by setting $T_2^i = T_1^i$ on $E \setminus E_i^1$ and define

$$T_2 = \sum_{i=1}^{2n} T_2^i \chi_{E_i^1}.$$

Then

$$T_2 = T_1 + \sum_{i=1}^{2n} (T_2^i - T_1^i)\chi_{E_i^1}.$$

Since the second term at the right-hand side is in $W_0^{1,\infty}(E, \mathbb{R}^n)$, T_2 belongs to $T_1 + W_0^{1,\infty}(E, \mathbb{R}^n) = T + W_0^{1,\infty}(E, \mathbb{R}^n)$ and has the same properties of T_1. This procedure can be iterated $n-k-1$ times to obtain the piecewise-affine transformation $T_r(E) = T_{n-k-1}$ belonging to $T + W_0^{1,\infty}(E, \mathbb{R}^n)$.

Now, for any $h = 1, \dots, m$, set $T_r^h = T^h$ if $d\Phi(\nabla T^h) \neq 0$ and $T_r^h = T_r(\Omega_h)$, defined as above, if $d\Phi(\nabla T^h) = 0$. Extend T_r^h on Ω by setting $T_r^h = T^h$ on $\Omega \setminus \Omega_h$ and define T_r as

$$T_r = \sum_{h=1}^{m} T_r^h \chi_{\Omega_h}.$$

We have

$$T_r = T_B + \sum_{h=1}^{m} \left(T_r^h - T^h\right) \chi_{\Omega_h},$$

and since the second term on the right-hand side is in $W_0^{1,\infty}(\Omega, \mathbb{R}^n)$, T_r belongs to $T_B + W_0^{1,\infty}(\Omega, \mathbb{R}^n)$. Moreover, by the above, $\Phi(\nabla T_r) = \Phi(\nabla T_B)$ and $d\Phi(\nabla T_r) \neq 0$ a.e. in Ω. □

THEOREM 3.1. *Let $\Phi : \mathcal{M}_n \to \mathbb{R}$ be nonconstant and quasi-affine. Let $T_B : \mathbb{R}^n \to \mathbb{R}^n$ be a piecewise-affine transformation such that $\Phi(\nabla T_B)$ is constant. Let $\alpha, \beta \in \mathbb{R}$ $(\alpha < \beta)$ $\lambda \in]0,1[$ be such that*

$$\Phi(\nabla T_B) = \lambda\alpha + (1-\lambda)\beta.$$

Then there exist two open regular disjoint subsets of Ω, Ω^α, and Ω^β, a null set N, and a piecewise-affine transformation $T_{\alpha,\beta} \in T_B + W_0^{1,\infty}(\Omega, \mathbb{R}^n)$, such that $\Omega = N \bigcup \Omega^\alpha \bigcup \Omega^\beta$ and

(i) $\Phi(\nabla T_{\alpha,\beta}) = \alpha\chi_{\Omega^\alpha} + \beta\chi_{\Omega^\beta}$.

(ii) $\int_\Omega \Phi(\nabla T_{\alpha,\beta}(x))dx = \mu(\Omega)\Phi(\nabla T_B)$,

or, in other words,

(ii′) $\frac{\mu(\Omega^\alpha)}{\mu(\Omega)} = \lambda$, $\frac{\mu(\Omega^\beta)}{\mu(\Omega)} = 1-\lambda$.

Proof. By the previous lemma there exists a piecewise-affine transformation $T_r \in T_B + W_0^{1,\infty}(\Omega, \mathbb{R}^n)$ such that

$$T_r = \sum_{i=1}^{m} T^i \chi_{\Omega_i},$$

where $\{\Omega_i, i = 1, \dots, m\}$ is a regular partition of Ω, each T^i is affine, $\Phi(\nabla T^i) = \Phi(\nabla T_B)$, and $d\Phi(\nabla T^i) \neq 0$ for every index i.

Fix an index $j \in \{1, \dots, m\}$ and consider the transformation T^j. Apply Lemma 3.1 with $D = \nabla T^j$, $\gamma_1 = \alpha - \Phi(\nabla T^j)$, $\gamma_2 = \beta - \Phi(\nabla T^j)$ to obtain a vector b and an

n-dimensional polytope $P \subset \mathbb{R}^n$ with vertices $\{v_i^\alpha, v_i^\beta, i = 1, \dots, 2n\}$ containing zero in its interior such that

$$\langle\!\langle d\Phi(\nabla T^j), b \otimes v_i^\alpha \rangle\!\rangle_n = \alpha - \Phi(\nabla T^j)$$

and

$$\langle\!\langle d\Phi(\nabla T^j), b \otimes v_i^\beta \rangle\!\rangle_n = \beta - \Phi(\nabla T^j).$$

Define $u \in W_0^{1,\infty}(\Omega_j, \mathbb{R})$ as in Lemma 2.1 with $V = V(P)$ and $E = \Omega_j$. Consider the transformation

$$S^j(x) = T^j(x) + u(x)b.$$

S^j is piecewise-affine and belongs to $T^j + W_0^{1,\infty}(\Omega_j, \mathbb{R}^n)$; moreover, by Lemma 2.1 and by Theorem 2.1 (iii), there exist two open regular disjoint subsets of Ω_j, Ω_j^α, and Ω_j^β and a null set N, such that $\Omega_j = N \bigcup \Omega_j^\alpha \bigcup \Omega_j^\beta$ and

$$\Phi(\nabla S^j) = \Phi(\nabla T^j + b \otimes \nabla u)$$

$$= \Phi(\nabla T^j) + \langle\!\langle d\Phi(\nabla T^j), b \otimes \nabla u \rangle\!\rangle_n = \alpha \chi_{\Omega_j^\alpha} + \beta \chi_{\Omega_j^\beta}.$$

Repeat this construction for any index $j \in \{1, \dots, m\}$ and extend S^j as T^j on $\Omega \setminus \Omega_j$.

Finally define $T_{\alpha,\beta}$ by

$$T_{\alpha,\beta} = \sum_{j=1}^m S^j \chi_{\Omega_j}.$$

$T_{\alpha,\beta}$ is piecewise affine and can be written as

$$T_{\alpha,\beta} = T_r + \sum_{j=1}^m \left(S^j - T^j\right) \chi_{\Omega_j}.$$

Since the second term on the right-hand side is in $W_0^{1,\infty}(\Omega, \mathbb{R}^n)$, $T_{\alpha,\beta}$ belongs to $T_r + W_0^{1,\infty}(\Omega, \mathbb{R}^n) = T_B + W_0^{1,\infty}(\Omega, \mathbb{R}^n)$. Setting

$$\Omega^\alpha = \bigcup_{j=1}^m \Omega_j^\alpha \quad \text{and} \quad \Omega^\beta = \bigcup_{j=1}^m \Omega_j^\beta,$$

we have, since Φ is quasi-affine,

$$\int_\Omega \Phi(\nabla T_{\alpha,\beta}(x))dx = \sum_{j=1}^m \mu(\Omega_j)\Phi(\nabla S^j)$$

$$= \sum_{j=1}^m \alpha\mu(\Omega_j^\alpha) + \beta\mu(\Omega_j^\beta) = \alpha\mu(\Omega^\alpha) + \beta\mu(\Omega^\beta) = \mu(\Omega)\Phi(\nabla T_{\alpha,\beta}),$$

i.e., point (ii). □

We are ready to prove the main result.

THEOREM 3.2. *Let Ω be an open bounded subset of $\mathbb{R}^n$ and let $g : \mathbb{R} \to \mathbb{R} \cup \{+\infty\}$ be a proper lower-semicontinuous function satisfying*

$$\lim_{|t|\to\infty} \frac{g(t)}{|t|} = +\infty.$$

Then for any piecewise-affine transformation $T_B : \mathbb{R}^n \to \mathbb{R}^n$ such that $\Phi(\nabla T_B)$ is constant and belongs to co(dom(g))*, the problem*

$$\mathcal{P}: \ \textit{Minimize} \int_\Omega g(\Phi(\nabla T(x)))dx, \qquad T \in T_B + W_0^{1,\infty}(\Omega, \mathbb{R}^n)$$

admits at least one solution.

Proof. Consider g^{**}, the bipolar of g, as defined in [ET] and call dom(g) the set in which g is strictly less than infinity.

(a) Consider first the case in which $\Phi(\nabla T_B) \in \partial(\text{co}(\text{dom}(g)))$. We claim that T_B is a solution of $\mathcal{P}$. In the case $\Phi(\nabla T_B) = \sup(\text{co}(\text{dom}(g)))$ we remark that for any $T \in T_B + W_0^{1,\infty}(\Omega, \mathbb{R}^n)$, by Theorem 1 of [D2] and the following remark, we have, since Φ is quasi-affine,

$$\int_\Omega (\Phi(\nabla T(x)) - \Phi(\nabla T_B(x)))\, dx = 0.$$

For any T such that

$$\int_\Omega g\,(\Phi(\nabla T(x)))\, dx < +\infty,$$

for a.e. $x \in \Omega$ it must be that $\Phi(\nabla T(x)) \leq \Phi(\nabla T_B(x))$. Hence $\Phi(\nabla T(x)) = \Phi(\nabla T_B(x))$ for a.e. $x \in \Omega$, so that T_B is a solution and any other solution T_1 of $\mathcal{P}$ must be such that $\Phi(\nabla T_1) = \Phi(\nabla T_B)$.

The case $\Phi(\nabla T_B) = \inf(\text{co}(\text{dom}(g)))$ is analogous.

(b) Consider now the case $\Phi(\nabla T_B) \in \text{int}(\text{co}(\text{dom}(g)))$ (where this set can be the whole $\mathbb{R}$). There exists a line ρ separating the point $(\Phi(\nabla T_B), g^{**}(\Phi(\nabla T_B)))$ from the closed convex set epi(g^{**}). Since $\Phi(\nabla T_B)$ is in the interior of dom(g^{**}), ρ cannot be vertical, i.e., there exist $\gamma, \delta \in \mathbb{R}$ such that, for $t \in \text{dom}(g^{**})$,

$$g(t) \geq g^{**}(t) \geq \gamma t + \delta$$

and

$$g^{**}(\Phi(\nabla T_B)) = \gamma \Phi(\nabla T_B) + \delta.$$

Let $T \in T_B + W_0^{1,\infty}(\Omega, \mathbb{R}^n)$. We have

$$\int_\Omega g\,(\Phi\,(\nabla T(x)))\, dx \geq \int_\Omega g^{**}\,(\Phi\,(\nabla T(x)))\, dx \geq \int_\Omega (\gamma\Phi\,(\nabla T(x)) + \delta)\, dx;$$

by the previous remark, since Φ is quasi-affine,

$$\int_\Omega (\gamma\Phi\,(\nabla T(x)) + \delta)\, dx = \int_\Omega (\gamma\Phi\,(\nabla T_B(x)) + \delta)\, dx = \int_\Omega g^{**}\,(\Phi\,(\nabla T_B(x)))\, dx.$$

When $g\,(\Phi\,(\nabla T_B)) = g^{**}\,(\Phi\,(\nabla T_B))$ the above argument shows that T_B is a solution. Otherwise, by a slight modification of IX.3.3 of [ET], taking into account the superlinear growth condition on g, we can say that there exist $\alpha, \beta \in \mathbb{R}$, $\lambda \in]0, 1[$ such that

$$\Phi\,(\nabla T_B) = \lambda\alpha + (1 - \lambda)\beta$$

and

$$g^{**}\left(\Phi\left(\nabla T_B\right)\right) = \lambda g(\alpha) + (1-\lambda)g(\beta).$$

In this case,

$$\int_{\Omega} g^{**}\left(\Phi\left(\nabla T_B(x)\right)\right) dx = \lambda\mu(\Omega)g(\alpha) + (1-\lambda)\mu(\Omega)g(\beta).$$

Hence the transformation $T_{\alpha,\beta}$ given by Theorem 3.1 is a solution of $\mathcal{P}$. □

Remark. Since the function Φ is real valued, the problem of finding a solution is underdetermined and, in general, one cannot expect uniqueness of the solution. Actually, in the case of the Jacobian determinant, it is easy to see that the problem admits infinitely many solutions. Indeed when T_B is not a solution, the assertion follows easily from the construction of the solution defined in Theorem 3.1, since it depends on a scalar function v, which can be defined in infinite ways (depending on the choice of the set of vectors which constitute the range of the gradient of v). When T_B is a solution of $\mathcal{P}$ we simply notice that, given a regular transformation $J : \Omega \to \Omega$, different from the identity, such that $\det(\nabla J) = 1$ on Ω and $J\big|_{\partial\Omega} = I\big|_{\partial\Omega}$ (I denotes the identity),

$$\det\left(\nabla\left(T_B \circ J\right)\right) = \det(\nabla T_B)$$

and

$$T_B \circ J\big|_{\partial\Omega} = T_B\big|_{\partial\Omega}.$$

Hence $T \equiv T_B \circ J$ is a solution. Since there exist infinitely many transformations J with such properties (see [DM]), the assertion is proved.

REFERENCES

[B] J. M. Ball, *Convexity conditions and existence theorems in nonlinear elasticity*, Arch. Rational Mech. Anal., 63 (1977), pp. 337–403.

[C] A. Cellina, *On minima of a functional of the gradient: sufficient conditions*, Nonlinear Anal., 20 (1993), pp. 337–341.

[D1] B. Dacorogna, *Direct Methods in Calculus of Variations*, Springer-Verlag, Berlin 1989.

[D2] B. Dacorogna, *A relaxation theorem and its application to the equilibrium of gases*, Arch. Rational Mech. Anal., 77 (1981), pp. 359–386.

[DM] B. Dacorogna and J. Moser, *On a partial differential equation involving the Jacobian determinant*, Ann. Inst. H. Poincaré Anal. Non Linéaire, 7 (1990), pp. 1–26.

[ET] I. Ekeland and R. Temam, *Convex Analysis and Variational Problems*, North Holland, Amsterdam, 1976.

[M] J. Moser, *On the volume elements on a manifold*, Trans. Amer. Math. Soc., 120 (1965), pp. 286–294.

[MS] E. Mascolo and R. Schianchi, *Existence theorems for nonconvex problems.* J. Math. Pures Appl. 62 (1983), pp. 349–359.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 33, No. 4, pp. 971-999, July 1995

© 1995 Society for Industrial and Applied Mathematics
001

FEEDBACK LAWS FOR NONLINEAR DISTRIBUTED CONTROL PROBLEMS VIA TROTTER-TYPE PRODUCT FORMULAE*

CĂTĂLIN POPA†

Abstract. A feedback law expressed by means of a Trotter product formula approximation of the dynamic programming equation, and which provides approximately optimal controls, is established for the control systems governed by a certain class of variational inequalities of parabolic type. To this purpose, two general Lie–Trotter formulas for the dynamic programming equation are proposed, and corresponding convergence results (generalizing previous results of the author) are proved. The framework also includes the control systems described by the parabolic obstacle problem as well as those governed by semilinear parabolic equations.

Key words. optimal control, variational inequalities, Hamilton–Jacobi equation, Trotter product formula, feedback law

AMS subject classifications. 49L10, 49L20, 49N35, 93B52, 93C20

1. Introduction. The aim of this paper is to show how a Trotter product formula treatment of the dynamic programming equation leads to a certain discrete feedback law for the optimal control problem, which seems very useful in effective synthesis of optimal control.

Our framework is given by the following class of nonlinear distributed optimal control problems:

(P) minimize

$$\int_0^T (h(u(t)) + g(y(t)))dt + l(y(T)) \tag{1.1}$$

over all $u \in L^2(0, T; \mathcal{U})$, where $y \in C([0, T]; \mathcal{H})$ satisfies the state equation

$$y'(t) + Ay(t) + \partial\phi(y(t)) \ni Bu(t) \tag{1.2}$$

and the initial condition

$$y(0) = y^0. \tag{1.3}$$

Here $\mathcal{H}$ and $\mathcal{U}$ are two real Hilbert spaces whose scalar products and norms are denoted by the same symbols, $(\cdot, \cdot)$ and $|\cdot|$, respectively.

We impose the following hypotheses on the data:

(i) $h : \mathcal{U} \to (-\infty, +\infty]$ is convex, lower semicontinuous, not identically $+\infty$, and satisfies

$$h(u) \geq c_1|u|^2 - c_2 \quad \text{for all } u \in \mathcal{U}, \tag{1.4}$$

where $c_1 > 0$ and $c_2 \in \mathbf{R}$.

(ii) $g, l : \mathcal{H} \to \mathbf{R}$ are Lipschitz continuous on bounded subsets and bounded from below by affine functions.

(iii) $A : \mathcal{V} \to \mathcal{V}'$ is linear, continuous, and symmetric, $\mathcal{V}$ being a Hilbert space continuously and densely imbedded in $\mathcal{H}$ with $\mathcal{V}'$ its dual space. Identifying $\mathcal{H}$ with its own dual, we have $\mathcal{V} \subset \mathcal{H} \subset \mathcal{V}'$. Denote by $(\cdot, \cdot)$ the pairing between $\mathcal{V}$ and $\mathcal{V}'$, and by $|\cdot|_{\mathcal{V}}$ the norm of $\mathcal{V}$. These being specified, we assume, in addition, that A satisfies

$$(Ay, y) \geq \omega|y|_{\mathcal{V}}^2 \quad \text{for all } y \in \mathcal{V}, \tag{1.5}$$

* Received by the editors January 11, 1993; accepted for publication (in revised form) January 5, 1994.

† Facultatea de Matematică, Universitatea "Al.I. Cuza," B-dul Copou 11, 6600 Iaşi, Romania. (cpopa@uaic.ro).

where $\omega > 0$, and the inclusion $\mathcal{V} \subset \mathcal{H}$ is compact. Finally, set $D(A) = \{y \in \mathcal{V} : Ay \in \mathcal{H}\}$.

(iv) $\phi : \mathcal{H} \to (-\infty, +\infty]$ is convex, lower semicontinuous, not identically $+\infty$, and such that $A + \partial\phi$ is maximal monotone in $\mathcal{H} \times \mathcal{H}$. (Here $\partial\phi$ is the subdifferential of ϕ.)

(v) $B : \mathcal{U} \to \mathcal{H}$ is a linear continuous operator.

To assure the maximal monotonicity of $A + \partial\phi$ in $\mathcal{H} \times \mathcal{H}$ it suffices to have (see [2, Thm. 1.10])

$$(Ay, (\partial\phi)_\eta(y)) \geq -c(1 + |(\partial\phi)_\eta(y)|)(1 + |y|) \quad \text{for all } y \in D(A) \text{ and } \eta > 0, \tag{1.6}$$

where $c > 0$ and $(\partial\phi)_\eta$ is the Yosida approximation of $\partial\phi$, i.e., $(\partial\phi)_\eta = \eta^{-1}(I - (I + \eta\partial\phi)^{-1})$, I being the identity operator of $\mathcal{H}$.

By a standard existence result, for any $y^0 \in \overline{D(A) \cap D(\partial\phi)}$, problem (1.2), (1.3) has a unique weak (integral) solution. Moreover, under (1.6), if $y^0 \in \mathcal{V} \cap D(\phi)$, then (1.2), (1.3) has a unique solution in $W^{1,2}([0,T];\mathcal{H}) \cap C([0,T];\mathcal{V}) \cap L^2(0,T;D(A))$ (see [2, Thm. 4.3]).

According to the dynamic programming method, we associate with the optimal control problem (P) the corresponding optimal value function

$$\begin{aligned} V(t,y) = \inf\bigg\{ \int_t^T (h(u(s)) + g(y(s)))ds + l(y(T)) : y' + Ay + \partial\phi(y) \ni Bu, \\ y(t) = y, u \in L^2(t,T;\mathcal{U}) \bigg\}, \qquad (t,y) \in [0,T] \times \overline{D(A) \cap D(\partial\phi)}. \end{aligned} \tag{1.7}$$

The function V formally satisfies the following infinite-dimensional Hamilton–Jacobi equation, called, in this context, the dynamic programming equation associated with problem (P):

$$\begin{aligned} D_tV(t,y) - h^*(-B^*D_yV(t,y)) - (Ay + \partial\phi(y), D_yV(t,y)) \\ = -g(y) \quad \text{in } [0,T] \times \mathcal{H}, \end{aligned} \tag{1.8}$$

together with the final condition

$$V(T,y) = l(y), \qquad y \in \mathcal{H}. \tag{1.9}$$

Here h^* denotes the convex conjugate of h and B^* is the adjoint of B. As a matter of fact, one can show that V is the unique solution of (1.8), (1.9) in a certain viscosity sense (see [12]).

Let $u^* \in L^2(0,T;\mathcal{U})$ be an arbitrary optimal control for problem (P) and $y^* \in C([0,T];\mathcal{H})$ the corresponding optimal arc. Heuristic considerations based on the Hamilton–Jacobi equation (1.8) lead to the following feedback law:

$$u^*(t) \in \partial h^*(-B^*D_yV(t,y^*(t))), \qquad t \in (0,T). \tag{1.10}$$

However, this formula (if we should prove it) is almost inapplicable to the effective synthesis of optimal control for (P), mainly because it is very difficult to calculate V: the Hamilton–Jacobi equation (1.8) is a very complicated mathematical object even in the finite-dimensional case. At present there are no constructive theories or adequate approximation methods for such equations, except in some special cases.

A new perspective on this topic is offered by an idea of V. Barbu introduced in [3], [4]. This consists of decoupling the terms containing $D_yV(t,y)$ in the dynamic programming equation by decomposing the corresponding Cauchy problem into a sequence of several simpler problems. Performing this operation successively on small time intervals, we obtain an approximate solution to the considered Cauchy problem, which may be interpreted as a Lie–Trotter product formula for the dynamic programming equation viewed as an evolution equation on a suitable space of function on $\mathcal{H}$. As V. Arnăutu has shown in [1], this method turns out to be very efficacious in numerical computation of V, at least in the finite-dimensional case ($n = 2, 3$).

In a previous paper [11], the author decomposed the Cauchy problem (1.8), (1.9) into two problems and expressed their solution by a suitable approximation and a Hopf–Oleinik–Lax representation formula (i.e., an explicit formula for the solution of problem (1.8), (1.9) in which $A + \partial\phi = 0$ and $g = 0$ given by a variant of (1.7) where the infimum is taken over $u \in \mathcal{U}$; see [9], [10] for the original formulas). In the case when ϕ in (1.2) is the indicator function I_K of a closed convex subset K of $\mathcal{H}$, this approach yields more general convergence results (see [11, Thms. 1 and 2]) as well as simpler formulas for calculating V.

In the present paper, we shall use one of the Trotter product formulas proposed in [11] to obtain a discrete variant of the feedback law (1.10), which will offer a reasonable way for synthesis of optimal control. The aforementioned formula expresses the convergence of the scheme

(1.11)

$$\begin{cases} D_tV^\varepsilon(t,y) - h^*(-B^*D_yV^\varepsilon(t,y)) = 0 \quad \text{in}\, [i\varepsilon,(i+1)\varepsilon) \times \mathcal{H}, \\ V^\varepsilon((i+1)\varepsilon - 0, y) = V^\varepsilon((i+1)\varepsilon, (I+\varepsilon\partial\phi)^{-1}(I+\varepsilon A)^{-1}y) \\ \qquad\qquad + \varepsilon g((I+\varepsilon\partial\phi)^{-1}(I+\varepsilon A)^{-1}y) \quad \text{on}\, \mathcal{H}, i = 0,1,\dots,N-1, \\ V^\varepsilon(N\varepsilon, y) = l(y), y \in \mathcal{H}, \end{cases}$$

to the solution of the dynamic programming equation (1.8) with the final condition (1.9). Here $\varepsilon = T/N$, where N is a positive integer. Solving (1.11) by a Lax-type formula, we obtain an approximate solution to problem (1.8), (1.9). For simplicity, let us indicate it only for $t = i\varepsilon$:

(1.12)

$$\begin{cases} V^\varepsilon(i\varepsilon, y) = \inf\{\varepsilon h(u) + \varepsilon g((I+\varepsilon\partial\phi)^{-1}(I+\varepsilon A)^{-1}(y+\varepsilon Bu)) \\ \qquad + V^\varepsilon((i+1)\varepsilon, (I+\varepsilon\partial\phi)^{-1}(I+\varepsilon A)^{-1}(y+\varepsilon Bu)) : u \in \mathcal{U}\}, \quad y \in \mathcal{H}, \\ \qquad i = 0,1,\dots,N-1, \\ V^\varepsilon(T,y) = l(y), y \in \mathcal{H}. \end{cases}$$

First of all, we shall prove the following Trotter-type product formula (Theorem 3.1):

$$\lim_{\varepsilon \to 0} V^\varepsilon(t,y) = V(t,y), \qquad (t,y) \in [0,T] \times \overline{D(A) \cap D(\partial\phi)}. \tag{1.13}$$

This is a generalization of a result from [11]: the indicator function I_K from [11] is replaced by a general lower semicontinuous convex function ϕ. So we also cover the case of the control systems governed by semilinear parabolic equations (see [5] for a related result) besides that of the parabolic obstacle problem. An alternative Trotter formula is given by Theorem 3.2.

Scheme (1.12) and formula (1.13) suggest to us that problem (P) could be approximated by the following family of discrete problems:

(P^ε) minimize

$$\sum_{i=1}^{N} \varepsilon(h(u_i) + g(y_i)) + l(y_N)$$

over all $(u_1, u_2, \dots, u_N) \in \mathcal{U}^N$, where $(y_1, y_2, \dots, y_N) \in \mathcal{H}^N$ satisfies the scheme

$$\begin{cases} y_i = (I+\varepsilon\partial\phi)^{-1}(I+\varepsilon A)^{-1}(y_{i-1}+\varepsilon Bu_i), \quad i = 1,2,\dots,N, \\ y_0 = y^0. \end{cases}$$

We shall show (Theorem 5.1) that this is indeed the case: every optimal N-tuple $(u_1^\varepsilon, u_2^\varepsilon, \dots, u_N^\varepsilon)$ for problem (P^ε) may be viewed as an approximately optimal control for problem (P).

The main result of this paper (Theorem 6.1) is established for a little more restricted class of nonlinear control problems: in (P), $\mathcal{H} = L^2(\Omega)$ and the state equation is

$$y' + Ay + \beta(y) \ni Bu,$$

where β is a maximal monotone graph in $\mathbf{R}^2$. In this context, any optimal N-tuple $(u_1^\varepsilon, u_2^\varepsilon, \dots, u_N^\varepsilon)$ for problem (P^ε) satisfies the following feedback law:

$$u_i^\varepsilon \in \partial h^*(-B^*\partial_y V^\varepsilon((i-1)\varepsilon, y_{i-1}^\varepsilon)), \qquad i = 1, 2, \dots, N, \tag{1.14}$$

where

$$\begin{cases} y_i^\varepsilon = (I + \varepsilon\beta)^{-1}(I + \varepsilon A)^{-1}(y_{i-1}^\varepsilon + \varepsilon B u_i^\varepsilon), \\ y_0^\varepsilon = y^0. \end{cases} \tag{1.15}$$

Here $\partial_y V^\varepsilon$ is the Clarke generalized gradient of V^ε with respect to y. Formula (1.14) may be interpreted as an approximate feedback law for problem (P).

System (1.14), (1.15) offers us a set of necessary optimality conditions for (P^ε), which can be useful in constructing a suboptimal control for (P), but on the condition that V^ε is known. The numerical calculation of V^ε is another problem, but a glance at (1.12) or (1.11) clearly shows how much easier it is to get V^ε than V. (For the finite-dimensional case we refer to [1].)

2. Trotter schemes for the dynamic programming equation. By a simple change of unknown function, we can transform problem (1.8), (1.9) into a Cauchy problem but with initial condition, which is more convenient in our further considerations. Indeed, we define

$$W(t,y) = V(T-t,y) \quad \text{for } (t,y) \in [0,T] \times \overline{D(A) \cap D(\partial\phi)}.$$

It is easy to check that

$$\begin{aligned} W(t,y) = \inf\Big\{ &\int_0^t (h(u(s)) + g(y(s)))ds + l(y(T)) : y' + Ay + \partial\phi(y) \ni Bu, \\ &y(0) = y,\, u \in L^2(0,t;\mathcal{U})\Big\}, \qquad (t,y) \in [0,T] \times \overline{D(A) \cap D(\partial\phi)}. \end{aligned} \tag{2.1}$$

The function W is the unique viscosity solution (in the sense of [12]) of the Hamilton–Jacobi equation

$$D_t W(t,y) + h^*(-B^* D_y W(t,y)) + (Ay + \partial\phi(y),\, D_y W(t,y)) = g(y) \tag{2.2}$$

with the initial condition

$$W(0,y) = l(y), \qquad y \in \mathcal{H}. \tag{2.3}$$

Let $\varepsilon = T/N$, where N is a sufficiently large positive integer. On each subinterval $((i-1)\varepsilon, i\varepsilon]$, we decompose the Cauchy problem (2.2), (2.3) into two problems, the second of these being the following:

$$\begin{aligned} &D_t W_\varepsilon(t,y) + h^*(-B^* D_y W_\varepsilon(t,y)) = 0 \quad \text{in } ((i-1)\varepsilon, i\varepsilon] \times \mathcal{H}, \\ &W_\varepsilon((i-1)\varepsilon + 0, y) = W_\varepsilon((i-1)\varepsilon, (I + \varepsilon\partial\phi)^{-1}(I + \varepsilon A)^{-1} y) \\ &\qquad + \varepsilon g((I + \varepsilon\partial\phi)^{-1}(I + \varepsilon A)^{-1} y), \qquad y \in \mathcal{H}, \quad i = 1, 2, \dots, N, \end{aligned}$$

where

$$W_\varepsilon(0,y) = l(y), \qquad y \in \mathcal{H}.$$

Here $y \mapsto W_\varepsilon((i-1)\varepsilon, (I+\varepsilon\partial\phi)^{-1}(I+\varepsilon A)^{-1}y) + \varepsilon g((I+\varepsilon\partial\phi)^{-1}(I+\varepsilon A)^{-1}y)$ represents an approximate solution of the first problem of the decomposition (which contains the unbounded term—corresponding to the dynamics of (P)—and the right-hand side g) with $W_\varepsilon((i-1)\varepsilon,\cdot)$ as initial datum. (Note that for $\partial\phi = 0$ and $g \equiv 0$, $y \mapsto W_\varepsilon((i-1)\varepsilon, e^{-A\varepsilon}y)$ is just the exact solution of the first problem at $t = i\varepsilon$.) Now expressing the exact solution of the second problem of the decomposition by a Lax-type representation formula (cf. [10, Eq. (2.12)]), we obtain

$$\begin{aligned} W_\varepsilon(t,y) = \inf\{&(t-(i-1)\varepsilon)h(u) + \varepsilon g((I+\varepsilon\partial\phi)^{-1}(I+\varepsilon A)^{-1}(y+(t-(i-1)\varepsilon)Bu)) \\ &+ W_\varepsilon((i-1)\varepsilon, (I+\varepsilon\partial\phi)^{-1}(I+\varepsilon A)^{-1}(y+(t-(i-1)\varepsilon)Bu)) : u \in \mathcal{U}\}, \\ &(t,y) \in ((i-1)\varepsilon, i\varepsilon] \times \mathcal{H}, \quad i = 1,2,\ldots,N, \\ W_\varepsilon(0,y) = l(y)&, y \in \mathcal{H}. \end{aligned}$$

Obviously, it is natural to interpret W_ε as an approximate solution to (2.2), (2.3), but we must prove that we may do this. For technical reasons, we shall slightly modify the above scheme to obtain

$$W^\varepsilon(t,y) = \begin{cases} \inf\{\varepsilon h(u) + \varepsilon g((I+\varepsilon\partial\phi)^{-1}(I+\varepsilon A)^{-1}(y+\varepsilon Bu)) \\ \qquad + W^\varepsilon(t-\varepsilon, (I+\varepsilon\partial\phi)^{-1}(I+\varepsilon A)^{-1}(y+\varepsilon Bu)) : u \in \mathcal{U}\}, \\ (t,y) \in (\varepsilon, T] \times \mathcal{H}, \\ \inf\{th(u) + \varepsilon g((I+\varepsilon\partial\phi)^{-1}(I+\varepsilon A)^{-1}(y+tBu)) \\ \qquad + l((I+\varepsilon\partial\phi)^{-1}(I+\varepsilon A)^{-1}(y+tBu)) : u \in \mathcal{U}\}, \\ (t,y) \in (0,\varepsilon] \times \mathcal{H}, \end{cases} \tag{2.4}$$

$$W^\varepsilon(0,y) = l(y), \qquad y \in \mathcal{H}.$$

This scheme gives the same values as the preceding scheme at $t = i\varepsilon : V_\varepsilon(i\varepsilon, y) = V^\varepsilon(i\varepsilon, y)$. As we shall see in §6, only these values will matter in expressing the discrete feedback law.

We obtain the following alternative scheme by interchanging the resolvents of A and $\partial\phi$ in (2.4):

$$\overline{W}^\varepsilon(t,y) = \begin{cases} \inf\{\varepsilon h(u) + \varepsilon g((I+\varepsilon A)^{-1}(I+\varepsilon\partial\phi)^{-1}(y+\varepsilon Bu)) \\ \qquad + \overline{W}^\varepsilon(t-\varepsilon, (I+\varepsilon A)^{-1}(I+\varepsilon\partial\phi)^{-1}(y+\varepsilon Bu)) : u \in \mathcal{U}\}, \\ (t,y) \in (\varepsilon, T] \times \mathcal{H}, \\ \inf\{th(u) + \varepsilon g((I+\varepsilon A)^{-1}(I+\varepsilon\partial\phi)^{-1}(y+tBu)) \\ \qquad + l((I+\varepsilon A)^{-1}(I+\varepsilon\partial\phi)^{-1}(y+tBu)) : u \in \mathcal{U}\}, \\ (t,y) \in (0,\varepsilon] \times \mathcal{H}, \end{cases} \tag{2.5}$$

$$\overline{W}^\varepsilon(0,y) = l(y), \qquad y \in \mathcal{H}.$$

3. Convergence of Trotter product formulae. As we already mentioned in the introduction, the following two results represent generalizations of those in [11]. As a matter of fact, we shall show that the arguments from [11] also work in the present, more general framework.

THEOREM 3.1. *Under hypotheses* (i) – (v), *suppose that the following additional assumption holds*:

(vi) $(I+\varepsilon\partial\phi)^{-1}$ *maps* $\mathcal{V}$ *into itself and*

$$(A(I+\varepsilon\partial\phi)^{-1}z, (I+\varepsilon\partial\phi)^{-1}z) \leq (Az, z) \quad \textit{for all } z \in \mathcal{V}. \tag{3.1}$$

Then for every $(t,y) \in [0,T] \times \overline{D(A) \cap D(\partial\phi)}$, *we have*

$$\lim_{\varepsilon\to 0} W^\varepsilon(t,y) = W(t,y). \tag{3.2}$$

Proof. Although the proof is very similar to that of Theorem 1 in [11], for the sake of completeness (and for the reader's convenience) we outline it, emphasizing the specific points of the present situation.

Recall that $\varepsilon = T/N$. Fix $(t,y) \in [0,T] \times \overline{D(A) \cup D(\partial\phi)}$. Let us suppose that t/ε is not an integer. The successive application of (2.4) on the intervals $(t-\varepsilon, t], (t-2\varepsilon, t-\varepsilon], \dots, (t-[t/\varepsilon]\varepsilon, t-([t/\varepsilon]-1)\varepsilon], (0, t-[t/\varepsilon]\varepsilon]$ leads (after some convenient rewriting) to the following representation for W^ε:

$$\begin{aligned} W^\varepsilon(t,y) = \inf\Bigg\{ & \sum_{i=1}^{[t/\varepsilon]} \varepsilon h(u(i\varepsilon)) + \left(t - \left[\frac{t}{\varepsilon}\right]\varepsilon\right) h\left(u\left(\left(\left[\frac{t}{\varepsilon}\right]+1\right)\varepsilon\right)\right) \\ & + \sum_{i=1}^{[t/\varepsilon]} \varepsilon g(y_u^\varepsilon(i\varepsilon)) + \varepsilon g(y_u^\varepsilon(t)) \\ & + l(y_u^\varepsilon(t)) : u \text{ is a step function from } [0,T] \text{ to } \mathcal{U} \Bigg\}, \end{aligned} \tag{3.3}$$

where for any (step) function $u : [0,T] \to \mathcal{U}$, y_u^ε is defined by

$$\begin{cases} y_u^\varepsilon(t) = (I+\varepsilon\partial\phi)^{-1}(I+\varepsilon A)^{-1}(y_u^\varepsilon((i-1)\varepsilon) + (t-(i-1)\varepsilon)Bu(i\varepsilon)) \\ \qquad \text{for } t \in ((i-1)\varepsilon, i\varepsilon], \quad i = 1,2,\dots,N, \\ y_u^\varepsilon(0) = y. \end{cases} \tag{3.4}$$

If t/ε is an integer, then in (3.3) we must eliminate the two terms containing the values of h and g which are not under the sum signs.

The statement of the theorem now becomes more transparent if we look at (3.3) and recognize in (3.4) a Trotter scheme for (1.2) with the initial state y. So, the following lemma represents an important stage in the proof.

Let u be a step function from $[0,T]$ to $\mathcal{U}$ which takes constant values on $(0,\varepsilon_0], (\varepsilon_0, 2\varepsilon_0], \dots, ((N_0-1)\varepsilon_0, N_0\varepsilon_0] = (T-\varepsilon_0, T]$. Denote by y_u the unique weak solution of the Cauchy problem

$$\begin{cases} y_u' + Ay_u + \partial\phi(y_u) \ni Bu, \\ y_u(0) = y. \end{cases}$$

LEMMA 3.1. *For every* $t \in [0,T]$, *we have*

$$\lim_{\varepsilon\to 0} y_u^\varepsilon(t) = y_u(t) \quad \textit{strongly in } \mathcal{H}.$$

The main tool in the proof of Lemma 3.1 is the nonlinear version of the Chernoff formula (see [11, Lem. 2]).

Using Lemma 3.1, we obtain without serious difficulties (see also [11, Rem. 1])

$$\limsup_{\varepsilon\to 0} W^\varepsilon(t,y) \le W(t,y).$$

The other inequality, i.e.,

$$\liminf_{\varepsilon\to 0} W^\varepsilon(t,y) \ge W(t,y),$$

is somewhat more difficult. We shall treat it in detail.

For any step function $u : [0,T] \to \mathcal{U}$, denote

$$W^\varepsilon(t,y;u) = \begin{cases} \displaystyle\sum_{i=1}^{[t/\varepsilon]} \varepsilon h(u(i\varepsilon)) + \left(t - \left[\frac{t}{\varepsilon}\right]\varepsilon\right) h\left(u\left(\left(\left[\frac{t}{\varepsilon}\right]+1\right)\varepsilon\right)\right) \\ \qquad\qquad\qquad \displaystyle + \sum_{i=1}^{[t/\varepsilon]} \varepsilon g(y^\varepsilon_u(i\varepsilon)) \\ \qquad\qquad\qquad + \ \varepsilon g(y^\varepsilon_u(t)) + l(y^\varepsilon_u(t)) \quad \text{if } t/\varepsilon \text{ is not an integer,} \\ \displaystyle\sum_{i=1}^{[t/\varepsilon]} \varepsilon(h(u(i\varepsilon)) + g(y^\varepsilon_u(i\varepsilon))) + l(y^\varepsilon_u(t)) \quad \text{otherwise.} \end{cases}$$

Clearly, we may select a subsequence of $\{\varepsilon\}$, also denoted $\{\varepsilon\}$, and a corresponding sequence of step functions u^ε, each u^ε taking constant values on $(0,\varepsilon], (\varepsilon, 2\varepsilon], \ldots, (T-\varepsilon, T]$, such that

$$\lim_{\varepsilon\to 0} W^\varepsilon(t,y;u^\varepsilon) = \liminf_{\varepsilon\to 0} W^\varepsilon(t,y) \le \limsup_{\varepsilon\to 0} W^\varepsilon(t,y) \le W(t,y).$$

Hence, by using hypotheses (i) (actually (1.4)) and (ii) together with the estimate

$$|y^\varepsilon_{u^\varepsilon}(s)| \le |((I+\varepsilon\partial\phi)^{-1}(I+\varepsilon A)^{-1})^{[s/\varepsilon]+1} y| + \int_0^s |Bu^\varepsilon(\tau)|d\tau,$$

it follows that $\{u^\varepsilon\}$ is bounded in $L^2(0,t;\mathcal{U})$; consequently, on a subsequence, $u^\varepsilon \to u^0$ weakly in $L^2(0,t;\mathcal{U})$ as $\varepsilon \to 0$.

LEMMA 3.2. *Let $\{u^\varepsilon\}$ be a sequence of step functions as above and $y \in D(A) \cap D(\partial\phi)$. If $u^\varepsilon \to u^0$ weakly in $L^2(0,t;\mathcal{U})$, then for every $s \in [0,t]$,*

$$y^\varepsilon_{u^\varepsilon}(s) \to y_{u^0}(s) \quad \textit{strongly in } \mathcal{H}.$$

Proof. First we shall show that $y^\varepsilon_{u^\varepsilon}(s)$ converges strongly in $\mathcal{H}$ (possible on a subsequence) for every $s \in [0,t]$. To this end we introduce the auxiliary function $z^\varepsilon_{u^\varepsilon}$ given by

$$z^\varepsilon_{u^\varepsilon}(s) = (I+\varepsilon A)^{-1}(y^\varepsilon_{u^\varepsilon}((i-1)\varepsilon) + (s-(i-1)\varepsilon)Bu^\varepsilon(i\varepsilon)) \quad \text{for } s \in ((i-1)\varepsilon, i\varepsilon], \\ i = 1,2,\ldots,N.$$

Obviously, $y^\varepsilon_{u^\varepsilon}(s) = (I+\varepsilon\partial\phi)^{-1} z^\varepsilon_{u^\varepsilon}(s)$. For simplicity we set $z^\varepsilon_{u^\varepsilon} = z^\varepsilon$ and $y^\varepsilon_{u^\varepsilon} = y^\varepsilon$. One easily verifies that z^ε satisfies the following difference scheme:

$$\frac{1}{\varepsilon}((I+\varepsilon\partial\phi)^{-1}z^\varepsilon(i\varepsilon) - (I+\varepsilon\partial\phi)^{-1}z^\varepsilon((i-1)\varepsilon)) \\ + Az^\varepsilon(i\varepsilon) + \partial\phi_\varepsilon(z^\varepsilon(i\varepsilon)) = Bu^\varepsilon(i\varepsilon), i = 2,3,\ldots,\left[\frac{s}{\varepsilon}\right], \qquad s \in [0,t].$$

Here ϕ_ε is the convex regularization of ϕ. (Recall that $(\partial\phi)_\varepsilon = \partial\phi_\varepsilon$.) Multiplying scalarly in $\mathcal{H}$ by $z^\varepsilon(i\varepsilon) - z^\varepsilon((i-1)\varepsilon)$, we obtain after some calculation

$$\begin{aligned} &\frac{1}{\varepsilon}|y^\varepsilon(i\varepsilon) - y^\varepsilon((i-1)\varepsilon)|^2 + \frac{1}{2}(Az^\varepsilon(i\varepsilon), z^\varepsilon(i\varepsilon)) - \frac{1}{2}(Az^\varepsilon((i-1)\varepsilon), z^\varepsilon((i-1)\varepsilon)) \\ &\quad + \phi_\varepsilon(z^\varepsilon(i\varepsilon)) - \phi_\varepsilon(z^\varepsilon((i-1)\varepsilon)) \\ &\le \frac{1}{2\varepsilon}|y^\varepsilon((i-1)\varepsilon) - y^\varepsilon((i-2)\varepsilon)|^2 + 2\varepsilon|Bu^\varepsilon(i\varepsilon)|^2 + \frac{\varepsilon}{2}|Bu^\varepsilon((i-1)\varepsilon)|^2, \\ &\qquad\qquad i = 2,3,\ldots,\left[\frac{s}{\varepsilon}\right]. \end{aligned} \tag{3.5}$$

If s/ε is not an integer, we have as above

$$\begin{aligned}&\frac{1}{\varepsilon}\left|y^\varepsilon(s)-y^\varepsilon\left(\left[\frac{s}{\varepsilon}\right]\varepsilon\right)\right|^2+\frac{1}{2}(Az^\varepsilon(s),z^\varepsilon(s))-\frac{1}{2}\left(Az^\varepsilon\left(\left[\frac{s}{\varepsilon}\right]\varepsilon\right),z^\varepsilon\left(\left[\frac{s}{\varepsilon}\right]\varepsilon\right)\right)\\&\quad+\phi_\varepsilon(z^\varepsilon(s))-\phi_\varepsilon\left(z^\varepsilon\left(\left[\frac{s}{\varepsilon}\right]\varepsilon\right)\right)\\&\quad\le\frac{1}{2\varepsilon}\left|y^\varepsilon\left(\left[\frac{s}{\varepsilon}\right]\varepsilon\right)-y^\varepsilon\left(\left(\left[\frac{s}{\varepsilon}\right]-1\right)\varepsilon\right)\right|^2\\&\qquad+2\left(s-\left[\frac{s}{\varepsilon}\right]\varepsilon\right)\left|Bu^\varepsilon\left(\left(\left[\frac{s}{\varepsilon}\right]+1\right)\varepsilon\right)\right|^2+\frac{\varepsilon}{2}\left|Bu^\varepsilon\left(\left[\frac{s}{\varepsilon}\right]\varepsilon\right)\right|^2.\end{aligned}\tag{3.6}$$

Adding inequalities (3.5) ($i=2,3,\dots,[s/\varepsilon]$) and (3.6), we have

$$\begin{aligned}&\sum_{i=2}^{[s/\varepsilon]}\frac{1}{2\varepsilon}|y^\varepsilon(i\varepsilon)-y^\varepsilon((i-1)\varepsilon)|^2+\frac{1}{\varepsilon}\left|y^\varepsilon(s)-y^\varepsilon\left(\left[\frac{s}{\varepsilon}\right]\varepsilon\right)\right|^2+\frac{1}{2}(Az^\varepsilon(s),z^\varepsilon(s))\\&\quad-\frac{1}{2}(Az^\varepsilon(\varepsilon),z^\varepsilon(\varepsilon))+\phi_\varepsilon(z^\varepsilon(s))-\phi_\varepsilon(z^\varepsilon(\varepsilon))\\&\quad\le\frac{1}{2\varepsilon}|y^\varepsilon(\varepsilon)-y|^2+\frac{1}{2}\varepsilon|Bu^\varepsilon(\varepsilon)|^2+\frac{5}{2}\varepsilon\sum_{i=2}^{[s/\varepsilon]}|Bu^\varepsilon(i\varepsilon)|^2\\&\qquad+2\left(s-\left[\frac{s}{\varepsilon}\right]\varepsilon\right)\left|Bu^\varepsilon\left(\left(\left[\frac{s}{\varepsilon}\right]+1\right)\varepsilon\right)\right|^2.\end{aligned}\tag{3.7}$$

Using the definition of $y^\varepsilon(\varepsilon)$ and the fact that $(I+\varepsilon\partial\phi)^{-1}$ and $(I+\varepsilon A)^{-1}$ are contractions, we derive

$$\frac{1}{\varepsilon}|y^\varepsilon(\varepsilon)-y|^2\le 3\varepsilon|Bu^\varepsilon(\varepsilon)|^2+3\varepsilon|Ay|^2+3\varepsilon|\partial\phi(y)|^2,\tag{3.8}$$

since $y\in D(A)\cap D(\partial\phi)$. Next we have

$$z^\varepsilon(\varepsilon)-y+\varepsilon Az^\varepsilon(\varepsilon)=\varepsilon Bu^\varepsilon(\varepsilon),$$

which, after a scalar multiplication by $z^\varepsilon(\varepsilon)-y$ and a suitable estimate of the right-hand side, yields

$$\frac{1}{2}(Az^\varepsilon(\varepsilon),z^\varepsilon(\varepsilon))-\frac{1}{2}(Ay,y)\le\frac{\varepsilon}{4}|Bu^\varepsilon(\varepsilon)|^2.\tag{3.9}$$

Further,

$$\phi_\varepsilon(z^\varepsilon(s))\ge\phi((I+\varepsilon\partial\phi)^{-1}z^\varepsilon(s))=\phi(y^\varepsilon(s)).\tag{3.10}$$

Regarding $\phi_\varepsilon(z^\varepsilon(\varepsilon))$, by the definition of ϕ_ε (see [6, p. 121]), we have

$$\phi_\varepsilon(z^\varepsilon(\varepsilon))\le\frac{1}{2\varepsilon}|z^\varepsilon(\varepsilon)-y|^2+\phi(y),$$

whence

$$\phi_\varepsilon(z^\varepsilon(\varepsilon))\le\varepsilon|Bu^\varepsilon(\varepsilon)|^2+\varepsilon|Ay|^2+\phi(y).\tag{3.11}$$

Now, adding (3.8), (3.9), and (3.11) to (3.7), and taking (3.10) into account as well, we obtain

$$\sum_{i=1}^{[s/\varepsilon]} \frac{1}{\varepsilon}|y^\varepsilon(i\varepsilon) - y^\varepsilon((i-1)\varepsilon)|^2 + \frac{1}{\varepsilon}\left|y^\varepsilon(s) - y^\varepsilon\left(\left[\frac{s}{\varepsilon}\right]\varepsilon\right)\right|^2 + (Az^\varepsilon(s), z^\varepsilon(s)) \leq \text{const.},$$

where the constant is independent of ε. Hence, on the one hand, using hypothesis (3.1) and condition (1.5), we ascertain that the sequence $\{y^\varepsilon(s)\}$ is bounded in $\mathcal{V}$ for any $s \in [0,t]$. On the other hand, taking $s = t$ in the above sum, we get

$$\sum_{i=1}^{[t/\varepsilon]} |y^\varepsilon(i\varepsilon) - y^\varepsilon((i-1)\varepsilon)| + \left|y^\varepsilon(t) - y^\varepsilon\left(\left[\frac{t}{\varepsilon}\right]\varepsilon\right)\right| \leq \text{const.}$$

But we also have (it is easy to verify)

$$|y^\varepsilon(s) - y^\varepsilon(s')| \leq |s - s'||Bu^\varepsilon(i\varepsilon)| \quad \text{for}\, s, s' \in ((i-1)\varepsilon, i\varepsilon], \quad i = 1, 2, \ldots, N.$$

Thus we have obtained that functions y^ε are of bounded variation on $[0,t]$ and $V_0^t y^\varepsilon \leq$ const. Since the inclusion $\mathcal{V} \subset \mathcal{H}$ is compact, by an infinite-dimensional version of the Helly theorem (see, for instance, [6, Rem. 3.2, p. 60]) we conclude that there exists a bounded variation function $y^0 : [0,t] \to \mathcal{H}$ such that on a subsequence, also denoted $\{y^\varepsilon\}$, we have

$$y^\varepsilon_{u^\varepsilon}(s) \to y^0(s) \quad \text{strongly in}\,\mathcal{H}\,\text{for all}\, s \in [0,t]. \tag{3.12}$$

We shall now prove that $y^0(s) = y_{u^0}(s)$ for $s \in [0,t]$. For any $\eta > 0$, choose a step function $u = u_\eta$ which takes constant values on $(0, \varepsilon_\eta], (\varepsilon_\eta, 2\varepsilon_\eta], \ldots, (T - \varepsilon_\eta, T]$ (where $\varepsilon_\eta = T/N_\eta$) such that

$$\left(\int_0^t |u_\eta(s) - u^0(s)|^2\, ds\right)^{1/2} < \eta \quad \text{and} \quad |y_{u_\eta}(s) - y_{u^0}(s)| < \eta \quad \text{for all}\, s \in [0,t]. \tag{3.13}$$

It is easy to see that $y^\varepsilon_{u^\varepsilon}$ and y^ε_u satisfy the following schemes:

$$\begin{aligned}&(I+\varepsilon A)^{-1} y^\varepsilon_{u^\varepsilon}(i\varepsilon) - (I+\varepsilon A)^{-1} y^\varepsilon_{u^\varepsilon}((i-1)\varepsilon) + \varepsilon A_\varepsilon y^\varepsilon_{u^\varepsilon}(i\varepsilon) + \varepsilon\partial\phi(y^\varepsilon_{u^\varepsilon}(i\varepsilon))\\ &\quad \ni \varepsilon(I+\varepsilon A)^{-1} Bu^\varepsilon(i\varepsilon),\\ &(I+\varepsilon A)^{-1} y^\varepsilon_u(i\varepsilon) - (I+\varepsilon A)^{-1} y^\varepsilon_u((i-1)\varepsilon) + \varepsilon A_\varepsilon y^\varepsilon_u(i\varepsilon) + \varepsilon\partial\phi(y^\varepsilon_u(i\varepsilon))\\ &\quad \ni \varepsilon(I+\varepsilon A)^{-1} Bu(i\varepsilon),\end{aligned}$$

$i = 1, 2, \ldots, [t/\varepsilon]$, A_ε being the Yosida approximation of A.

Let $s \in [0,t]$ such that $s \neq [s/\varepsilon]\varepsilon$. Subtract the preceding two inclusions and multiply the difference by $y^\varepsilon_{u^\varepsilon}(i\varepsilon) - y^\varepsilon_u(i\varepsilon)$. Then, using the monotonicity of A_ε, $\partial\phi$, and adding the obtained inequalities with respect to i from 1 to $[s/\varepsilon]$, we get

$$\begin{aligned}&\frac{1}{2}\left|(I+\varepsilon A)^{-1} y^\varepsilon_{u^\varepsilon}\left(\left[\frac{s}{\varepsilon}\right]\varepsilon\right) - (I+\varepsilon A)^{-1} y^\varepsilon_u\left(\left[\frac{s}{\varepsilon}\right]\varepsilon\right)\right|^2\\ &\quad \leq \sum_{i=1}^{[s/\varepsilon]} \varepsilon(Bu^\varepsilon(i\varepsilon) - Bu(i\varepsilon), (I+\varepsilon A)^{-1} y^\varepsilon_{u^\varepsilon}(i\varepsilon) - (I+\varepsilon A)^{-1} y^\varepsilon_u(i\varepsilon)).\end{aligned}$$

Hence

$$\begin{aligned}
&\frac{1}{2}|y^\varepsilon_{u^\varepsilon}(s)-y^\varepsilon_u(s)|^2\\
&\quad\le 3\sum_{i=1}^{[s/\varepsilon]}\varepsilon(u^\varepsilon(i\varepsilon)-u(i\varepsilon),B^*(I+\varepsilon A)^{-1}y^\varepsilon_{u^\varepsilon}(i\varepsilon)-B^*(I+\varepsilon A)^{-1}y^\varepsilon_u(i\varepsilon))\\
&\qquad+\frac{3}{2}\left(s-\left[\frac{s}{\varepsilon}\right]\varepsilon\right)^2\left|Bu^\varepsilon\left(\left(\left[\frac{s}{\varepsilon}\right]+1\right)\varepsilon\right)\right|^2\\
&\qquad+\frac{3}{2}\left(s-\left[\frac{s}{\varepsilon}\right]\varepsilon\right)^2\left|Bu\left(\left(\left[\frac{s}{\varepsilon}\right]+1\right)\varepsilon\right)\right|^2.
\end{aligned}\tag{3.14}$$

In the case when $s=[s/\varepsilon]\varepsilon$, we add the same inequalities as above but only from 1 to $[s/\varepsilon]-1$ (the computations being similar).

To conclude, we write the above sum as an integral:

$$\begin{aligned}
&\sum_{i=1}^{[s/\varepsilon]}\varepsilon(u^\varepsilon(i\varepsilon)-u(i\varepsilon),B^*(I+\varepsilon A)^{-1}y^\varepsilon_{u^\varepsilon}(i\varepsilon)-B^*(I+\varepsilon A)^{-1}y^\varepsilon_u(i\varepsilon))\\
&\quad=\int_0^{[s/\varepsilon]\varepsilon}(u^\varepsilon(\tau)-\tilde{u}_\varepsilon(\tau),B^*(I+\varepsilon A)^{-1}\tilde{y}^\varepsilon_{u^\varepsilon}(\tau)-B^*(I+\varepsilon A)^{-1}\tilde{y}^\varepsilon_u(\tau))d\tau,
\end{aligned}$$

where

$$\begin{gathered}
\tilde{y}^\varepsilon_{u^\varepsilon}(\tau)=y^\varepsilon_{u^\varepsilon}(i\varepsilon),\qquad \tilde{y}^\varepsilon_u(\tau)=y^\varepsilon_u(i\varepsilon),\quad\text{and}\quad \tilde{u}_\varepsilon(\tau)=u(i\varepsilon)\\
\text{for}\,\tau\in((i-1)\varepsilon,i\varepsilon],\quad i=1,2,\dots,\left[\frac{s}{\varepsilon}\right].
\end{gathered}$$

In addition to this, since $|\tilde{y}^\varepsilon_{u^\varepsilon}(\tau)-y^\varepsilon_{u^\varepsilon}(\tau)|\longrightarrow 0$ (because $\{u^\varepsilon\}$ is bounded in $L^2(0,t;\mathcal{U})$), $|\tilde{y}^\varepsilon_u(\tau)-y^\varepsilon_u(\tau)|\longrightarrow 0$, by (3.12) and Lemma 3.1 we have

$$\tilde{y}^\varepsilon_{u^\varepsilon}(\tau)\longrightarrow y^0(\tau)\text{ (on a subsequence)},\qquad \tilde{y}^\varepsilon_u(\tau)\longrightarrow y_u(\tau)\quad\text{strongly in }\mathcal{H}\text{ for all }\tau\in[0,s].$$

Also,

$$\tilde{u}_\varepsilon(\tau)\longrightarrow u(\tau)\quad\text{for}\,\tau\in[0,s]\text{ except the possible discontinuity points of }u.$$

Letting $\varepsilon\longrightarrow 0$ in (3.14), we obtain

$$\frac{1}{2}|y^0(s)-y_{u_\eta}(s)|^2\le 3\int_0^s(Bu^0(\tau)-Bu_\eta(\tau),y^0(\tau)-y_{u_\eta}(\tau))dt,$$

which together with (3.13) (η being arbitrary) gives $y^0\equiv y_{u^0}$, and the proof of Lemma 3.2 is finished.

Now it is easy to derive the assertion of Theorem 3.1 for $y\in D(A)\cap D(\partial\phi)$. Indeed, we may write

$$W^\varepsilon(t,y;u^\varepsilon)=\int_0^t h(u^\varepsilon(s))ds+\int_0^{[t/\varepsilon]\varepsilon}g(\tilde{y}^\varepsilon_{u^\varepsilon}(s))ds+\varepsilon g(y^\varepsilon_{u^\varepsilon}(t))+l(y^\varepsilon_{u^\varepsilon}(t)),$$

where $\tilde{y}^\varepsilon_{u^\varepsilon}$ has been defined above. Hence, using the weak lower semicontinuity in $L^2(0,t;\mathcal{U})$ of $u\mapsto\int_0^t h(u(s))ds$ and the Lebesgue dominated convergence theorem combined with Lemma 3.2, we see that

$$\lim_{\varepsilon\to 0}W^\varepsilon(t,y;u^\varepsilon)\ge W(t,y).$$

To extend this conclusion to the case when $y \in \overline{D(A) \cap D(\partial\phi)}$, we need the following continuity result.

LEMMA 3.3. *Under the hypotheses of Theorem* 3.1, *for any* $t \in [0,T]$, *the function* $\overline{D(A) \cap D(\partial\phi)} \ni y \mapsto W(t,y)$ *is Lipschitz continuous on bounded subsets, and the functions* $\mathcal{H} \ni y \mapsto W^\varepsilon(t,y)$ *are Lipschitz continuous on bounded subsets, uniformly with respect to* ε.

Proof. We shall sketch only the proof of the second part of the lemma (the proof of the first is similar).

Consider y, $z \in \mathcal{H}$, arbitrary, such that $|y|, |z| \le r$. Let $\eta > 0$, also arbitrary. We choose a step function $u^\varepsilon = u^\varepsilon_\eta$ such that

$$\begin{aligned} &W^\varepsilon(t,z) - W^\varepsilon(t,y) \\ &\quad \le \sum_{i=1}^{[t/\varepsilon]} \varepsilon(g(z^\varepsilon_{u^\varepsilon}(i\varepsilon)) - g(y^\varepsilon_{u^\varepsilon}(i\varepsilon))) + \varepsilon(g(z^\varepsilon_{u^\varepsilon}(t)) - g(y^\varepsilon_{u^\varepsilon}(t))) \\ &\quad + l(z^\varepsilon_{u^\varepsilon}(t)) - l(y^\varepsilon_{u^\varepsilon}(t)) + \eta. \end{aligned}$$

Here $z^\varepsilon_{u^\varepsilon}$ is defined as in (3.4) but with z instead of y.

Taking (1.4) into account, we easily obtain

$$\sum_{i=1}^{[t/\varepsilon]} \varepsilon|u^\varepsilon_\eta(i\varepsilon)|^2 + \left(t - \left[\frac{t}{\varepsilon}\right]\varepsilon\right)\left|u^\varepsilon_\eta\left(\left(\left[\frac{t}{\varepsilon}\right]+1\right)\varepsilon\right)\right|^2 \le \text{const.},$$

where the constant depends only on r. Consequently, $y^\varepsilon_{u^\varepsilon}(i\varepsilon), y^\varepsilon_{u^\varepsilon}(t), z^\varepsilon_{u^\varepsilon}(i\varepsilon)$, and $z^\varepsilon_{u^\varepsilon}(t)$ also are bounded by a constant which depends only on r. We may now use assumption (ii) to derive the Lipschitz continuity on bounded subsets of $y \mapsto W^\varepsilon(t,y)$, uniformly with respect to ε. This completes the proof of Lemma 3.3.

A density argument based on Lemma 3.3 finishes the proof of Theorem 3.1.

Let us now establish the convergence of scheme (2.5).

THEOREM 3.2. *In addition to assumptions* (i)–(v) *we suppose that at least one of the following two hypotheses holds*:

(vii) $(I + \varepsilon\partial\phi)^{-1} = P$ *for all* $\varepsilon > 0$ (*i.e., the resolvent of* ϕ *is independent of* ε);

(viii) $(I + \varepsilon A)^{-1}D(\partial\phi) \subset \overline{D(\partial\phi)}$ *for all* $\varepsilon > 0$.

Then for every $(t,y) \in [0,T] \times \overline{D(A) \cap D(\partial\phi)}$, *we have*

$$\lim_{\varepsilon\to 0} \overline{W}^\varepsilon(t,y) = W(t,y). \tag{3.15}$$

The proof of Theorem 3.2 is very similar to that of the preceding theorem. The only differences appear in the proof of the corresponding variant of Lemma 3.2.

In the present case, for a (step) function $u : [0,T] \to \mathcal{U}$, y^ε_u is defined by

$$\begin{cases} y^\varepsilon_u(t) = (I+\varepsilon A)^{-1}(I+\varepsilon\partial\phi)^{-1}(y^\varepsilon_u((i-1)\varepsilon) + (t-(i-1)\varepsilon)Bu(i\varepsilon)) \\ \qquad\qquad \text{for } t \in ((i-1)\varepsilon, i\varepsilon], \quad i = 1,2,\ldots,N, \\ y^\varepsilon_u(0) = y. \end{cases}$$

LEMMA 3.4. *Let* $\{u^\varepsilon\}$ *be a sequence of step functions from* $[0,T]$ *to* $\mathcal{U}$ *such that each* u^ε *takes constant values on* $(0,\varepsilon], (\varepsilon, 2\varepsilon], \ldots, (T-\varepsilon, T]$, *and let* $y \in D(A) \cap D(\partial\phi)$.

If $u^\varepsilon \to u^0$ *weakly in* $L^2(0,t;\mathcal{U})$, *then for every* $s \in [0,t]$,

$$y^\varepsilon_{u^\varepsilon}(s) \to y_{u^0}(s) \quad \textit{strongly in } \mathcal{H}.$$

Proof. Set $y^\varepsilon_{u^\varepsilon} = y^\varepsilon$. As in the proof of Lemma 3.2, we have

$$\begin{aligned}&\frac{1}{\varepsilon}|(I+\varepsilon\partial\phi)^{-1}y^\varepsilon(i\varepsilon) - (I+\varepsilon\partial\phi)^{-1}y^\varepsilon((i-1)\varepsilon)|^2 + \frac{1}{2}(Ay^\varepsilon(i\varepsilon), y^\varepsilon(i\varepsilon))\\ &\quad - \frac{1}{2}(Ay^\varepsilon((i-1)\varepsilon), y^\varepsilon((i-1)\varepsilon)) + \phi_\varepsilon(y^\varepsilon(i\varepsilon)) - \phi_\varepsilon(y^\varepsilon((i-1)\varepsilon))\\ &\quad \le |Bu^\varepsilon(i\varepsilon)||y^\varepsilon(i\varepsilon) - y^\varepsilon((i-1)\varepsilon)|, \qquad i = 1,2,\ldots,\left[\frac{s}{\varepsilon}\right], \quad s\in[0,t].\end{aligned}$$

To eliminate the operator $(I+\varepsilon\partial\phi)^{-1}$ above, we use the following inequalities (easy to prove):

$$\begin{aligned}|y^\varepsilon((i+1)\varepsilon) - y^\varepsilon(i\varepsilon)|^2 &\le 3|(I+\varepsilon\partial\phi)^{-1}y^\varepsilon(i\varepsilon) - (I+\varepsilon\partial\phi)^{-1}y^\varepsilon((i-1)\varepsilon)|^2\\ &\quad + 3\varepsilon^2|Bu^\varepsilon((i+1)\varepsilon)|^2 + 3\varepsilon^2|Bu^\varepsilon(i\varepsilon)|^2,\\ &\quad i = 1,2,\ldots,\left[\frac{s}{\varepsilon}\right]-1.\end{aligned}$$

We obtain

$$\begin{aligned}&\frac{1}{3\varepsilon}|y^\varepsilon((i+1)\varepsilon) - y^\varepsilon(i\varepsilon)|^2 + \frac{1}{2}(Ay^\varepsilon(i\varepsilon), y^\varepsilon(i\varepsilon))\\ &\quad - \frac{1}{2}(Ay^\varepsilon((i-1)\varepsilon), y^\varepsilon((i-1)\varepsilon)) + \phi_\varepsilon(y^\varepsilon(i\varepsilon)) - \phi_\varepsilon(y^\varepsilon((i-1)\varepsilon))\\ &\quad \le |Bu^\varepsilon(i\varepsilon)||y^\varepsilon(i\varepsilon) - y^\varepsilon((i-1)\varepsilon)| + \varepsilon|Bu^\varepsilon((i+1)\varepsilon)|^2 + \varepsilon|Bu^\varepsilon(i\varepsilon)|^2,\\ &\qquad i = 1,2,\ldots,\left[\frac{s}{\varepsilon}\right]-1.\end{aligned}\tag{3.16}$$

Similarly, in the case when s/ε is not an integer, we have

$$\begin{aligned}&\frac{1}{3\varepsilon}\left|y^\varepsilon(s) - y^\varepsilon\left(\left[\frac{s}{\varepsilon}\right]\varepsilon\right)\right|^2 + \frac{1}{2}\left(Ay^\varepsilon\left(\left[\frac{s}{\varepsilon}\right]\varepsilon\right), y^\varepsilon\left(\left[\frac{s}{\varepsilon}\right]\varepsilon\right)\right)\\ &\quad - \frac{1}{2}\left(Ay^\varepsilon\left(\left(\left[\frac{s}{\varepsilon}\right]-1\right)\varepsilon\right), y^\varepsilon\left(\left(\left[\frac{s}{\varepsilon}\right]-1\right)\varepsilon\right)\right) + \phi_\varepsilon\left(y^\varepsilon\left(\left[\frac{s}{\varepsilon}\right]\varepsilon\right)\right)\\ &\quad - \phi_\varepsilon\left(y^\varepsilon\left(\left(\left[\frac{s}{\varepsilon}\right]-1\right)\varepsilon\right)\right)\\ &\quad \le \left|Bu^\varepsilon\left(\left[\frac{s}{\varepsilon}\right]\varepsilon\right)\right|\left|y^\varepsilon\left(\left[\frac{s}{\varepsilon}\right]\varepsilon\right) - y^\varepsilon\left(\left(\left[\frac{s}{\varepsilon}\right]-1\right)\varepsilon\right)\right|\\ &\qquad + \left(s - \left[\frac{s}{\varepsilon}\right]\varepsilon\right)\left|Bu^\varepsilon\left(\left(\left[\frac{s}{\varepsilon}\right]+1\right)\varepsilon\right)\right|^2 + \varepsilon\left|Bu^\varepsilon\left(\left[\frac{s}{\varepsilon}\right]\varepsilon\right)\right|^2,\end{aligned}\tag{3.17}$$

and also

$$\begin{aligned}&\frac{1}{2}(Ay^\varepsilon(s), y^\varepsilon(s)) - \frac{1}{2}\left(Ay^\varepsilon\left(\left[\frac{s}{\varepsilon}\right]\varepsilon\right), y^\varepsilon\left(\left[\frac{s}{\varepsilon}\right]\varepsilon\right)\right)\\ &\quad + \phi_\varepsilon(y^\varepsilon(s)) - \phi_\varepsilon\left(y^\varepsilon\left(\left[\frac{s}{\varepsilon}\right]\varepsilon\right)\right)\\ &\quad \le \left(\frac{s}{\varepsilon} - \left[\frac{s}{\varepsilon}\right]\right)\left|Bu^\varepsilon\left(\left(\left[\frac{s}{\varepsilon}\right]+1\right)\varepsilon\right)\right|\left|y^\varepsilon(s) - y^\varepsilon\left(\left[\frac{s}{\varepsilon}\right]\varepsilon\right)\right|.\end{aligned}\tag{3.18}$$

Adding inequalities (3.16) – (3.18) and (3.8) (the last multiplied by 1/3), we get

$$\begin{aligned}&\sum_{i=1}^{[s/\varepsilon]}\frac{1}{12\varepsilon}|y^\varepsilon(i\varepsilon) - y^\varepsilon((i-1)\varepsilon)|^2 + \frac{1}{12\varepsilon}\left|y^\varepsilon(s) - y^\varepsilon\left(\left[\frac{s}{\varepsilon}\right]\varepsilon\right)\right|^2\\ &\quad + \frac{1}{2}(Ay^\varepsilon(s), y^\varepsilon(s)) - \frac{1}{2}(Ay, y) + \phi_\varepsilon(y^\varepsilon(s)) - \phi_\varepsilon(y)\\ &\quad \le \varepsilon|Ay|^2 + \varepsilon|\partial\phi(y)|^2 + 3\sum_{i=1}^{[s/\varepsilon]}\varepsilon|Bu^\varepsilon(i\varepsilon)|^2 + 2\left(s - \left[\frac{s}{\varepsilon}\right]\varepsilon\right)\left|Bu^\varepsilon\left(\left(\left[\frac{s}{\varepsilon}\right]+1\right)\varepsilon\right)\right|^2.\end{aligned}$$

Then, proceeding as in the proof of Lemma 3.2, by the same infinite-dimensional variant of the Helly theorem we find a bounded variation function $y^0 : [0,t] \to \mathcal{H}$ such that, on a subsequence,

$$y^\varepsilon_{u^\varepsilon}(s) \to y^0(s) \quad \text{strongly in } \mathcal{H} \text{ for all } s \in [0,t]. \tag{3.19}$$

Next, for $\eta > 0$, arbitrary, we select a step function $u = u_\eta$ which takes constant values on the subintervals $((i-1)\varepsilon_\eta, i\varepsilon_\eta]$, $i = 1,2,\dots,N_\eta$, and satisfies (3.13). Define the function $z^\varepsilon_{u^\varepsilon}$ by

$$\begin{gathered} z^\varepsilon_{u^\varepsilon}(s) = (I + \varepsilon\partial\phi)^{-1}(y^\varepsilon_{u^\varepsilon}((i-1)\varepsilon) + (s-(i-1)\varepsilon)Bu^\varepsilon(i\varepsilon)) \\ \text{for } s \in ((i-1)\varepsilon, i\varepsilon], \quad i = 1,2,\dots,N, \end{gathered}$$

and define z^ε_u in the same way, but with u replacing u^ε. Obviously, $y^\varepsilon_{u^\varepsilon}(s) = (I+\varepsilon A)^{-1} z^\varepsilon_{u^\varepsilon}(s)$ and $y^\varepsilon_u(s) = (I+\varepsilon A)^{-1} z^\varepsilon_u(s)$, $s \in (0,T]$. We have

$$\begin{aligned} &y^\varepsilon_{u^\varepsilon}(i\varepsilon) - y^\varepsilon_{u^\varepsilon}((i-1)\varepsilon) + \varepsilon A_\varepsilon z^\varepsilon_{u^\varepsilon}(i\varepsilon) + \varepsilon\partial\phi(z^\varepsilon_{u^\varepsilon}(i\varepsilon)) \ni \varepsilon Bu^\varepsilon(i\varepsilon), \\ &y^\varepsilon_u(i\varepsilon) - y^\varepsilon_u((i-1)\varepsilon) + \varepsilon A_\varepsilon z^\varepsilon_u(i\varepsilon) + \varepsilon\partial\phi(z^\varepsilon_u(i\varepsilon)) \ni \varepsilon Bu(i\varepsilon), \qquad i = 2,3,\dots,\left[\frac{s}{\varepsilon}\right]. \end{aligned}$$

Multiplying the difference of these inclusions by $z^\varepsilon_{u^\varepsilon}(i\varepsilon) - z^\varepsilon_u(i\varepsilon)$, we obtain

$$\begin{aligned} &\frac{1}{2}(y^\varepsilon_{u^\varepsilon}(i\varepsilon) - y^\varepsilon_u(i\varepsilon), z^\varepsilon_{u^\varepsilon}(i\varepsilon) - z^\varepsilon_u(i\varepsilon)) \\ &\quad - \frac{1}{2}(y^\varepsilon_{u^\varepsilon}((i-1)\varepsilon) - y^\varepsilon_u((i-1)\varepsilon), z^\varepsilon_{u^\varepsilon}((i-1)\varepsilon) - z^\varepsilon_u((i-1)\varepsilon)) \\ &\quad \le \varepsilon(Bu^\varepsilon(i\varepsilon) - Bu(i\varepsilon), z^\varepsilon_{u^\varepsilon}(i\varepsilon) - z^\varepsilon_u(i\varepsilon)), \qquad i = 2,3,\dots,\left[\frac{s}{\varepsilon}\right]. \end{aligned} \tag{3.20}$$

Similarly, in the case when s/ε is not an integer,

$$\begin{aligned} &\frac{1}{2}(y^\varepsilon_{u^\varepsilon}(s) - y^\varepsilon_u(s), z^\varepsilon_{u^\varepsilon}(s) - z^\varepsilon_u(s)) \\ &\quad - \frac{1}{2}\left(y^\varepsilon_{u^\varepsilon}\left(\left[\frac{s}{\varepsilon}\right]\varepsilon\right) - y^\varepsilon_u\left(\left[\frac{s}{\varepsilon}\right]\varepsilon\right), z^\varepsilon_{u^\varepsilon}\left(\left[\frac{s}{\varepsilon}\right]\varepsilon\right) - z^\varepsilon_u\left(\left[\frac{s}{\varepsilon}\right]\varepsilon\right)\right) \\ &\quad \le \left(s - \left[\frac{s}{\varepsilon}\right]\varepsilon\right)\left(Bu^\varepsilon\left(\left(\left[\frac{s}{\varepsilon}\right]+1\right)\varepsilon\right)\right. \\ &\qquad \left. -Bu\left(\left(\left[\frac{s}{\varepsilon}\right]+1\right)\varepsilon\right), z^\varepsilon_{u^\varepsilon}(s) - z^\varepsilon_u(s)\right). \end{aligned} \tag{3.21}$$

The summation of (3.20) and (3.21) gives

$$\begin{aligned} &\frac{1}{2}|y^\varepsilon_{u^\varepsilon}(s) - y^\varepsilon_u(s)|^2 - \frac{1}{2}(y^\varepsilon_{u^\varepsilon}(\varepsilon) - y^\varepsilon_u(\varepsilon), z^\varepsilon_{u^\varepsilon}(\varepsilon) - z^\varepsilon_u(\varepsilon)) \\ &\quad \le \sum_{i=2}^{[s/\varepsilon]} \varepsilon(u^\varepsilon(i\varepsilon) - u(i\varepsilon), B^*(z^\varepsilon_{u^\varepsilon}(i\varepsilon) - z^\varepsilon_u(i\varepsilon))) \\ &\qquad + \left(s - \left[\frac{s}{\varepsilon}\right]\varepsilon\right)\left(u^\varepsilon\left(\left(\left[\frac{s}{\varepsilon}\right]+1\right)\varepsilon\right)\right. \\ &\qquad \left. - u\left(\left(\left[\frac{s}{\varepsilon}\right]+1\right)\varepsilon\right), B^*(z^\varepsilon_{u^\varepsilon}(s) - z^\varepsilon_u(s))\right). \end{aligned} \tag{3.22}$$

(If s/ε is an integer, we add only inequalities (3.20) to obtain a similar conclusion.)

But we may write (3.22) in the following form:

(3.23)

$$\begin{aligned}\frac{1}{2}|y^\varepsilon_{u^\varepsilon}(s) - y^\varepsilon_u(s)|^2 \leq{}& \frac{1}{2}(y^\varepsilon_{u^\varepsilon}(\varepsilon) - y^\varepsilon_u(\varepsilon), z^\varepsilon_{u^\varepsilon}(\varepsilon) - z^\varepsilon_u(\varepsilon))\\ &+ \int_0^{([s/\varepsilon]-1)\varepsilon} (u^\varepsilon(\tau+\varepsilon) - \tilde{u}_\varepsilon(\tau), B^*(\tilde{z}^\varepsilon_{u^\varepsilon}(\tau) - \tilde{z}^\varepsilon_u(\tau)))d\tau\\ &+ \varepsilon\Big(u^\varepsilon\Big(\Big(\Big[\frac{s}{\varepsilon}\Big]+1\Big)\varepsilon\Big) - u\Big(\Big(\Big[\frac{s}{\varepsilon}\Big]+1\Big)\varepsilon\Big), B^*(z^\varepsilon_{u^\varepsilon}(s) - z^\varepsilon_u(s))\Big),\end{aligned}$$

where

$$\begin{gathered}\tilde{z}^\varepsilon_{u^\varepsilon}(\tau) = z^\varepsilon_{u^\varepsilon}((i+1)\varepsilon), \quad \tilde{z}^\varepsilon_u(\tau) = z^\varepsilon_u((i+1)\varepsilon), \quad \text{and} \quad \tilde{u}_\varepsilon(\tau) = u((i+1)\varepsilon)\\ \text{for}\, \tau \in ((i-1)\varepsilon, i\varepsilon], \quad i = 1, 2, \ldots, \Big[\frac{s}{\varepsilon}\Big] - 1.\end{gathered}$$

By obvious estimates, $|\tilde{z}^\varepsilon_{u^\varepsilon}(\tau) - (I + \varepsilon\partial\phi)^{-1}y^\varepsilon_{u^\varepsilon}(\tau)|$ and $|\tilde{z}^\varepsilon_u(\tau) - (I + \varepsilon\partial\phi)^{-1}y^\varepsilon_u(\tau)|$ tend to zero as $\varepsilon \longrightarrow 0$ for all $\tau \in [0, s]$, so that we can use (3.19) and Lemma 3.1 to find the limits of $\tilde{z}^\varepsilon_{u^\varepsilon}$ and $\tilde{z}^\varepsilon_u$. In this way, under hypothesis (vii) it immediately follows that

$$\tilde{z}^\varepsilon_{u^\varepsilon}(\tau) \longrightarrow Py^0(\tau)\,\text{(on a subsequence)} \quad \text{and}$$

$$z^\varepsilon_u(\tau) \longrightarrow Py_u(\tau)\,\text{strongly in}\,\mathcal{H}\,\text{for all}\,\tau \in [0, s].$$

Under the other additional hypothesis, we have for every $\tau \in [0, s]$

$$\tilde{z}^\varepsilon_{u^\varepsilon}(\tau) \longrightarrow y^0(\tau)\,\text{(on a subsequence)} \quad \text{and} \quad z^\varepsilon_u(\tau) \longrightarrow y_u(\tau) \quad \text{strongly in}\,\mathcal{H}.$$

This happens because $y^0(\tau) \in \overline{D(\partial\phi)}$ (since $y^\varepsilon_{u^\varepsilon}(\tau) \in \overline{D(\partial\phi)}$ for any τ), and consequently $(I + \varepsilon\partial\phi)^{-1}y^0(\tau) \longrightarrow y^0(\tau)$ as $\varepsilon \longrightarrow 0$.

Now, to conclude that in the first case the integral in (3.23) converges to $\int_0^s (u^0(\tau) - u(\tau), B^*(Py^0(\tau) - Py_u(\tau)))d\tau$ as $\varepsilon \longrightarrow 0$, it suffices to observe that

$$\int_0^{([s/\varepsilon]-1)\varepsilon} (u^\varepsilon(\tau+\varepsilon), B^*((\tilde{z}^\varepsilon_{u^\varepsilon}(\tau) - \tilde{z}^\varepsilon_u(\tau)) - (Py^0(\tau) - Py_u(\tau))))d\tau \longrightarrow 0$$

and

$$\int_\varepsilon^{[s/\varepsilon]\varepsilon} (u^\varepsilon(\tau), B^*(Py^0(\tau-\varepsilon) - Py_u(\tau)))d\tau \longrightarrow \int_0^s (u^0(\tau), B^*(Py^0(\tau) - Py_u(\tau)))d\tau.$$

For the last limit, we also use the continuity of y^0 on $[0, s]$ except a countable subset (since y^0 is a function of bounded variation). In the second case, it follows similarly that the integral in (3.23) converges to $\int_0^s (u^0(\tau) - u(\tau), B^*(y^0(\tau) - y_u(\tau)))d\tau$ as $\varepsilon \longrightarrow 0$.

Thus, letting $\varepsilon \longrightarrow 0$ in (3.23), we obtain in both situations that

$$\frac{1}{2}|y^0(s) - y_{u_\eta}(s)|^2 \leq \int_0^s |Bu^0(\tau) - Bu_\eta(\tau)||y^0(\tau) - y_{u_\eta}(\tau)|d\tau,$$

and so, by (3.13), we have $y^0 \equiv y_{u^0}$ on $[0, t]$, since η is arbitrary. Taking (3.19) into account, this completes the proof.

4. Examples. We shall now indicate some control systems to which the results of the preceding section may be applied. The problem here is whether the additional hypotheses (vi) (on the one hand) and (vii) or (viii) (on the other) are verified in the cases of interest in applications. We shall see that among the significant situations which are covered by our results we find systems governed by semilinear parabolic equations and by the parabolic obstacle problem.

First, let us see a general situation in which hypothesis (vi) is verified (so that Theorem 3.1 applies). Let Ω be an open and bounded subset of $\mathbf{R}^n$ having a sufficiently smooth boundary, and A_0 be the elliptic differential operator defined by

$$A_0 y = -\sum_{i,j=1}^{n} \frac{\partial}{\partial x_j}\left(a_{ij}(x)\frac{\partial y}{\partial x_i}\right) + a_0(x)y,$$

where $a_{ij} \in C^1(\Omega)$, $a_0 \in L^\infty(\Omega)$, $a_{ij} = a_{ji}$ for all $i, j, a_0(x) \geq 0$ a.e. $x \in \Omega$, and for a certain $\omega > 0$,

$$\sum_{i,j=1}^{n} a_{ij}(x)\xi_i\xi_j \geq \omega \sum_{i=1}^{n} |\xi_i|^2, \qquad x \in \Omega, \quad (\xi_1, \xi_2, \ldots, \xi_n) \in \mathbf{R}^n. \tag{4.1}$$

Finally, let β be a maximal monotone graph in $\mathbf{R}^2$.

Consider the control system described by the following mixed boundary value problem:

$$\begin{cases} \dfrac{\partial y}{\partial t} + A_0 y + \beta(y) \ni Bu & \text{a.e. in}\, Q = (0,T) \times \Omega, \\ y = 0 & \text{in}\,(0,T) \times \partial\Omega, \\ y(0,x) = y^0(x) & \text{in}\,\Omega, \end{cases} \tag{4.2}$$

where $y^0 \in L^2(\Omega)$ and $B : \mathcal{U} \to L^2(\Omega)$ satisfies (v).

Let us observe that problem (4.2) may be written as an evolution equation of the form (1.2) if we put $\mathcal{H} = L^2(\Omega), \mathcal{V} = H_0^1(\Omega)$, and take A and ϕ as follows (see [2, §4.2]).

The linear operator $A : H_0^1(\Omega) \to H^{-1}(\Omega)$ is defined by

$$(Ay, z) = \sum_{i,j=1}^{n} \int_\Omega a_{ij} \frac{\partial y}{\partial x_i}\frac{\partial z}{\partial x_j} dx + \int_\Omega a_0\, yz\, dx \quad \text{for all}\, y, z \in H_0^1(\Omega). \tag{4.3}$$

Clearly, $D(A) = H_0^1(\Omega) \cap H^2(\Omega)$.

Then taking $j : \mathbf{R} \to (-\infty, +\infty]$ such that $\beta = \partial j$, we define the convex and lower semicontinuous function $\phi : L^2(\Omega) \to (-\infty, +\infty]$ in the following way:

$$\phi(y) = \int_\Omega j(y(x))dx. \tag{4.4}$$

Hence (see [2, Prop. 1.9]),

$$\partial\phi(y) = \{w \in L^2(\Omega) : w(x) \in \beta(y(x))\, \text{a.e.}\, x \in \Omega\}.$$

If condition (1.6) also holds, then assumptions (iii) and (iv) are wholly satisfied. But in the present case, (1.6) takes the following form:

$$(Ay, \beta_\eta(y)) \geq -c(1 + |\beta_\eta(y)|)(1 + |y|) \quad \text{for all}\, y \in D(A)\, \text{and}\, \eta > 0, \tag{4.5}$$

where $\beta_\eta(r) = \eta^{-1}(r - (I + \eta\beta)^{-1}r)$ for $r \in \mathbf{R}$, and $c > 0$. Indeed, one easily verifies that $((\partial\phi)_\eta(y))(x) = \beta_\eta(y(x))$ a.e. $x \in \Omega$. When $0 \in \beta(0)$, one obtains without difficulty that

$$(Ay, \beta_\eta(y)) \geq 0 \quad \text{for all } y \in H_0^1(\Omega),\ \eta > 0,$$

so that (1.6) holds.

PROPOSITION 4.1. *Suppose in addition that* $0 \in \beta(0)$. *Then* A *and* ϕ *defined by* (4.3) *and* (4.4), *respectively, satisfy* (vi) *besides* (iii) *and* (iv).

Proof. It is easy to see that

$$((I + \varepsilon\partial\phi)^{-1}z)(x) = (I + \varepsilon\beta)^{-1}(z(x)) \quad \text{a.e. } x \in \Omega.$$

Thus we have to show that $(I + \varepsilon\beta)^{-1}$ maps $H_0^1(\Omega)$ into itself and

$$\begin{aligned} &\sum_{i,j=1}^{n} \int_\Omega a_{ij} \frac{\partial}{\partial x_i}(I + \varepsilon\beta)^{-1}(z) \frac{\partial}{\partial x_j}(I + \varepsilon\beta)^{-1}(z) dx \\ &\quad + \int_\Omega a_0((I + \varepsilon\beta)^{-1}(z))^2\, dx \\ &\quad \leq \sum_{i,j=1}^{n} \int_\Omega a_{ij} \frac{\partial z}{\partial x_i}\frac{\partial z}{\partial x_j}\, dx + \int_\Omega a_0 z^2\, dx \quad \text{for all } z \in H_0^1(\Omega). \end{aligned} \tag{4.6}$$

We set $(I + \varepsilon\beta)^{-1} = \gamma$. Obviously, γ is a contraction on $\mathbf{R}$ and $\gamma(0) = 0$. For any positive integer m, we define

$$\gamma_m(r) = \int_{-\infty}^{\infty} \left(\gamma\left(r - \frac{1}{m}s\right) - \gamma\left(-\frac{1}{m}s\right)\right) \rho(s) ds,$$

where ρ is a C_0^∞-mollifier on $\mathbf{R}$, i.e., $\rho \in C^\infty(\mathbf{R})$, $\rho(r) = 0$ for $|r| > 1$, and $\int_{-\infty}^{\infty} \rho(r) dr = 1$. One readily checks that $\gamma_m \in C^\infty(\mathbf{R})$, $|\gamma_m'(r)| \leq 1$ for $r \in \mathbf{R}$, $\gamma_m(0) = 0$, and γ_m tend to γ uniformly on $\mathbf{R}$ as $m \to \infty$.

Let $z \in H_0^1(\Omega)$, arbitrary. Clearly,

$$\gamma_m(z) \to \gamma(z) \quad \text{strongly in } L^2(\Omega) \text{ (in fact in } L^\infty(\Omega)) \text{ as } m \to \infty. \tag{4.7}$$

Since $\gamma_m(0) = 0$ and $|\gamma_m'(r)| \leq 1$ for $r \in \mathbf{R}$, using the chain rule in $H_0^1(\Omega)$ we see that $\gamma_m(z) \in H_0^1(\Omega)$ and

$$\begin{aligned} &\sum_{i,j=1}^{n} \int_\Omega a_{ij} \frac{\partial}{\partial x_i}\gamma_m(z) \frac{\partial}{\partial x_j}\gamma_m(z) dx + \int_\Omega a_0(\gamma_m(z))^2\, dx \\ &\quad \leq \sum_{i,j=1}^{n} \int_\Omega a_{ij} \frac{\partial z}{\partial x_i}\frac{\partial z}{\partial x_j}\, dx + \int_\Omega a_0 z^2\, dx. \end{aligned} \tag{4.8}$$

But this means (by the ellipticity condition (4.1)) that $\gamma_m(z)$ is bounded in $H_0^1(\Omega)$; therefore, there exists $y \in H_0^1(\Omega)$ such that, on a subsequence of $\{m\}$,

$$\gamma_m(z) \to y \quad \text{weakly in } H_0^1(\Omega).$$

Hence, taking (4.7) into account, we have $\gamma(z) = y \in H_0^1(\Omega)$. Consequently,

$$\gamma_m(z) \to \gamma(z) \quad \text{weakly in } H_0^1(\Omega) \text{ as } m \to \infty. \tag{4.9}$$

Now, combining (4.8) with (4.9) and the weak lower semicontinuity of the norm of $H_0^1(\Omega)$ induced by (4.3) (which obviously is equivalent to the usual norm of $H_0^1(\Omega)$), we obtain (4.6), which finishes the proof.

The most interesting situations in which (vi) is fulfilled are offered by the control systems governed on the one hand by semilinear parabolic equations, i.e., in (4.2) β is a continuous monotonically increasing function on **R** such that $\beta(0) = 0$ (see [5] for a related but less general result), and on the other hand by the parabolic obstacle problem with obstacle $\psi \equiv 0$. In the last case β is given by

$$\beta(r) = \begin{cases} 0 & \text{for}\, r > 0, \\ (-\infty, 0] & \text{for}\, r = 0, \\ \emptyset & \text{for}\, r < 0, \end{cases}$$

and (4.2) can be written as (see [2, p. 138])

$$\begin{cases} y \geq 0 & \text{a.e. in}\, Q, \\ \dfrac{\partial y}{\partial t} + A_0 y = Bu & \text{a.e. in}\, \{(t,x) \in Q : y(t,x) > 0\}, \\ \dfrac{\partial y}{\partial t} = \max\{Bu, 0\} & \text{a.e. in}\, \{(t,x) \in Q : y(t,x) = 0\}, \\ y = 0 & \text{in}\, (0,T) \times \partial\Omega, \\ y(0,x) = y^0(x) & \text{in}\, \Omega, \end{cases}$$

where $y^0(x) \geq 0$ a.e. $x \in \Omega$.

Remark 4.1. We may apply Theorem 3.1 only to those control systems (4.2) (we emphasize that here y takes only the boundary values 0) where, in addition, $0 \in \beta(0)$. In the case of the obstacle problem with boundary values 0, scheme (2.4) converges only for obstacle 0. We can say nothing concerning the validity of (vi) for any other obstacle $\psi \not\equiv 0$.

Remark 4.2. The problem of the validity of (vi) in the case of Neumann boundary conditions remains open.

For Theorem 3.2, one easily observes that (vii) is verified for the general control system described by the variational inequality (see also [11, Thm. 2]).

$$y' + Ay + \partial I_K(y) \ni Bu,$$

where I_K is the indicator function of a closed convex subset K of $\mathcal{H}$, and $A, \partial I_K, B$ satisfy hypotheses (iii) – (v). Indeed, we have

$$\partial I_K(y) = \{p \in \mathcal{H} : (p, y - z) \geq 0\, \text{for all}\, z \in K\}, \qquad y \in \mathcal{H},$$

and

$$(I + \varepsilon \partial I_K)^{-1} = P_K \quad \text{for all}\, \varepsilon > 0,$$

where P_K is the projection operator of $\mathcal{H}$ into K, so that (vii) holds.

This framework wholly covers the case of the systems governed by the parabolic obstacle problem, but this time for any obstacle $\psi \in H^2(\Omega)$ and for boundary conditions of the form

$$\alpha_1 y + \alpha_2 \frac{\partial y}{\partial \nu} = 0 \quad \text{in}\, (0,T) \times \partial\Omega, \tag{4.10}$$

where α_1, α_2 are two nonnegative real numbers which are nonsimultaneously null. In this case,

$$K = \{y \in L^2(\Omega) : y \geq \psi \text{ a.e. in } \Omega\}$$

and the operator A is properly defined (see [2, §4.2]).

Finally, alternative hypothesis (viii) is obviously implied by the following simpler condition:

$$D(\partial\phi) \text{ is dense in } \mathcal{H}. \tag{4.11}$$

If we denote, as usual, $D(\phi) = \{y \in \mathcal{H} : \phi(y) < +\infty\}$, then (4.11) is equivalent to the condition

$$\overline{D(\phi)} = \mathcal{H}, \tag{4.11$'$}$$

since $\overline{D(\phi)} = \overline{D(\partial\phi)}$ (see for instance [6, Cor. 2.2, p. 110]). Here is a significant situation in which (4.11) is easy to verify. In (4.2) (where instead of the Dirichlet condition we may take the more general boundary condition (4.10)), β is a monotonically increasing function on **R**. Then (4.11) is satisfied with $\mathcal{H} = L^2(\Omega)$. Indeed, one readily shows that $C_0(\Omega) \subset D(\phi)$, where ϕ is given by (4.4), whence (4.11)$'$ follows immediately. If, in addition, $\alpha_2 \neq 0$ in (4.10), then it is easy to see that (4.5) is fulfilled (see [2, p. 137]), and Theorem 3.2 applies. Otherwise, as we have already seen, it suffices to have $\beta(0) = 0$ (see also [2, p. 137] for a less restrictive condition).

Remark 4.3. In the case of control systems governed by semilinear parabolic equations or the parabolic obstacle problem, Theorem 3.2 gives more general results than Theorem 3.1 (and also than the related results from [4], [5]). Indeed, the Trotter scheme (2.5) converges for more general boundary conditions (including Neumann conditions), and also for any obstacle $\psi \not\equiv 0$ in the case of the obstacle problem. Nevertheless, as we shall see in the following sections (see Remark 6.1), scheme (2.4) presents certain advantages in comparison with (2.5).

5. The approximating discrete control problem. According to (3.3), (3.4), consider the following sequence of discrete control problems:

(P^ε) minimize

$$\sum_{i=1}^{N} \varepsilon(h(u_i) + g(y_i)) + l(y_N) \tag{5.1}$$

over all N-tuples $(u_1, u_2, \ldots, u_N) \in \mathcal{U}^N$, where $(y_1, y_2, \ldots, y_N) \in \mathcal{H}^N$ satisfies the scheme

$$\begin{cases} y_i = (I + \varepsilon\partial\phi)^{-1}(I + \varepsilon A)^{-1}(y_{i-1} + \varepsilon B u_i), & i = 1, 2, \ldots, N, \\ y_0 = y^0. \end{cases} \tag{5.2}$$

Obviously y_i depends on ε, but for the sake of simplicity we shall not indicate this explicitly. It is easy to see that

$$|y_i| \leq |((I + \varepsilon\partial\phi)^{-1}(I + \varepsilon A)^{-1})^i y^0| + \sum_{j=1}^{i} \varepsilon|Bu_j|, \qquad i = 1, 2, \ldots, N. \tag{5.3}$$

PROPOSITION 5.1. *Suppose that* (i)–(vi) *hold. Then, for every* $\varepsilon > 0$, *problem* (P^ε) *has at least one solution* $(u_1^\varepsilon, u_2^\varepsilon, \ldots, u_N^\varepsilon) \in \mathcal{U}^N$.

Proof. Let $\{(u_{1,m}, u_{2,m}, \ldots, u_{N,m})\} \subset \mathcal{U}^N$ be a minimizing sequence for (P^ε). Using condition (1.4) and taking assumption (ii) together with (5.3) into account, we deduce that the sequences $\{u_{i,m}\}, i = 1, 2, \ldots, N$, are bounded in $\mathcal{U}$; so, on a certain subsequence of $\{m\}$, $u_{i,m}$ converge to u_i^ε weakly in $\mathcal{U}$ as $m \to \infty$.

Now set

$$z_{i,m} = (I + \varepsilon A)^{-1}(y_{i-1,m} + \varepsilon B u_{i,m}), \qquad i = 1, 2, \ldots, N,$$

where $y_{i,m}$ is given by (5.2) with $u_i = u_{i,m}$. Obviously, $y_{i,m} = (I + \varepsilon \partial\phi)^{-1} z_{i,m}$. We easily get

$$\frac{1}{2}|z_{i,m}|^2 + \varepsilon(A z_{i,m}, z_{i,m}) \le |y_{i-1,m}|^2 + \varepsilon^2 |B u_{i,m}|^2, \qquad i = 1, 2, \ldots, N,$$

whence, by (3.1) and (1.5), we obtain that $\{y_{i,m}\}, i = 1, 2, \ldots, N$, are bounded in $\mathcal{V}$. Since the inclusion $\mathcal{V} \subset \mathcal{H}$ is compact, by extracting a subsequence, $y_{i,m}$ converges to y_i^ε strongly in $\mathcal{H}$ as $m \to \infty$.

We may rewrite (5.2) where $u_i = u_{i,m}, y_i = y_{i,m}$ as follows:

$$y_{i,m} + \varepsilon \partial\phi(y_{i,m}) \ni (I + \varepsilon A)^{-1}(y_{i-1,m} + \varepsilon B u_{i,m}).$$

Since $\partial\phi$ is demiclosed in $\mathcal{H} \times \mathcal{H}$, letting $m \to \infty$ we obtain that y_i^ε satisfies (5.2) with $u_i = u_i^\varepsilon$.

Now let $m \to \infty$ in (5.1) where $u_i = u_{i,m}$; using the weak lower semicontinuity of h as well as the continuity of g and l, we conclude that $(u_1^\varepsilon, u_2^\varepsilon, \ldots, u_N^\varepsilon)$ solves (P^ε), and the proof is complete.

A sequence of optimal N-tuples for discrete problems (P^ε) yields a sequence of approximately optimal controls (a minimizing sequence) for problem (P). Indeed, let $(u_1^\varepsilon, u_2^\varepsilon, \ldots, u_N^\varepsilon)$ be an optimal N-tuple of (P^ε). Define

$$u^\varepsilon(t) = u^\varepsilon(i\varepsilon) = u_i^\varepsilon \quad \text{for } t \in ((i-1)\varepsilon, i\varepsilon], \quad i = 1, 2, \ldots, N. \tag{5.4}$$

We shall show that if we take these functions as controls in the initial system (1.2), (1.3), the corresponding values of the performance index (1.1) approach the infimum as $\varepsilon \to 0$. We set

$$J(u) = \int_0^T (h(u(t)) + g(y(t)))dt + l(y(T)) \quad \text{for } u \in L^2(0, T; \mathcal{U}),$$

where y is the solution of (1.2), (1.3).

THEOREM 5.1. *Let $y^0 \in \mathcal{V} \cap D(\phi)$. Under hypotheses* (i) – (iii), (1.6), (v), *and* (vi), *we have*

$$\lim_{\varepsilon \to 0} J(u^\varepsilon) = \inf\{J(u) : u \in L^2(0, T; \mathcal{U})\},$$

where the controls u^ε correspond by (5.4) *to solutions of problems (P^ε). Moreover, every weak limit point of $\{u^\varepsilon\}$ in $L^2(0, T; \mathcal{U})$ is an optimal control for problem (P).*

Proof. Denote by $J_\varepsilon(u_1, u_2, \ldots, u_N)$ the value of functional (5.1) at $(u_1, u_2, \ldots, u_N) \in \mathcal{U}^N$ (where, of course, y_i is given by (5.2)) and set $J_\varepsilon = \inf\{J_\varepsilon(u_1, u_2, \ldots, u_N) : (u_1, u_2, \ldots, u_N) \in \mathcal{U}^N\}$. We have $J_\varepsilon(u_1^\varepsilon, u_2^\varepsilon, \ldots, u_N^\varepsilon) = J_\varepsilon = W^\varepsilon(T, y^0)$. We shall compare $J(u^\varepsilon)$ with $J_\varepsilon(u_1^\varepsilon, u_2^\varepsilon, \ldots, u_N^\varepsilon)$.

Let $u \in \mathcal{U}$ such that $h(u) < \infty$. Clearly,

$$J_\varepsilon(u_1^\varepsilon, u_2^\varepsilon, \ldots, u_N^\varepsilon) \le J_\varepsilon(u, u, \ldots, u) \le \text{const.},$$

where the constant is independent of ε. Hence, using (1.4) together with (ii) and (5.3), we deduce that

$$\int_0^T |u^\varepsilon(t)|^2\,dt = \sum_{i=1}^N \varepsilon |u_i^\varepsilon|^2 \leq \text{const.}$$

Consequently, on a subsequence of $\{\varepsilon\}$, $u^\varepsilon \longrightarrow u^0$ weakly in $L^2(0,T;\mathcal{U})$ as $\varepsilon \longrightarrow 0$.

Let $(y_1^\varepsilon, y_2^\varepsilon, \dots, y_N^\varepsilon)$ be the solution of (5.2) corresponding to $u_i = u_i^\varepsilon$. We may write

$$J_\varepsilon(u_1^\varepsilon, u_2^\varepsilon, \dots, u_N^\varepsilon) = \int_0^T (h(u^\varepsilon(t)) + g(y^\varepsilon(t)))dt + l(y^\varepsilon(T)),$$

where

$$y^\varepsilon(t) = y^\varepsilon(i\varepsilon) = y_i^\varepsilon \quad \text{for}\, t \in ((i-1)\varepsilon, i\varepsilon], \quad i = 1, 2, \dots, N.$$

By virtue of Lemma 3.2 (see also the end of its proof),

$$\lim_{\varepsilon \longrightarrow 0} y^\varepsilon(t) = y_{u^0}(t) \quad \text{strongly in}\, \mathcal{H}\, \text{for all}\, t \in [0,T],$$

where y_{u^0} is the solution of (1.2), (1.3) corresponding to $u = u^0$.

On the other hand, multiplying (scalarly in $\mathcal{H}$) equation (1.2), where $u = u^\varepsilon$, first by the solution $y_{u^\varepsilon} \in W^{1,2}([0,T];\mathcal{H}) \cap C([0,T];\mathcal{V}) \cap L^2(0,T;D(A))$ of (1.2), (1.3) and second by y'_{u^ε}, then integrating on $[0,T]$, after some calculation we obtain

$$|y_{u^\varepsilon}(t)| \leq \text{const.} \quad \text{and} \quad \int_0^T |y'_{u^\varepsilon}(s)|^2\,ds + |y_{u^\varepsilon}(t)|_{\mathcal{V}}^2 \leq \text{const.} \quad \text{for}\, t \in [0,T],$$

where the constants are independent of ε. Now applying the Arzelà–Ascoli theorem (do not forget that the inclusion $\mathcal{V} \subset \mathcal{H}$ is compact), we infer that, on a subsequence, $y_{u^\varepsilon} \longrightarrow y_0$ strongly in $C([0,T];\mathcal{H})$ as $\varepsilon \longrightarrow 0$. Subtracting the differential equations satisfied by $y_{u^\varepsilon}, y_{u^0}$, and multiplying the difference by $y_{u^\varepsilon} - y_{u^0}$, we get

$$\frac{1}{2}|y_{u^\varepsilon}(t) - y_{u^0}(t)|^2 \leq \int_0^t (Bu^\varepsilon - Bu^0, y_{u^\varepsilon} - y_{u^0})ds \quad \text{for all}\, t \in [0,T],$$

whence, letting $\varepsilon \longrightarrow 0$, it follows that $y_0 \equiv y_{u^0}$. Hence,

$$y_{u^\varepsilon} \longrightarrow y_{u^0} \quad \text{strongly in}\, C([0,T];\mathcal{H}).$$

Since g and l are continuous, by the Lebesgue dominated convergence theorem we have

$$\lim_{\varepsilon \longrightarrow 0} (J(u^\varepsilon) - J_\varepsilon(u_1^\varepsilon, u_2^\varepsilon, \dots, u_N^\varepsilon)) = 0.$$

But Theorem 3.1 asserts that

$$\lim_{\varepsilon \longrightarrow 0} J_\varepsilon(u_1^\varepsilon, u_2^\varepsilon, \dots, u_N^\varepsilon) = \lim_{\varepsilon \longrightarrow 0} W^\varepsilon(T, y^0) = W(T, y^0) = \inf\{J(u) : u \in L^2(0,T;\mathcal{U})\},$$

which proves the first statement of Theorem 5.1.

Finally, let u^0 be a weak limit point in $L^2(0,T;\mathcal{U})$ of $\{u^\varepsilon\}$. Then take limits in (5.1) where $u_i = u_i^\varepsilon$. Using the weak lower semicontinuity in $L^2(0,T;\mathcal{U})$ of $u \longmapsto \int_0^T h(u(t))dt$ and the first part of Theorem 5.1 (just proved), we conclude that u^0 is an optimal control for problem (P), which finishes the proof.

6. The discrete feedback law. Let Ω be an open bounded subset of $\mathbf{R}^n$. In this section, we are forced to restrict ourselves to systems governed by

$$y' + Ay + \beta(y) \ni Bu \tag{6.1}$$

with the initial condition

$$y(0) = y^0, \tag{6.2}$$

where A, B satisfy (iii), (v) with $\mathcal{H} = L^2(\Omega)$, respectively, and β is a maximal monotone graph in $\mathbf{R}^2$. As we have seen in the preceding section, (6.1) can be set in the form (1.2) where ϕ is given by (4.4). Suppose that (4.5) holds as well.

Consider the following optimal control problem:

(P_1) minimize functional (1.1) over all $u \in L^2(0,T;\mathcal{U})$, where $y \in C([0,T];\mathcal{H})$ satisfies (6.1), (6.2).

We impose on the data of (P_1) hypotheses (i) and (ii).

Let $\varepsilon = T/N$, where N is a positive integer. Define the function $V^\varepsilon : [0,T] \times \mathcal{H} \to \mathbf{R}$ by

$$V^\varepsilon(t,y) = W^\varepsilon(T-t,y), \qquad (t,y) \in [0,T] \times \mathcal{H}, \tag{6.3}$$

where in (2.4) β replaces $\partial\phi$. We may regard V^ε as an approximately optimal value function associated with problem (P_1). Indeed, if (vi) also holds, then by virtue of Theorem 3.1,

$$\lim_{\varepsilon \to 0} V^\varepsilon(t,y) = V(t,y) \quad \text{for all } (t,y) \in [0,T] \times \overline{D(A) \cap D(\partial\phi)},$$

where in (1.7) ϕ is given by (4.4). Note that only the values of V^ε calculated at $t = i\varepsilon$ $(i = 0, 1, \ldots, N)$ will be relevant in our further considerations. By (6.3) and (2.4) we have

$$\left\{ \begin{aligned} & V^\varepsilon(i\varepsilon, y) = \inf\{\varepsilon h(u) + \varepsilon g((I+\varepsilon\beta)^{-1}(I+\varepsilon A)^{-1}(y+\varepsilon Bu)) \\ & \qquad + V^\varepsilon((i+1)\varepsilon, (I+\varepsilon\beta)^{-1}(I+\varepsilon A)^{-1}(y+\varepsilon Bu)) : u \in \mathcal{U}\}, \quad y \in \mathcal{H}, \\ & \qquad i = 0, 1, \ldots, N-1, \\ & V^\varepsilon(T,y) = l(y), y \in \mathcal{H}. \end{aligned} \right. \tag{6.4}$$

Finally, let us denote by (P_1^ε) the discrete problem (P^ε) where ϕ is given by (4.4). The relationship between the components $u_1^\varepsilon, u_2^\varepsilon, \ldots, u_N^\varepsilon$ of a solution of (P_1^ε) and the corresponding $y_1^\varepsilon, y_2^\varepsilon, \ldots, y_N^\varepsilon$ via V^ε (a genuine feedback law for the discrete system) forms the content of the following theorem.

THEOREM 6.1. *Suppose that hypotheses* (i)–(iii), (4.5), (v), *and* (vi) *hold, and let* $y^0 \in \mathcal{V} \cap D(\phi)$. *If* $(u_1^\varepsilon, u_2^\varepsilon, \ldots, u_N^\varepsilon) \in \mathcal{U}^N$ *is an optimal N-tuple for* (P_1^ε) *and* $y_0^\varepsilon, y_1^\varepsilon, \ldots, y_N^\varepsilon$ *are given by* (5.2) *where* $u_i = u_i^\varepsilon, i = 1, 2, \ldots, N$ (*and* $\partial\phi$ *is replaced by* β), *then the following discrete feedback law holds*:

$$u_i^\varepsilon \in \partial h^*(-B^*\partial_y V^\varepsilon((i-1)\varepsilon, y_{i-1}^\varepsilon)), \qquad i = 1, 2, \ldots, N. \tag{6.5}$$

Here $\partial_y V^\varepsilon(t,y)$ represents the generalized gradient (in Clarke's sense) of $y \mapsto V^\varepsilon(t,y)$. (Recall that, by virtue of Lemma 3.3, V^ε is Lipschitz continuous on bounded subsets of $\mathcal{H}$ with respect to y.)

Proof. Inspired by an idea of F. H. Clarke and R. B. Vinter (see [8, Lem. 8.4]), we consider the following perturbed discrete system:

$$\begin{cases} y_i^- &= (I+\varepsilon\beta)^{-1}(I+\varepsilon A)^{-1}(y_{i-1}^+ + \varepsilon Bu_i), \\ y_{i-1}^+ &= y_{i-1}^- + v_i, \quad i = 1,2,\dots,N, \\ y_0^- &= y^0, \end{cases}$$

where $(u_1, u_2, \dots, u_N) \in \mathcal{U}^N$ and $(v_1, v_2, \dots, v_N) \in \mathcal{H}^N$. Obviously,

$$y_i^- = (I+\varepsilon\beta)^{-1}(I+\varepsilon A)^{-1}(y_{i-1}^- + \varepsilon Bu_i + v_i).$$

Fix $\delta > 0$. Let $(u_1, u_2, \dots, u_N) \in \mathcal{U}^N, (v_1, v_2, \dots, v_N) \in \mathcal{H}^N$ such that

$$|y_i^- - y_i^\varepsilon| \le \delta, |y_i^+ - y_i^\varepsilon| \le \delta \quad \text{for } i = 1,2,\dots,N-1.$$

From the definition of V^ε (see (6.4)) we have

$$\varepsilon h(u_i) + \varepsilon g(y_i^-) + V^\varepsilon(i\varepsilon, y_i^-) - V^\varepsilon((i-1)\varepsilon, y_{i-1}^+) \ge 0, \qquad i = 1,2,\dots,N.$$

For simplicity we set $y_i = y_i^-$. Adding the above inequalities, we obtain

$$(6.6)\quad \sum_{i=1}^N \varepsilon(h(u_i) + g(y_i)) + l(y_N) - \sum_{i=1}^N (V^\varepsilon((i-1)\varepsilon, y_{i-1} + v_i) - V^\varepsilon((i-1)\varepsilon, y_{i-1})) \\ \ge V^\varepsilon(0, y^0).$$

By the mean-value theorem for generalized gradient (see, for instance, [2, Cor. 1.2]),

$$V^\varepsilon((i-1)\varepsilon, y_{i-1} + v_i) - V^\varepsilon((i-1)\varepsilon, y_{i-1}) = (p_{i-1}, v_i),$$

where $p_{i-1} \in \partial_y V^\varepsilon((i-1)\varepsilon, z_{i-1})$ with $|z_{i-1} - y_{i-1}^\varepsilon| \le \delta, i = 1,2,\dots,N$. We define

$$k_{i,\delta}(v) = \sup\{(p,v) : p \in \partial_y V^\varepsilon(i\varepsilon, y), |y - y_i^\varepsilon| \le \delta\} \\ \text{for all } v \in \mathcal{H}, i = 1,2,\dots,N, \text{ and } \delta > 0.$$

Clearly, $k_{i,\delta}$ is convex, lower semicontinuous and finite everywhere (since V^ε is Lipschitz continuous on bounded subsets); therefore, it is a continuous convex function on $\mathcal{H}$. Returning to (6.6), we have

$$\sum_{i=1}^N \varepsilon(h(u_i) + g(y_i)) + \sum_{i=1}^N k_{i-1,\delta}(-v_i) + l(y_N) \ge V^\varepsilon(0, y^0).$$

Thus we have proved that $(u_1^\varepsilon, u_2^\varepsilon, \dots, u_N^\varepsilon, 0, 0, \dots, 0) \in \mathcal{U}^N \times \mathcal{H}^N$ and $(y_1^\varepsilon, y_2^\varepsilon, \dots, y_N^\varepsilon)$ solve the following problem:

(P_δ) minimize

$$(6.7)\qquad \sum_{i=1}^N \varepsilon(h(u_i) + g(y_i)) + \sum_{i=1}^N k_{i-1,\delta}(-v_i) + l(y_N)$$

over all $(u_1, u_2, \dots, u_N, v_1, v_2, \dots, v_N) \in \mathcal{U}^N \times \mathcal{H}^N$, where

$$(6.8)\qquad \begin{cases} y_i = (I+\varepsilon\beta)^{-1}(I+\varepsilon A)^{-1}(y_{i-1} + \varepsilon Bu_i + v_i), \quad i = 1,2,\dots,N, \\ y_0 = y^0, \end{cases}$$

and

$$|y_i - y_i^\varepsilon| \leq \delta, \quad |y_i + v_{i+1} - y_i^\varepsilon| \leq \delta, \quad i = 0, 1, \ldots, N-1.$$

Roughly speaking, we shall get (6.5) from a set of optimality conditions for (P_δ) as $\delta \longrightarrow 0$. To this end, we associate with (P_δ) the following family of smooth (and penalized) problems:

$(P_{\delta,\eta})$ minimize

$$\begin{aligned}(6.9)\qquad &\sum_{i=1}^N \varepsilon\left(h(u_i) + \frac{1}{2}|u_i - u_i^\varepsilon|^2 + g_\eta(y_i) + I_{i-1,\delta,\eta}(y_{i-1}) + I_{i-1,\delta,\eta}(y_{i-1} + v_i)\right) \\ &+ \sum_{i=1}^N \left(k_{i-1,\delta}(-v_i) + \frac{1}{2}|v_i|^2\right) + l_\eta(y_N)\end{aligned}$$

over all $(u_1, \ldots, u_N, v_1, \ldots, v_N) \in \mathcal{U}^N \times \mathcal{H}^N$, where y_i satisfies

$$(6.10)\qquad \begin{cases} y_i = (I + \varepsilon\tilde{\beta}_\eta)^{-1}(I + \varepsilon A)^{-1}(y_{i-1} + \varepsilon B u_i + v_i), & i = 1, 2, \ldots, N, \\ y_0 = y^0. \end{cases}$$

Here g_η and l_η are regularizations of g and l, respectively (see [2, p. 28]), $I_{i,\delta,\eta}$ is the convex regularization of the indicator function $I_{i,\delta}$ of the set $\{y \in \mathcal{H} : |y - y_i^\varepsilon| \leq \delta\}, i = 0, 1, \ldots, N-1$ (i.e., $I_{i,\delta,\eta}(y) = \inf\{|y - z|^2/2\eta : |z - y_i^\varepsilon| \leq \delta\}$), and $\tilde{\beta}_\eta$ is a regularization of $\beta_\eta = \eta^{-1}(I - (I + \eta\beta)^{-1})$ (see [2, p. 75]).

Now fix $\delta > 0$. The proof of the following lemma is similar to that of Proposition 5.1.

LEMMA 6.1. *For every* $\eta > 0$, *problem* $(P_{\delta,\eta})$ *has at least one solution* $(u_{1,\delta,\eta}, \ldots, u_{N,\delta,\eta}, v_{1,\delta,\eta}, \ldots, v_{N,\delta,\eta}) \in \mathcal{U}^N \times \mathcal{H}^N$.

For the sake of simplicity, we set $u_{i,\eta} = u_{i,\delta,\eta}, v_{i,\eta} = v_{i,\delta,\eta}$ for any optimal $2N$-tuple of $(P_{\delta,\eta})$ given by Lemma 6.1. The following result is an effect of the penalization terms.

LEMMA 6.2. *For each* $\eta > 0$, *let us consider a solution* $(u_{1,\eta}, \ldots, u_{N,\eta}, v_{1,\eta}, \ldots, v_{N,\eta})$ *for problem* $(P_{\delta,\eta})$ *and let* $(y_{1,\eta}, y_{2,\eta}, \ldots, y_{N,\eta})$ *be the solution of* (6.10) *corresponding to* $u_i = u_{i,\eta}$ *and* $v_i = v_{i,\eta}$. *We have*

$$\lim_{\eta \to 0} u_{i,\eta} = u_i^\varepsilon \quad \textit{strongly in } \mathcal{U},$$

$$\lim_{\eta \to 0} v_{i,\eta} = v_i^\varepsilon = 0 \quad \textit{strongly in } \mathcal{H},$$

$$\lim_{\eta \to 0} y_{i,\eta} = y_i^\varepsilon \quad \textit{strongly in } \mathcal{H},$$

$i = 1, 2, \ldots, N$.

Proof. Let us denote by $J_\delta(u_1, \ldots, u_N, v_1, \ldots, v_N)$ functional (6.7) and by $J_{\delta,\eta}(u_1, \ldots, u_N, v_1, \ldots, v_N)$ functional (6.9). Also, we denote by $\tilde{J}_{\delta,\eta}(u_1, \ldots, u_N, v_1, \ldots, v_N)$ the functional obtained from (6.9) by eliminating the penalization terms $(1/2)|u_i - u_i^\varepsilon|^2$ and $(1/2)|v_i|^2$.

Arguing as below, we can show that the solution of (6.10) corresponding to $u_i = u_i^\varepsilon$, $v_i = 0\,(i = 1, 2, \ldots, N)$ converges to $(y_1^\varepsilon, y_2^\varepsilon, \ldots, y_N^\varepsilon)$ as $\eta \longrightarrow 0$. But since $I_{i,\delta,\eta}(y) = 0$ when $|y - y_i^\varepsilon| \leq \delta$, for each $\delta > 0$, we can find $\eta_\delta > 0$ such that

$$(6.11)\qquad \begin{aligned} J_{\delta,\eta}(u_{1,\eta}, \ldots, u_{N,\eta}, v_{1,\eta}, \ldots, v_{N,\eta}) &\leq J_{\delta,\eta}(u_1^\varepsilon, \ldots, u_N^\varepsilon, v_1^\varepsilon, \ldots, v_N^\varepsilon) \\ &\leq \text{const.} \quad \text{for } \eta \in (0, \eta_\delta], \end{aligned}$$

where the constant is independent of δ. Hence, since $k_{i,\delta}(v) \geq -L_i|v|$ (where L_i is the Lipschitz constant of $y \mapsto V^\varepsilon(i\varepsilon, y)$ corresponding to a certain ball with y_i^ε as center), it follows that there exists $M > 0$, also independent of δ, such that

$$(6.12)\qquad |u_{i,\eta}| \leq M \quad \text{and} \quad |v_{i,\eta}| \leq M \quad \text{for } \eta \in (0, \eta_\delta], \quad i = 1, 2, \ldots, N.$$

Consequently, on a subsequence of $\{\eta\}$, $u_{i,\eta} \to u_{i,0}$ weakly in $\mathcal{U}$ and $v_{i,\eta} \to v_{i,0}$ weakly in $\mathcal{H}$ as $\eta \to 0, i = 1, 2, \ldots, N$. (Note that the independence of δ of the boundedness constants of $\{u_{i,\delta,\eta}\}$ and $\{v_{i,\delta,\eta}\}$ will be important when we let $\delta \to 0$.)

As in Proposition 5.1, we get that, on a subsequence, $y_{i,\eta} \to y_{i,0}$ strongly in $\mathcal{H}$ as $\eta \to 0, i = 1, 2, \ldots, N$. Let $j_\eta(r) = \int_0^r \beta_\eta(s)ds$. Define $\phi_\eta(y) = \int_\Omega j_\eta(y(x))dx$ for $y \in L^2(\Omega)$. We have $\nabla\phi_\eta(y) = \beta_\eta(y)$ a.e. in Ω; therefore, we may write

$$y_{i,\eta} + \varepsilon\nabla\phi_\eta(y_{i,\eta}) = (I + \varepsilon A)^{-1}(y_{i-1,\eta} + \varepsilon Bu_{i,\eta} + v_{i,\eta}) + \varepsilon(\beta_\eta(y_{i,\eta}) - \tilde{\beta}_\eta(y_{i,\eta})).$$

Since $\nabla\phi_\eta(y_{i,\eta})$ is the Yosida approximation of $\partial\phi$ calculated at y and $|\beta_\eta(r) - \tilde{\beta}_\eta(r)| \leq 2\eta$ for all $r \in \mathbf{R}$ (see [2, p. 78]), we obtain as $\eta \to 0$,

$$y_{i,0} + \varepsilon\partial\phi(y_{i,0}) \ni (I + \varepsilon A)^{-1}(y_{i-1,0} + \varepsilon Bu_{i,0} + v_{i,0}).$$

So we have shown that, if $u_{i,\eta} \to u_{i,0}$ weakly in $\mathcal{U}$ and $v_{i,\eta} \to v_{i,0}$ weakly in $\mathcal{H}$ as $\eta \to 0$, then the components $y_{i,\eta}$ converge to the components of the solution of scheme (6.8) corresponding to $u_i = u_{i,0}$ and $v_i = v_{i,0}$ strongly in $\mathcal{H}$.

We still need the following semicontinuity property:

$$\text{(6.13)} \quad \liminf_{\eta\to 0} I_{i-1,\delta,\eta}(y_{i-1,\eta} + v_{i,\eta}) \geq I_{i-1,\delta}(y_{i-1,0} + v_{i,0}), \qquad i = 1, 2, \ldots, N.$$

To see this, let us observe that, in view of (6.11), $I_{i-1,\delta,\eta}(y_{i-1,\eta} + v_{i,\eta})$ is bounded with respect to η. Hence, by a well-known expression of the convex regularization (see [6, p. 121]), it readily follows that

$$(I + \eta I_{i-1,\delta})^{-1}(y_{i-1,\eta} + v_{i,\eta}) - (y_{i-1,\eta} + v_{i,\eta}) \to 0 \quad \text{strongly in } \mathcal{H},$$

whence by the lower semicontinuity of $I_{i-1,\delta}$ we have (6.13).

Now, by the convergence of $\{u_{i,\eta}\}$, $\{v_{i,\eta}\}$, and $\{y_{i,\eta}\}$, together with (6.13) and the weak lower semicontinuity of $u \mapsto h(u) + (1/2)|u - u_i^\varepsilon|^2, v \mapsto k_{i-1,\delta}(-v) + (1/2)|v|^2, i = 1, 2, \ldots, N$, we obtain the following chain of inequalities:

$$\begin{aligned}
&\liminf_{\eta\to 0} J_{\delta,\eta}(u_{1,\eta}, \ldots, u_{N,\eta}, v_{1,\eta}, \ldots, v_{N,\eta}) \\
&\quad \geq \liminf_{\eta\to 0} \tilde{J}_{\delta,\eta}(u_{1,\eta}, \ldots, u_{N,\eta}, v_{1,\eta}, \ldots, v_{N,\eta}) \\
&\quad \geq J_\delta(u_{1,0}, \ldots, u_{N,0}, v_{1,0}, \ldots, v_{N,0}) \\
&\quad \geq J_\delta(u_1^\varepsilon, \ldots, u_N^\varepsilon, 0, \ldots, 0) \\
&\quad = \lim_{\eta\to 0} J_{\delta,\eta}(u_1^\varepsilon, \ldots, u_N^\varepsilon, 0, \ldots, 0) \\
&\quad \geq \limsup_{\eta\to 0} J_{\delta,\eta}(u_{1,\eta}, \ldots, u_{N,\eta}, v_{1,\eta}, \ldots, v_{N,\eta}) \\
&\quad \geq \limsup_{\eta\to 0} \tilde{J}_{\delta,\eta}(u_{1,\eta}, \ldots, u_{N,\eta}, v_{1,\eta}, \ldots, v_{N,\eta}).
\end{aligned}$$

So we infer that

$$\begin{aligned}
\lim_{\eta\to 0} J_{\delta,\eta}(u_{1,\eta}, \ldots, u_{N,\eta}, v_{1,\eta}, \ldots, v_{N,\eta}) &= \lim_{\eta\to 0} \tilde{J}_{\delta,\eta}(u_{1,\eta}, \ldots, u_{N,\eta}, v_{1,\eta}, \ldots, v_{N,\eta}) \\
&= J_\delta(u_1^\varepsilon, \ldots, u_N^\varepsilon, 0, \ldots, 0),
\end{aligned}$$

and hence

$$\lim_{\eta\to 0} |u_{i,\eta} - u_i^\varepsilon| = 0, \quad \lim_{\eta\to 0} |v_{i,\eta}| = 0, \qquad i = 1, 2, \ldots, N.$$

The proof is thus complete.

In order to obtain the required optimality conditions for $(P_{\delta,\eta})$, we shall compare the optimal value with the values corresponding to $u_i = u_{i,\eta} + \lambda\bar{u}_i, v_i = v_{i,\eta} + \lambda\bar{v}_i, i = 1,2,\dots,N$, where $\lambda > 0$, and $\bar{u}_i \in \mathcal{U}, \bar{v}_i \in \mathcal{H}$ are arbitrary. We have

$$\begin{aligned}(6.14)\quad & J_{\delta,\eta}(u_{1,\eta}+\lambda\bar{u}_1,\dots,u_{N,\eta}+\lambda\bar{u}_N,v_{1,\eta}+\lambda\bar{v}_1,\dots,v_{N,\eta}+\lambda\bar{v}_N)\\ & \quad - J_{\delta,\eta}(u_{1,\eta},\dots,u_{N,\eta},v_{1,\eta},\dots,v_{N,\eta}) \geq 0 \quad \text{for all } \lambda > 0 \text{ and } \bar{u}_i \in \mathcal{U}, \bar{v}_i \in \mathcal{H}.\end{aligned}$$

Denote by $y_{i,\eta,\lambda}$ the solution of (6.10) corresponding to $u_i = u_{i,\eta} + \lambda\bar{u}_i, v_i = v_{i,\eta} + \lambda\bar{v}_i$. It is easy to check that

$$\lim_{\lambda\to 0}\frac{1}{\lambda}(y_{i,\eta,\lambda} - y_{i,\eta}) = z_{i,\eta} \quad \text{strongly in } \mathcal{H},$$

where $z_{i,\eta}$ satisfies the scheme

$$(6.15)\quad \begin{cases} z_{i,\eta} = (1+\varepsilon\tilde{\beta}'_\eta(y_{i,\eta}))^{-1}(I+\varepsilon A)^{-1}(z_{i-1,\eta}+\varepsilon B\bar{u}_i+\bar{v}_i), & i = 1,2,\dots,N,\\ z_{0,\eta} = 0. \end{cases}$$

Dividing (6.14) by λ and letting $\lambda \to 0$, we get

$$\begin{aligned}(6.16)\quad & \sum_{i=1}^{N}\varepsilon(h'(u_{i,\eta};\bar{u}_i) + (u_{i,\eta}-u_i^\varepsilon,\bar{u}_i) + (\nabla g_\eta(y_{i,\eta}),z_{i,\eta})\\ &\quad +(\nabla I_{i-1,\delta,\eta}(y_{i-1,\eta}) + \nabla I_{i-1,\delta,\eta}(y_{i-1,\eta}+v_{i,\eta}),z_{i-1,\eta}))\\ &\quad +\sum_{i=1}^{N}(k'_{i-1,\delta}(-v_{i,\eta};-\bar{v}_i) + (v_{i,\eta},\bar{v}_i) + (\varepsilon\nabla I_{i-1,\delta,\eta}(y_{i-1,\eta}+v_{i,\eta}),\bar{v}_i))\\ &\quad +(\nabla l_\eta(y_{N,\eta}),z_{N,\eta}) \geq 0 \quad \text{for all } (\bar{u}_1,\bar{u}_2,\dots,\bar{u}_N) \in \mathcal{U}^N,\\ &\qquad\qquad\qquad\qquad\qquad\qquad (\bar{v}_1,\bar{v}_2,\dots,\bar{v}_N) \in \mathcal{H}^N,\end{aligned}$$

where $h'(u_{i,\eta};\bar{u}_i)$ and $k'_{i-1,\delta}(-v_{i,\eta};-\bar{v}_i)$ denote the directional derivatives of h and $k_{i-1,\delta}$ at $u_{i,\eta}$ and $-v_{i,\eta}$ in the directions $\bar{u}_i$ and $-\bar{v}_i$, respectively. Note that, for sufficiently small $\eta > 0$, we have

$$\nabla I_{i-1,\delta,\eta}(y_{i-1,\eta}) = 0, \quad \nabla I_{i-1,\delta,\eta}(y_{i-1,\eta}+v_{i,\eta}) = 0, \qquad i = 1,2,\dots,N,$$

since $y_{i,\eta} \to y_i^\varepsilon$ and $v_{i,\eta} \to 0$ strongly in $\mathcal{H}$ as $\eta \to 0$.

Now define $p_{i,\eta}$ by

$$(6.17)\quad \begin{cases} p_{i-1,\eta} = (I+\varepsilon A)^{-1}((1+\varepsilon\tilde{\beta}'_\eta(y_{i,\eta}))^{-1}(p_{i,\eta}-\varepsilon\nabla g_\eta(y_{i,\eta}))), & i = 1,2,\dots,N,\\ p_{N,\eta} = -\nabla l_\eta(y_{N,\eta}). \end{cases}$$

Some calculation in (6.16) involving (6.15), (6.17) together with the identity

$$(p_{i,\eta} - p_{i-1,\eta}, z_{i,\eta}) = (p_{i,\eta}, z_{i,\eta}) - (p_{i-1,\eta}, z_{i-1,\eta}) - (p_{i-1,\eta}, z_{i,\eta} - z_{i-1,\eta}),$$

gives, for sufficiently small $\eta > 0$,

$$\begin{aligned}&\sum_{i=1}^{N}\varepsilon(h'(u_{i,\eta};\bar{u}_i) + (u_{i,\eta}-u_i^\varepsilon - B^*p_{i-1,\eta},\bar{u}_i))\\ &\quad +\sum_{i=1}^{N}(k'_{i-1,\delta}(-v_{i,\eta};-\bar{v}_i) + (v_{i,\eta}-p_{i-1,\eta},\bar{v}_i))\\ &\quad \geq 0 \quad \text{for all } (\bar{u}_1,\bar{u}_2,\dots,\bar{u}_N) \in \mathcal{U}^N, (\bar{v}_1,\bar{v}_2,\dots,\bar{v}_N) \in \mathcal{H}^N,\end{aligned}$$

which yields, for $\eta > 0$ small enough,

$$B^* p_{i-1,\eta} \in \partial h(u_{i,\eta}) + u_{i,\eta} - u_i^\varepsilon, \tag{6.18}$$

$$-p_{i-1,\eta} \in \partial k_{i-1,\delta}(-v_{i,\eta}) - v_{i,\eta}, \qquad i = 1, 2, \dots, N. \tag{6.19}$$

Obviously, like $u_{i,\eta} = u_{i,\delta,\eta}$, $v_{i,\eta} = v_{i,\delta,\eta}$, and $y_{i,\eta} = y_{i,\delta,\eta}$, $p_{i,\eta}$ also depends on $\delta > 0$ (even if we have purposely omitted, for simplicity, the subscript δ). So, in what follows we set $p_{i,\delta,\eta} = p_{i,\eta}$. As (6.12) shows, $u_{i,\eta}$ and $v_{i,\eta}$ are bounded by constants which are independent of δ; it easily follows that $y_{i,\eta}$, $\nabla g_\eta(y_{i,\eta})$, and $\nabla l_\eta(y_{N,\eta})$ satisfy similar boundedness conditions. Hence $p_{i,\delta,\eta}$ is also bounded in the sense of (6.12) by a constant $M > 0$ which is independent of δ, i.e., for each $\delta > 0$, we find $\eta_\delta > 0$ such that

$$|p_{i,\delta,\eta}| \leq M \quad \text{for } \eta \in (0, \eta_\delta], \quad i = 0, 1, \dots, N-1.$$

(Moreover, the above condition also holds in the norm of $\mathcal{V}$, but with a different constant M.) So, for each $\delta > 0$, on a subsequence of $\{\eta\}$,

$$p_{i,\delta,\eta} \to p_{i,\delta} \quad \text{weakly (or strongly) in } \mathcal{H} \quad \text{as } \eta \to 0.$$

This, in conjunction with the boundedness condition for $p_{i,\delta,\eta}$, yields

$$|p_{i,\delta}| \leq M \quad \text{for all } \delta > 0, \quad i = 0, 1, \dots, N-1. \tag{6.20}$$

Letting $\eta \to 0$ in (6.18), (6.19), since ∂h and $\partial k_{i,\delta}$ are demiclosed we get

$$B^* p_{i-1,\delta} \in \partial h(u_i^\varepsilon), \tag{6.21}$$

$$-p_{i-1,\delta} \in \partial k_{i-1,\delta}(0) \tag{6.22}$$

for $\delta > 0, i = 1, 2, \dots, N$.

On the other hand, by (6.20) we have on a subsequence of $\{\delta\}$

$$p_{i,\delta} \to p_i^\varepsilon \quad \text{weakly in } \mathcal{H} \text{ as } \delta \to 0, \quad i = 0, 1, \dots, N-1. \tag{6.23}$$

Now, the assertion of Theorem 6.1 will follow by taking limits in (6.21), as soon as we show by interpreting (6.22) the following relationship among p_i^ε, y_i^ε, and V^ε:

$$-p_{i-1}^\varepsilon \in \partial_y V^\varepsilon((i-1)\varepsilon, y_{i-1}^\varepsilon), \qquad i = 1, 2, \dots, N. \tag{6.24}$$

This inclusion is a discrete variant of the well-known connection between the maximum principle and dynamic programming. Its proof is a simple adaptation for our discrete (but infinite-dimensional) case of the proof of Theorem 3.1 from [8].

Let us first clarify the meaning of inclusion (6.22). Observe that

$$-p_{i-1,\delta} \in \overline{co} \bigcup_{|y - y_{i-1}^\varepsilon| \leq \delta} \partial_y V^\varepsilon((i-1)\varepsilon, y), \qquad i = 1, 2, \dots, N.$$

Otherwise, there exists a hyperplane strictly separating $-p_{i-1,\delta}$ and the above closed convex set (see [6, Thm. 1.15, p. 21]), i.e., we find $v \in \mathcal{H}$ such that

$$\begin{aligned}(-p_{i-1,\delta}, -v) &> \sup\left\{(p, -v) : p \in \overline{co} \bigcup_{|y - y_{i-1}^\varepsilon| \leq \delta} \partial_y V^\varepsilon((i-1)\varepsilon, y)\right\} \\ &= \sup\left\{(p, -v) : p \in \bigcup_{|y - y_{i-1}^\varepsilon| \leq \delta} \partial_y V^\varepsilon((i-1)\varepsilon, y)\right\} \\ &= k_{i-1,\delta}(-v),\end{aligned}$$

which contradicts (6.22) because $k_{i-1,\delta}(0) = 0$.

We let δ tend to 0 in (6.21), (6.22) by using (6.23), and we get

$$B^* p^\varepsilon_{i-1} \in \partial h(u^\varepsilon_i), \tag{6.25}$$

$$-p^\varepsilon_{i-1} \in \bigcap_{\delta>0} \overline{co} \bigcup_{|y-y^\varepsilon_{i-1}|\le\delta} \partial_y V^\varepsilon((i-1)\varepsilon, y), \qquad i = 1, 2, \ldots, N. \tag{6.26}$$

Now we shall show that (6.26) implies (6.24). Indeed, if this does not happen, then $-p^\varepsilon_{i-1}$ and $\partial_y V^\varepsilon((i-1)\varepsilon, y^\varepsilon_{i-1})$ can be strictly separated by a hyperplane, i.e., there exists $v \in \mathcal{H}$ and $\gamma > 0$ such that

$$(-p^\varepsilon_{i-1}, v) - \gamma > \max\{(p, v) : p \in \partial_y V^\varepsilon((i-1)\varepsilon, y^\varepsilon_{i-1})\} = (V^\varepsilon)^\circ((i-1)\varepsilon, y^\varepsilon_{i-1}; v).$$

Here $(V^\varepsilon)^\circ((i-1)\varepsilon, y^\varepsilon_{i-1}; v)$ is the generalized directional derivative of the function $y \mapsto V^\varepsilon((i-1)\varepsilon, y)$ at y^ε_{i-1} in the direction v. (For the above equality, we refer to [7, Prop. 2.1.2].) By the upper semicontinuity of $(V^\varepsilon)^\circ$ (see [7, Prop. 2.1.1]), we find $\delta' > 0$ such that

$$(-p^\varepsilon_{i-1}, v) - \gamma > (V^\varepsilon)^\circ((i-1)\varepsilon, y; v) \quad \text{for } |y - y^\varepsilon_{i-1}| \le \delta'.$$

Now using the definition of the generalized gradient, we find

$$\begin{aligned}(-p^\varepsilon_{i-1}, v) - \gamma &\ge \sup\left\{(p, v) : p \in \bigcup_{|y-y^\varepsilon_{i-1}|\le\delta'} \partial_y V^\varepsilon((i-1)\varepsilon, y)\right\} \\ &= \sup\left\{(p, v) : p \in \overline{co} \bigcup_{|y-y^\varepsilon_{i-1}|\le\delta'} \partial_y V^\varepsilon((i-1)\varepsilon, y)\right\}.\end{aligned}$$

So we have obtained that

$$-p^\varepsilon_{i-1} \notin \overline{co} \bigcup_{|y-y^\varepsilon_{i-1}|\le\delta'} \partial_y V^\varepsilon((i-1)\varepsilon, y),$$

which contradicts (6.26).

Finally, we may rewrite (6.25) as

$$u^\varepsilon_i \in \partial h^*(B^* p^\varepsilon_{i-1}), \qquad i = 1, 2, \ldots, N.$$

But the above inclusion combined with (6.24) gives (6.5), and the proof of Theorem 6.1 is complete.

Remark 6.1. The feedback law (6.5) is expressed with the aid of the Trotter scheme (2.4). A different feedback law is given, at least formally, by scheme (2.5) (via the function $\overline{W^\varepsilon}$). However, in this case, even the existence of an optimal N-tuple for the approximating problem (P^ε) remains an open problem.

7. Concluding remarks. Theorems 5.1 and 6.1 apply to the control systems governed by (4.2) where $0 \in \beta(0)$. Thus the case of the control of the parabolic obstacle problem with obstacle and boundary values 0 as well as that of the control of semilinear parabolic equations are covered (see again §4 for details).

Let us emphasize that Theorems 5.1 and 6.1 must be regarded together. So, the knowledge of the function V^ε allows us to construct an approximately optimal control for problem (P_1) by using feedback law (6.5). In this context, pay attention to the fact that formula (6.5) is merely a *necessary* condition of optimality for problem (P_1^ε). But so is the Pontryagin maximum principle and, nevertheless, in many specific situations, this offers us an effective tool for constructing an optimal control.

Let us now see how we can use formula (6.5) to obtain an approximately optimal control for problem (P_1). Clearly, (6.5) is an exact feedback law for problem (P_1^ε). It may also be regarded as an approximate feedback law for problem (P_1), which works as follows: Starting with the initial state y^0, we select $u_1^\varepsilon \in \partial h^*(-B^*\partial_y V^\varepsilon(0, y^0))$ (the first component of an optimal N-tuple for (P_1^ε)). Next we introduce the controller $u^\varepsilon(t) = u_1^\varepsilon$ on $(0, \varepsilon]$ in the real system. (For the real system we may choose either the continuous model described by equation (6.1) or the discrete model given by scheme (5.2).) Now we assimilate y_1^ε in (6.5) with the state of the real system observed at the moment $t = \varepsilon$; let us denote it by $y^\varepsilon(\varepsilon)$. With y_1^ε so specified, we further select $u_2^\varepsilon \in \partial h^*(-B^*\partial_y V^\varepsilon(\varepsilon, y^\varepsilon(\varepsilon)))$ and we proceed as above. Consequently, we have the following synthesis scheme:

$$(7.1) \qquad u^\varepsilon(t) = u_i^\varepsilon \in \partial h^*(-B^*\partial_y V^\varepsilon((i-1)\varepsilon, y^\varepsilon((i-1)\varepsilon))) \quad \text{on}\,((i-1)\varepsilon, i\varepsilon],$$

where y^ε is the state of the real system corresponding to the initial state y^0 and the input u^ε. So, the control u^ε obtained in this way is a feedback control.

Now let us look at formula (6.5) from another point of view (which does not require the external concept of observation). According to Theorem 5.1 we can construct an approximately optimal control for problem (P_1) by starting with a solution $(u_1^\varepsilon, u_2^\varepsilon, \ldots, u_N^\varepsilon)$ of problem (P_1^ε). But, by virtue of Theorem 6.1, any such solution verifies the inclusions (6.5). Therefore to calculate $(u_1^\varepsilon, u_2^\varepsilon, \ldots, u_N^\varepsilon)$, we consider the system

$$(7.2) \qquad \begin{cases} u_i^\varepsilon \in \partial h^*(-B^*\partial_y V^\varepsilon((i-1)\varepsilon, y_{i-1}^\varepsilon)), \\ y_i^\varepsilon = (I + \varepsilon\beta)^{-1}(I + \varepsilon A)^{-1}(y_{i-1}^\varepsilon + \varepsilon B u_i^\varepsilon), \quad i = 1, 2, \ldots, N, \\ y_0^\varepsilon = y^0. \end{cases}$$

Thus, starting with y^0, one obtains alternatively all the components of two N-tuples $(u_1^\varepsilon, u_2^\varepsilon, \ldots, u_N^\varepsilon)$ and $(y_1^\varepsilon, y_2^\varepsilon, \ldots, y_N^\varepsilon)$. Here is the order in which the calculations are made:

$$y^0, u_1^\varepsilon, y_1^\varepsilon, \ldots, y_{N-1}^\varepsilon, u_N^\varepsilon, y_N^\varepsilon.$$

Certainly, it is possible to find N-tuples $(u_1^\varepsilon, u_2^\varepsilon, \ldots, u_N^\varepsilon)$ verifying (7.2) (or (6.5)) that are not optimal for problem (P_1^ε), but all optimal N-tuples lie among the solutions of system (7.2). So, an adequate criterion of selection for the components u_i^ε must be established separately for each specific situation. Naturally, this is also true for synthesis scheme (7.1).

We point out that (7.2) is in essence an open-loop scheme: in order to determine the input u_i^ε, one uses not the value observed at the moment $t = (i-1)\varepsilon$ for the state of the real system, i.e., $y^\varepsilon((i-1)\varepsilon)$, but the value y_{i-1}^ε calculated by (5.2). Moreover, as we have seen in the proof of Theorem 6.1, $\partial_y V^\varepsilon((i-1)\varepsilon, y_{i-1}^\varepsilon) \ni -p_{i-1}^\varepsilon$, where the N-tuple $(p_0^\varepsilon, p_1^\varepsilon, \ldots, p_{N-1}^\varepsilon)$ may be viewed as a dual of $(y_1^\varepsilon, y_2^\varepsilon, \ldots, y_N^\varepsilon)$ in a Pontryagin-type maximum principle implicitly contained in the proof. In fact, system (7.2) is equivalent to a maximum principle for (P_1^ε). However, in comparison with the continuous case, the attempt to establish an explicit maximum principle for the discrete problem (P_1^ε) leads to some supplementary difficulties. In the discrete case, it is a real problem to give a sense to the adjoint scheme (which must be verified by p_i^ε) even for the control of the parabolic obstacle problem or that of semilinear parabolic equations (situations that were completely treated in the continuous case by Barbu in [2]).

Any optimal pair (u^*, y^*) for problem (P_1) formally satisfies the following continuous variant of (7.2):

$$\text{(7.3)} \qquad \begin{cases} u^*(t) \in \partial h^*(-B^* \partial_y V(t, y^*(t))) \quad \text{a.e.}\, t \in (0,T), \\ y^{*\prime} + Ay^* + \beta(y^*) \ni Bu^*, \\ y^*(0) = y^0, \end{cases}$$

where V is the optimal value function associated with (P_1). If we regard (7.3) as an open-loop system, we are led to a very complicated equation whose unknown is y^*. This happens because, in the first inclusion of (7.3), $u^*(t)$ depends on $y^*(t)$. In other words, the unknown $y^*(t)$ also appears in the right-hand side, incorporated in a strongly nonlinear term. The situation is different in the discrete case. Here u_i^ε depends on y_{i-1}^ε (previously calculated) and not on y_i^ε. Now substituting u_i^ε in (5.2), we no longer obtain the unknown y_i^ε in the right-hand side. Scheme (5.2) preserves its characteristic: the calculation of y_i^ε is reduced to the solution of an elliptic problem followed by the inversion of a graph in $\mathbf{R}^2$.

For V^ε, (1.11), (1.12) indicate two alternative ways to calculate it: one by a direct approach of a relatively simple type of Hamilton–Jacobi equation and the other by solving some minimization problems on $\mathcal{U}$.

REFERENCES

[1] V. Arnăutu, *Numerical results for a product formula approximation of Hamilton–Jacobi equation*, Internat. J. Computer Math., to appear.

[2] V. Barbu, *Optimal Control of Variational Inequalities*, Res. Notes Math., 100, Pitman, London, 1984.

[3] ———, *A product formula approach to nonlinear optimal control problems*, SIAM J. Control Optim., 26 (1988), pp. 496–520.

[4] ———, *Approximation of the Hamilton–Jacobi equations via Lie–Trotter product formula*, Control Theory Adv. Tech., 4 (1988), pp. 189–208.

[5] ———, *The fractional step method for a nonlinear distributed control problem*, in Differential Equations and Control Theory, V. Barbu, ed., Pitman Res. Notes Math. Ser., 250, Longman Scientific and Technical, Harlow, Essex, 1991, pp. 7–16.

[6] V. Barbu and T. Precupanu, *Convexity and Optimization in Banach Spaces*, 2nd ed., Editura Academiei, Bucureşti and D. Reidel, Dordrecht, Boston, Lancaster, 1986.

[7] F. H. Clarke, *Optimization and Nonsmooth Analysis*, John Wiley, New York, 1983.

[8] F. H. Clarke and R. B. Vinter, *The relationship between the maximum principle and dynamic programming*, SIAM J. Control Optim., 25 (1987), pp. 1291–1311.

[9] E. Hopf, *Generalized solutions of non-linear equations of first order*, J. Math. Mech., 14 (1965), pp. 951–973.

[10] P. D. Lax, *Hyperbolic systems of conservation laws* II, Comm. Pure Appl. Math., 10 (1957), pp. 537–566.

[11] C. Popa, *Trotter product formulae for Hamilton–Jacobi equations in infinite dimensions*, Differential Integral Equations, 4 (1991), pp. 1251–1268.

[12] D. Tătaru, *Viscosity solutions of Hamilton–Jacobi equations with unbounded nonlinear terms*, J. Math. Anal. Appl., 163 (1992), pp. 345–392.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 33, No. 4, pp. 1000–1016, July 1995

© 1995 Society for Industrial and Applied Mathematics
002

REGULARITY CONDITIONS FOR THE STABILITY MARGIN PROBLEM WITH LINEAR DEPENDENT PERTURBATIONS *

ANTONIO VICINO† AND ALBERTO TESI‡

Abstract. In this paper, the problem of continuity of the stability margin of a control system on problem input data is addressed. The case in which perturbations are linearly correlated is considered. It is shown that the existence of special points (called critical points) in the stability boundary manifold in parameter space plays a key role in the analysis of the problem. Several conditions, either sufficient or both necessary and sufficient, are given, ensuring continuity of the stability margin on problem data. The obtained conditions turn out to be easily checkable for practical applications. Numerical examples are presented to illustrate the proposed techniques.

Key words. regularity conditions, robust stability, stability margin, parametric perturbations

AMS subject classification. 93

1. Introduction.

The study of robustness properties of dynamic systems in the presence of parameter variations is a classical subject in control theory. The renewed and stronger interest of recent years in this area has produced many contributions (see [14] for a historical survey on the subject for both linear and nonlinear systems and [2], [10] for specific reference to linear systems). The (real) stability margin (or stability radius) of a control system is a key concept in this field. Roughly speaking, it represents a measure of the minimum perturbation which destabilizes a system designed to be stable in nominal operating conditions. Methods for determining this quantity for the case where plant transfer function coefficients are linear functions of physical real parameters may be found in [3], [7], [11], [15], [16], and [17], whereas [4], [12], and [19] provide algorithms for the case of nonlinear (polynomial or rational) dependence.

In this paper we make reference to the classic problem setting of Fig. 1. We assume that the plant model uncertainty is "highly structured," in the sense that plant coefficients depend on a vector of uncertain physical parameters p according to given relationships. The regularity problem addressed in this paper can be summarized as follows. Suppose that the designer has chosen a certain controller, i.e., a vector of controller coefficients λ^*, such that it guarantees a prescribed stability margin ensuring robust stability of the closed-loop system for a given uncertainty set in plant parameter space. Suppose that the controller parameters are slightly perturbed with respect to their nominal design value $\lambda = \lambda^*$. Can we predict whether the controller goes on doing its job, i.e., robustly stabilizes the system? Engineering intuition suggests that a negative answer to this question would be considered an indicator of a poor control design. In this perspective, the regularity assessment of a robust control problem with respect to input data, i.e., controller coefficients or parameters, becomes a mandatory requirement to perform a robust control design. Obviously, it would be particularly interesting to characterize real uncertainty structures for which the stability margin is continuous in the controller parameter space. Recently, results related to this problem have appeared in the literature. The possibility of existence of discontinuities of the stability margin on input data was pointed out in [1], where some examples are reported but no analysis is attempted. An analytical characterization of the behavior of the stability margin with respect to changes in system data has been given in [8] by employing the concept of strongly destabilizing perturbation. This property is strictly connected to the separating property used in the present

* Received by the editors July 20, 1992; accepted for publication (in revised form) January 18, 1994. This work was supported in part by funds of the Ministero della Università e della Ricerca Scientifica e Tecnologica.

† Facoltà di Ingegneria, Università di Siena, Via Roma 56, 53100 Siena, Italy.

‡ Dipartimento di Sistemi e Informatica, Università di Firenze, Via di Santa Marta 3, 50139 Firenze, Italy.

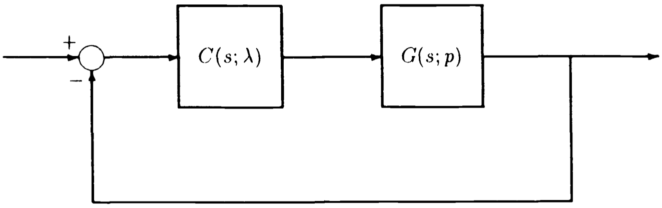


FIG. 1. *Linear feedback system.*

paper and in the preliminary conference paper [18]. Although the analysis in [8] covers general classes of perturbations, the results are not constructive, or at least how to build a technique for verifying the existence of a strongly destabilizing perturbation is not clear.

This paper represents a complete and detailed development of the theory originally presented in [18]. In particular, we exploit a geometric view of the obtained analytical results in the plant parameter space, allowing us to build an analysis technique for understanding how and why certain model parameterizations generate ill-conditioned problems, also suggesting ideas on changes to be made to regularize the problem. A distinguishing feature of this paper is that the different sufficient or necessary and sufficient conditions can be fruitfully applied to characterize completely the stability margin problem for several classes of real perturbations of continuous and discrete time systems widely investigated in the literature, such as interval feedback systems.

The uncertainty structure addressed in the present paper covers the case when plant transfer function coefficients depend affinely on the parameter vector p. Actually, this hypothesis, which may appear excessively restrictive at first glance, covers the vast majority of literature on parametric robust control. Even in this relatively simple situation, an analytic solution of the resulting stability margin nonlinear optimization problem is not available (see [3], [7], [11], [15], [16], and [17].

The regularity problem could be attacked by using general results on sensitivity analysis of nonlinear programming problems (see, e.g., [5]). Nevertheless, since the stability margin represents the minimum distance of the origin from the stability boundary manifold in parameter space, the analysis can be made considerably sharper. In fact, information on the behavior of the optimization problem can be obtained by studying the continuity of stability boundaries on problem data. Using this approach, we are able to prove a number of sufficient as well as necessary and sufficient conditions for the stability margin to be continuous on the problem data. The problem is strictly related to regularity and bifurcations of the boundary of stability in parameter space, an issue whose importance has long been recognized (see [13, pp. 404–406]). Actually, it happens that possible discontinuities of the problem solution with respect to data perturbations are strictly related to the presence of certain special points (called critical points) in the stability boundary manifold and to their behavior for small data perturbations. The basic results obtained show that under mild assumptions on the controller structure, the stability margin in the plant coefficient space is continuous on the controller parameters. One important implication of this result is that the robust stability problem for the class of interval feedback systems widely investigated in the recent literature on robust control is regular with respect to perturbations of the controller coefficients.

This paper is structured as follows. Section 2 introduces the stability margin problem and gives conditions for regularity of the boundary manifold. In §3 the main sufficient or necessary and sufficient conditions for continuity of the stability margin on problem data are given. In §4 it is shown how the results of §3 can be exploited to prove regularity of robustness problems for interval feedback systems. Section 5 reports some examples both taken from the literature and newly devised showing structural features of critical manifolds of the stability boundary in parameter space.

2. Stability margin and critical manifolds in parameter space.

2.1. The stability margin problem. Consider an nth-order polynomial whose coefficients depend on a perturbation parameter vector $p = (p_1, \dots, p_q)^T \in R^q$

$$\Delta(s;p) = s^n + c_{n-1}(p)s^{n-1} + \cdots + c_1(p)s + c_0(p) \tag{1}$$

(^{T} stands for transpose).

For notational simplicity, we will often identify p with the corresponding polynomial $\Delta(s;p)$. To avoid trivial degeneracies which are not of interest for our purposes, it is assumed that $n,\ q > 1$. We consider the case when polynomial coefficients are affine in the parameter p:

$$c(p) = T(p + p^o) + t \tag{2}$$

where $c(p) = [c_0(p), \dots, c_{n-1}(p)]^t \in R^n$ is the coefficient vector, T is a given (n, q) matrix, t is a given n-dimensional vector, and p^o is a given "nominal" parameter vector. We assume that the nominal polynomial $\Delta(s; 0)$ corresponding to the parameter p^o is Hurwitz (or stable); i.e., all its zeros have negative real part.

The (real) stability margin relative to the nominal parameter p^o is defined as the radius ρ of the maximal ball centered at p^o, such that all its interior points generate Hurwitz polynomials. If p^o is chosen as the origin of the parameter space, where parameter deviations p represent the coordinates, and the real and imaginary parts of $\Delta(j\omega; p)$ are denoted by $R(p;\omega)$ and $I(p;\omega)$, this problem can be formalized as a minimal-distance problem from the origin to the boundary manifold defined parametrically by $R(p;\omega) = 0$ and $I(p;\omega) = 0$ for $\omega \geq 0$ (see, e.g., [2] and [15]):

$$\begin{gathered}\rho = \min\nolimits_{p,\omega} \|p\| \\ \text{subject to} \\ \begin{cases} R(p;\omega) = 0 \\ I(p;\omega) = 0 \\ \omega \geq 0, \end{cases}\end{gathered} \tag{3}$$

where $R(p;\omega) = 0$ and $I(p;\omega) = 0$ are two equations affine in p whose coefficients are polynomials in ω,

$$R(p;\omega) = \sum_{i=1}^{q} a_i(\omega)p_i + a_0(\omega) = 0, \tag{4}$$

$$I(p;\omega) = \sum_{i=1}^{q} b_i(\omega)p_i + b_0(\omega) = 0. \tag{5}$$

Any point $(p^T, \omega) \doteq (p_1, \dots, p_q, \omega) \in R^{q+1}$ solving (3) will be called a *solution point* of the stability margin problem. By introducing a $(2, q+1)$ matrix $D(\omega)$, the boundary manifold equations can be written as

$$D(\omega)\pi = 0, \tag{6}$$

where

$$D(\omega) = \begin{bmatrix} a_0(\omega) & a_1(\omega) & \cdots & a_q(\omega) \\ b_0(\omega) & b_1(\omega) & \cdots & b_q(\omega) \end{bmatrix}, \qquad \pi = [1, p_1, \dots, p_q]^t. \tag{7}$$

We denote by $D_{ij}(\omega)$, $i = 0, \dots, q$, $i \neq j$ $(2,2)$ submatrices of the coefficient matrix $D(\omega)$

$$D_{ij}(\omega) = \begin{bmatrix} a_i(\omega) & a_j(\omega) \\ b_i(\omega) & b_j(\omega) \end{bmatrix} \tag{8}$$

and by $D_{0\omega}(p;\omega)$ the matrix

$$D_{0\omega}(p;\omega) = \begin{bmatrix} a_0(\omega) & \partial R(p;\omega)/\partial\omega \\ b_0(\omega) & \partial I(p;\omega)/\partial\omega \end{bmatrix}. \tag{9}$$

Note that since $a_0(\omega)$ and $b_0(\omega)$ are real and imaginary parts of the Hurwitz nominal polynomial $\Delta(s;0)$, there exist no ω such that $a_0(\omega) = b_0(\omega) = 0$, and hence the first column in D_{0j} and $D_{0\omega}$ cannot vanish.

With reference to the stability margin problem (3), we will consider entries of T, t, and p^o as the problem data. Obviously, these data may represent perturbations of the plant parameters p as well as perturbations induced by variations of the vector of controller coefficients λ. In the forthcoming subsection, we introduce the concept of the critical point of a manifold which plays a key role in assessing continuity of ρ as a function of the data.

2.2. Critical manifolds in parameter space. Let us introduce the manifold Γ_b^Ω in R^{q+1}

$$\Gamma_b^\Omega = \Gamma_o^\Omega \cup \Gamma^\Omega, \tag{10}$$

where

$$\Gamma_o^\Omega \doteq \{(p^T, \omega) \in R^{q+1} : R(p;\omega) = 0,\ \omega = 0\}, \tag{11}$$

$$\Gamma^\Omega \doteq \{(p^T, \omega) \in R^{q+1} : R(p;\omega) = 0,\ I(p;\omega) = 0,\ \omega > 0\}. \tag{12}$$

Γ_b^Ω will be referred to as the (stability) boundary manifold in R^{q+1}. The boundary manifold Γ_b in parameter space R^q, whose distance from the origin represents the stability margin, is strictly related to Γ_b^Ω. It is defined as

$$\Gamma_b = \Gamma_o \cup \Gamma, \tag{13}$$

where

$$\Gamma_o \doteq \{p \in R^q : c_0(p) = 0\}, \tag{14}$$

$$\Gamma \doteq \{p \in R^q : \exists\, \omega > 0 \text{ s.t. } (p^T, \omega) \in \Gamma^\Omega\}. \tag{15}$$

We notice that Γ (Γ_o) can be obtained as the image of Γ^Ω (Γ_o^Ω) under the identity rectangular operator $\mathcal{I} : R^{q+1} \to R^q$.

The study of sensitivity of ρ on problem data leads naturally to the investigation of the dependence of the manifolds Γ_o and Γ on the data; i.e., information about continuity of the distance ρ on problem data is strictly related to conditions under which these manifolds change smoothly with problem data. Since critical points of a manifold may undergo abrupt changes for arbitrarily small data perturbations, their study is an important issue for our problem.

Consider a manifold defined by the equation $f(x) = 0$, $x \in R^l$, $f : R^l \to R^m$, with $m \leq l$. Let ∇f_x be the Jacobian of the function $z = f(x)$.

$$\nabla f_x = \begin{bmatrix} \partial f_1/\partial x_1 & \cdots & \partial f_1/\partial x_l \\ \cdots & \cdots & \cdots \\ \partial f_m/\partial x_1 & \cdots & \partial f_m/\partial x_l \end{bmatrix}. \tag{16}$$

A point x^* of the manifold is said to be critical (or singular) if all the minors of order m of ∇f_x are null at that point. Points of a manifold which are not critical are called regular. A manifold whose points are all regular is called regular.

We study the regularity of Γ_b^Ω. However, since Γ_b is the representation of Γ_b^Ω as a parametric manifold in R^q, we observe that any regular (critical) point of Γ_b corresponds to a regular (critical) point of Γ_b^Ω.

A simple application of the above definitions to the maps defining manifolds Γ_o^Ω and Γ^Ω, given by (11) and (12), respectively, allows one to reach the following conclusions.

- Γ_o^Ω is regular (excluding the meaningless case $c_0(p) = 0$, $\forall p \in R^q$).
- Γ^Ω has singular points only if $\exists\,\omega^* > 0$ satisfying simultaneously the following algebraic equations.

$$|D_{0j}(\omega^*)| = 0, \qquad j = 1, \ldots, q, \tag{17}$$

where $|\cdot|$ denotes determinant. Notice that condition (17) implies that $R(p;\omega^*) = 0$ and $I(p;\omega^*) = 0$ represent the same $q-1$ dimensional hyperplane in parameter space. If (17) is solved for some strictly positive ω^*, according to the above definition, a critical point of Γ^Ω must satisfy the linear equation in p

$$|D_{0\omega}(p;\omega^*)| = 0. \tag{18}$$

Denoting by ω^* any real strictly positive solution of (17), critical points of Γ^Ω are the solutions of the system of two linear equations in p,

$$(p^T, \omega^*) \in R^{q+1} : \begin{cases} \sum_{i=1}^{q} a_i(\omega^*)p_i + a_0(\omega^*) = 0 \\ \sum_{i=1}^{q} \tilde{a}_i(\omega^*)p_i + \tilde{a}_0(\omega^*) = 0, \end{cases} \tag{19}$$

where the second equation represents (18) and

$$\tilde{a}_i(\omega) = a_0(\omega)b_i'(\omega) - b_0(\omega)a_i'(\omega), \qquad i = 0, 1, \ldots, q, \tag{20}$$

with $'$ denoting differentiation.

Assuming that $a_0(\omega^*) \neq 0$ in (19), we can characterize the set of critical points of Γ^Ω corresponding to each ω^* in terms of minors of order 2 of the coefficient matrix of the linear system (19). (If $a_0(\omega^*) = 0$, the first equation of (19) should be replaced by the second equation $I(p;\omega) = 0$ of (6).) Solutions of (19) in parameter space can be characterized according to three different cases.

Case 1.

$$\operatorname{rank} \begin{bmatrix} a_0(\omega^*) & a_j(\omega^*) \\ \tilde{a}_0(\omega^*) & \tilde{a}_j(\omega^*) \end{bmatrix} < 2, \qquad j = 1, \ldots, q. \tag{21}$$

In this case, corresponding to ω^*, a $q-1$ dimensional linear manifold of critical points is detected. Notice that condition (21) is equivalent to imposing that

$$|D_{0j}(\omega^*)|' = 0, \qquad j = 1, \ldots, q, \tag{22}$$

where $|D_{ij}(\omega^*)|'$ stands for $\left[\frac{d}{d\omega}|D_{ij}(\omega)|\right]_{\omega=\omega^*}$.

Case 2.

$$|D_{ij}(\omega^*)|' \neq 0 \quad \text{for some } i, j \neq 0,\ i \neq j. \tag{23}$$

In this case, it is concluded that system (19) determines a linear variety contained in Γ^Ω of dimension $q-2$.

Case 3.

$$|D_{ij}(\omega^*)|' = 0 \quad \forall i,j : i,j \neq 0,\, i \neq j \qquad \text{and} \tag{24}$$
$$|D_{0i}(\omega^*)|' \neq 0 \quad \text{for some } i \neq 0.$$

In this case, it can be easily verified that system (19) has no real solutions and hence Γ^Ω is a regular manifold.

Let us now introduce the sets of critical frequencies corresponding to the first and second cases examined above and the overall critical set Ω_c:

$$\Omega_{c_1} = \{\omega \in R^+ : |D_{0j}(\omega)| = 0, \text{ and } |D_{0j}(\omega)|' = 0,\ j = 1, \dots, q\}. \tag{25}$$

(26)
$$\Omega_{c_2} = \{\omega \in R^+ : |D_{0j}(\omega)| = 0,\ j = 1, \dots, q, \text{ and } |D_{ij}(\omega)|' \neq 0 \text{ for some } i,j \neq 0, i \neq j\},$$

$$\Omega_c = \Omega_{c_1} \cup \Omega_{c_2}. \tag{27}$$

We observe that Ω_{c_1}, Ω_{c_2}, and Ω_c have a finite number of elements. Denote by C_1 and C_2 conditions (25) and (26) defining the sets Ω_{c_1} and Ω_{c_2}, respectively, and by S_{c_1} and S_{c_2} the corresponding sets of critical points of Γ^Ω:

$$S_{c_{1(2)}} \doteq \{(p^T, \omega) \in \Gamma^\Omega \text{ s.t. } \omega \in \Omega_{c_{1(2)}}\}. \tag{28}$$

We say that a critical point of Γ^Ω is of type C_1 (C_2) if it belongs to S_{c_1} (S_{c_2}). The next lemma synthesizes the conclusions reached in the preceding discussion.

LEMMA 1. *Assume that the origin of parameter space is Hurwitz. The stability boundary manifold* Γ^Ω *is regular if and only if the set* Ω_c *is empty.*

Note that from a practical point of view, the determination of critical manifolds of Γ^Ω, if any, asks for the solution of an algebraic equation in ω. More precisely, one of the first q determinant equations of conditions C_1 or C_2 must be solved, checking successively if any of its strictly positive roots satisfies the remaining polynomial constraints of (25) or (26).

3. Regularity conditions for robust stability problems. To study regularity of the boundary manifold with respect to problem data, we introduce two vectors, x and y. The first vector contains problem variables, i.e., $x = (p_1, \dots, p_q, \omega)^T$, while the second vector $y \in R^u$ includes problem data consisting of the entries of matrix T, vector t, and nominal parameter p^o. We will be concerned only with the manifold Γ^Ω defined in (12). In fact, all solutions of the linear equation $R(p; 0) = 0$, i.e., points of the $q-1$ dimensional hyperplane Γ_o^Ω, are continuous on the problem data. Thus, in this section, with a certain abuse of notation, we will occasionally identify Γ_b^Ω with Γ^Ω. Moreover, we will refer to Γ or Γ^Ω according to the space (the parameter space R^q or the augmented space R^{q+1}) in which the boundary manifold is embedded. By taking into consideration the dependence on data y, the boundary manifold $\Gamma^\Omega(y)$ can be defined as follows:

$$\Gamma^\Omega(y) = \{x \in R^{q+1} : G(x; y) = 0,\ x_{q+1} > 0\}, \tag{29}$$

where the two components G_1 and G_2 of G are the real and imaginary parts of $\Delta(j\omega; p)$. We denote by $\rho(y)$ the stability margin as a function of input data y and assume that $G_1(x; y)$ and

$G_2(x;y)$ are continuous functions of the data y. We refer to $y=0$ as the nominal problem data, which means that our vector y is to be interpreted as a vector of data perturbations with respect to a fixed vector of data.

We say that the manifold $\Gamma^{\Omega}(y)$ is continuous at $y=0$ if for any point $x^* \in \Gamma^{\Omega}(y)$, there exist two variables x_i, x_j, defined implicitly by the two equations $G_1(x;y)=0$ and $G_2(x;y)=0$, which are continuous at $(x;y)=(x^*;0)$. We observe that, according to the previous definition, continuity of $\Gamma^{\Omega}(y)$ on y implies that a sufficiently small perturbation on the data y induces a small perturbation of points of the manifold $\Gamma^{\Omega}(y)$. We look for conditions under which the boundary manifold $\Gamma^{\Omega}(y)$ is continuous at $y=0$.

Let v_o be a vector in a given space R^l. We denote by $U_\varepsilon(v_o)$ a neighborhood of v_o of radius $\varepsilon>0$, i.e., $U_\varepsilon(v_o)=\{v \in R^l : \|v-v_o\| \leq \varepsilon\}$. Let $\nabla G_x(x;y)$ denote the Jacobian of the function G with respect to the variable x, and let $U_\varepsilon(0)$ be a neighborhood of the origin. Since the functions $G(x;y)$ and $\nabla G_x(x;y)$ are continuous in R^{q+u+1}, application of the implicit function theorem allows us to derive the following lemma.

LEMMA 2. *If $\nabla G_x(x;0)$ has row rank 2 for any $x \in \Gamma^{\Omega}(0)$, then $\Gamma^{\Omega}(y)$ is continuous at $y=0$.*

Proof. The proof is given in the appendix.

Remark 1. Note that continuity of $\Gamma_b^{\Omega}(y)$ at $y=0$ implies continuity of the boundary manifold in parameter space $\Gamma_b(y)$ at the same point. Since the stability margin $\rho(y)$ is the distance, according to some given norm, of the origin from $\Gamma_b(y)$ in parameter space, i.e., the minimum of the optimization problem (3), continuity of the manifold $\Gamma_b(y)$ implies continuity of $\rho(y)$ at $y=0$ whatever norm is used in (3).

From the above lemma and remark, we obtain the following theorem.

THEOREM 1. *The stability margin is continuous on input data at $y=0$ if Γ^{Ω} is a regular manifold.*

Remark 2. It is well known that the collection of critical values of a manifold mapped by a smooth map is a set of measure zero in the image space. This result is known as the Sard–Brown theorem (see, e.g., [9, pp. 10–11]). In our case, since we assume that $q>1$, the critical points in the boundary manifold Γ_b^{Ω} form a thin set in R^{q+1}. In particular, conditions C_1 and C_2 state that a necessary condition for the existence of critical points is that q algebraic equations in the unique unknown ω admit a common strictly positive solution. This implies that an arbitrarily small data perturbation (i.e., a perturbation of the coefficients of the q equations) generically destroys singularities. This does not mean that in general one cannot design a small perturbation affecting continuously critical points; rather, the set of perturbations destroying singularities is dense in the space of data.

The following theorem gives sufficient conditions for continuity of $\rho(y)$ improving those of Theorem 1.

THEOREM 2. *The stability margin $\rho(y)$ is continuous on the problem data at $y=0$ if at least one of the solution points of the stability margin problem for $y=0$ is either a regular point of the boundary manifold or a critical point of type C_2.*

Proof. The proof is given in the appendix.

Theorem 2 allows us to derive the following result on families of third-order polynomials.

COROLLARY 1. *Given a family of third-order polynomials defined by a matrix T of full column (or row) rank, the stability margin $\rho(y)$ is continuous on the data y.*

Proof. The proof is given in the appendix.

Before giving a theorem stating necessary and sufficient conditions for continuity of the stability margin on the problem data, we need a definition which recalls the concept of strongly destabilizing perturbation given in [8]. For the sake of clarity, in the rest of this section we will call a parameter p strictly unstable if at least one of the roots of the corresponding polynomial has strictly positive real part. Consider a solution point $x^* = (p^{*T}, \omega^*)$ of the stability margin problem.

DEFINITION. *A solution point $x^* = (p^{*T}, \omega^*)$ of the stability margin problem is called a separating point if any neighborhood $U_\varepsilon(p^*)$ contains both stable and strictly unstable points.*

The following lemma establishes a relationship between regularity and separating properties of solution points of the stability margin problem.

LEMMA 3. *A regular solution point of the stability margin problem is a separating point.*

Proof. The proof is given in the appendix.

We can now provide the following theorem.

THEOREM 3. *A necessary and sufficient condition for continuity of the stability margin at $y = 0$ is that the stability margin problem admit at least one separating solution point.*

Proof.

Necessity. Assume that $\rho(y)$ is continuous at $y = 0$. If there exists at least one regular solution point, then necessity follows immediately from Lemma 3. If the stability margin problem admits critical solution points only, necessity can be proven by contradiction. Consider an arbitrary solution point (p^{*T}, ω^*) and suppose that the critical manifold corresponding to $\omega = \omega^*$ does not separate stable from strictly unstable points in $U_\varepsilon(p^*)$. We show that there exists an arbitrarily small data perturbation which induces a noninfinitesimal perturbation of the stability margin. Consider the data perturbation generated by an arbitrarily small horizontal shift of the imaginary axis into the right half plane of the complex plane. This means to perform the change of variable $s \rightarrow \tilde{s} = s - \eta$, with $\eta > 0$ and arbitrarily small. For a given η, it is not difficult to realize that this perturbation can be generated as a sufficiently small polynomial coefficient perturbation. In this new situation, p^* becomes a stable parameter. In addition, this perturbation destroys the entire critical manifold. Thus, since points which are stable for $\eta = 0$ remain stable under the perturbation, $U_\varepsilon(p^*)$ contains only stable points. If several parameters p^* exist such that (p^{*T}, ω^*) is a solution point, the above argument holds for each of them. Hence, there exists $\eta > 0$ such that neighborhoods $U_\varepsilon(p^*)$ for all p^* contain only stable points. This means that all the solution points of the perturbed stability margin problem in parameter space must necessarily differ from the corresponding points for $\eta = 0$ for a noninfinitesimal quantity. This implies discontinuity of the stability margin.

Sufficiency. Let (p^{*T}, ω^*) be a separating solution point of the stability margin problem. For an arbitrarily small $\varepsilon > 0$, the separating property of (p^{*T}, ω^*) ensures that there exists a strictly unstable parameter $\hat{p} \in U_\varepsilon(p^*)$. Since the coefficients of the polynomial are continuous in the problem data and the roots of a polynomial are continuous in its coefficients, it follows that for any sufficiently small data perturbation, $\hat{p}$ remains strictly unstable. Thus, sufficiently small data perturbations preserve strictly unstable points $\hat{p}$ arbitrarily close to p^* and this proves that the stability margin $\rho(y)$ is continuous at $y = 0$.

Remark 3. Theorem 3 is somewhat similar to a result given in [8, p. 403]. Actually, since both results provide necessary and sufficient conditions for the same problem, they must be necessarily equivalent. However, it turns out that, as explained below, the separability concept used in Theorem 3 allows a direct characterization of continuity of the stability margin problem in terms of easily checkable conditions.

First, we notice that, if the stability margin problem admits at least one regular solution point or one critical solution point of type C_2, then Theorem 2 ensures continuity, without explicitly requiring the separating property. In the remaining cases, the condition of Theorem 3 requires that at least one of the $q - 1$ dimensional critical manifolds containing solution point(s) (p^{*T}, ω^*) of the stability margin problem (3) separates stable from strictly unstable points in a neighborhood of at least one point p^*. Since critical manifolds are hyperplanes in parameter space, verification of the separability property asked by Theorem 3 turns out to be easily tested in practice.

As a further observation, we notice that all the obtained conditions are easily checkable, and their verification provides information, if necessary, on how far the boundary manifold is from the case in which it has singularities. For example, the fact that there exist values of ω for which $|D_{0j}(\omega)| \approx 0$, $|D_{0j}(\omega)|' \approx 0$, $j = 1, \ldots, q$ in condition C_1 of (25) can be interpreted as an indicator that the problem may become critical to data perturbation.

Before concluding this section, we briefly summarize how the results obtained can be used to assess the regularity of the stability margin problem for a given family of perturbed polynomials. It is well known that many techniques for computing ρ, i.e., solving (3), are based on a one-dimensional search along the frequency positive axis (see, e.g., [2], [15], and [16]). We know that the domain of search $\Omega^+ \doteq [0, \infty)$ can be partitioned into three disjoint subsets, Ω_r, Ω_{c1}, and Ω_{c2}, such that

$$\Omega^+ = \Omega_r \cup \Omega_{c1} \cup \Omega_{c2}, \tag{30}$$

where Ω_r is the set of frequencies corresponding to regular points of Γ^Ω, while Ω_{c1} and Ω_{c2} are defined as in (25) and (26). Denote by ρ_r, ρ_{c1}, and ρ_{c2} the stability margins relative to the three sets of frequencies Ω_r, Ω_{c1}, and Ω_{c2}, i.e., solutions of problems like (3) with $\omega \in \Omega_r$, $\omega \in \Omega_{c1}$, and $\omega \in \Omega_{c2}$, respectively. It is clear that $\rho = \min\{\rho_r, \rho_{c1}, \rho_{c2}\}$. By applying Theorems 1 and 2, the following condition ensuring continuity of the stability margin on problem data is obtained:

$$\rho = \min\{\rho_r, \rho_{c2}\} \leq \rho_{c1}. \tag{31}$$

On the other hand, if

$$\rho = \rho_{c1} < \min\{\rho_r, \rho_{c2}\}, \tag{32}$$

Theorem 3 requires testing if certain $q - 1$ dimensional critical hyperplanes in parameter space separate stable from strictly unstable points around the corresponding solution point(s) of the stability margin problem relative to Ω_{c1} (see Remark 3).

4. Regularity conditions for the robust stability problem in plant coefficient space. In this section, we show how the results obtained in the previous sections can be exploited whenever the changes in system data are due to controller coefficient variations. In particular, we will study regularity of the robust stability problem in plant coefficient space. The case when plant coefficients are linear affine in the uncertain parameters falls in the general problem setting developed in the previous sections, because in this case, for any fixed controller, the closed-loop characteristic polynomial coefficients are affine linear in the parameters.

With reference to Fig. 1, we assume that $G(s;p)$ is a strictly proper plant and $C(s;\lambda)$ is a rational controller of given order

$$G(s;p) = \frac{N(s;p)}{D(s;p)} = \frac{\sum_{i=0}^{m}(p_{n+i+1} + p_{n+i+1}^o)s^i}{s^n + \sum_{i=1}^{n}(p_i + p_i^o)s^{i-1}}, \qquad m < n, \tag{33}$$

$$C(s;\lambda) = \frac{N^c(s;\lambda)}{D^c(s;\lambda)} = \frac{\sum_{i=0}^{l}\lambda_{r+i+1}s^i}{\sum_{i=0}^{r}\lambda_i s^i}, \qquad l \leq r, \tag{34}$$

where the parameters p coincide with the plant transfer function coefficients and $\lambda = (\lambda_0, \dots, \lambda_r, \lambda_{r+1}, \dots, \lambda_{r+l+1})^T$ is the controller coefficient vector. The closed-loop characteristic polynomial is

$$\Delta(s;p;\lambda) \doteq D^c(s;\lambda)D(s;p) + N^c(s;\lambda)N(s;p), \tag{35}$$

where, with a slight modification of notation with respect to (1), the arguments of Δ include the controller coefficient λ. This implies that T, t, and p^o are now suitable functions of λ. The nominal closed-loop characteristic polynomial and the controller numerator and denominator polynomials can be written as the sum of even and odd parts (subscripts e and o, respectively):

$$\begin{aligned} \Delta(s;0;\lambda) &\doteq \Delta_e(s^2;\lambda) + s\Delta_o(s^2;\lambda), \\ N^c(s;\lambda) &\doteq N_e^c(s^2;\lambda) + sN_o^c(s^2;\lambda), \\ D^c(s;\lambda) &\doteq D_e^c(s^2;\lambda) + sD_o^c(s^2;\lambda). \end{aligned} \tag{36}$$

According to these definitions, the coefficient matrix $D(\omega;\lambda)$ becomes

(37)

$$D(\omega;\lambda) = \begin{bmatrix} \Delta_e(\omega^2;\lambda) & D_e^c(\omega^2;\lambda) & -\omega^2 D_o^c(\omega^2;\lambda) & \dots & N_e^c(\omega^2;\lambda) & -\omega^2 N_o^c(\omega^2;\lambda) & \dots \\ \omega\Delta_o(\omega^2;\lambda) & \omega D_o^c(\omega^2;\lambda) & \omega D_e^c(\omega^2;\lambda) & \dots & \omega N_o^c(\omega^2;\lambda) & \omega N_e^c(\omega^2;\lambda) & \dots \end{bmatrix}.$$

Consider now the nominal characteristic polynomial $\Delta(s;0;\lambda)$ and define the stability domain Λ in controller coefficient space as the set of coefficients λ stabilizing the nominal plant, i.e.,

$$\Lambda \doteq \{\lambda \in R^{r+l+2} : \Delta(s;0;\lambda) \text{ is Hurwitz}\}. \tag{38}$$

Assuming that the set Λ is nonempty, we can give the following theorem on regularity of the stability margin $\rho = \rho(\lambda)$ in plant coefficient space.

THEOREM 4. 1) *If $m = 0$, i.e., the plant family is an all pole family, $\rho(\lambda)$ is continuous $\forall\lambda \in \Lambda$ such that $C(s;\lambda)$ does not have multiple purely imaginary poles.*

2) *If $m \geq 1$, $\rho(\lambda)$ is continuous $\forall\, \lambda \in \Lambda$.*

Proof. The proof is given in the appendix.

Observation. The above theorem holds independently of the norm used in coefficient space. In particular, when an l_∞ norm is taken into account, Theorem 4 ensures regularity of the stability margin problem for the well-known and widely investigated class of interval feedback systems.

5. Numerical examples. In this section we present three numerical examples. The first example is taken from [1]. In this case, lack of regularity is detected analytically by applying Theorem 3. A data perturbation producing a discontinuity in the stability margin is generated according to the technique used to prove necessity of Theorem 3. The second example is new. A fourth-order polynomial with two uncertain parameters is considered. A straight line of critical points is detected, showing that is is possible to generate a discontinuity of the stability margin ρ with respect to problem data. The third example reports the study of the behavior of the stability margin of a feedback system with an interval plant and a controller as a function of one coefficient of the controller. Of course, in this case, as predicted by Theorem 4, it is found that ρ is continuous as a function of the considered coefficient. In the second and third examples, perturbations on T, t, and p^o will be denoted by δT, δt, and δp^o, respectively, and will be assumed to enter additively in T, t, and p^o.

Example 1 [1]. Consider the polynomial

$$\begin{aligned} \Delta(s;p) = s^4 &+ 20(1-p_2)s^3 + (44 + 2a + 10p_1 - 40p_2)s^2 \\ &+ (20 + 8a + 20ap_1 - 20p_2)s + 5a^2(1+2p_1), \end{aligned} \tag{39}$$

where $a = 3 + \sqrt{2}$. The data are given by

(40)
$$T = \begin{bmatrix} 10a^2 & 20a & 10 & 0 \\ 0 & -20 & -40 & -20 \end{bmatrix}^T; \ t = [5a^2 \quad 20+8a \quad 2a+44 \quad 20]^T; \ p^o = [0 \quad 0]^T.$$

The boundary manifold Γ_b is made of points solving the following equation system with $\omega \geq 0$:

$$(41) \qquad \begin{cases} R(p;\omega) = 10(a^2 - \omega^2)p_1 + 40\omega^2 p_2 + 5a^2 - \omega^2(44+2a) + \omega^4 = 0, \\ I(p;\omega) = \omega[20ap_1 + 10(\omega^2 - 2)p_2 + 20 + 8a - 20\omega^2] = 0. \end{cases}$$

The manifolds Γ_o and Γ have the following expressions:

$$(42) \qquad \Gamma_o = \{p : p_1 + 0.5 = 0\},$$

$$(43) \qquad \Gamma = \{p : 5(a-1)p_1 + 20p_2 - 2(11-a) = 0\} \cup \{p : p_2 + a/10 - 1 = 0\}.$$

The stability domain is represented by the dashed area in Fig. 2(a). To investigate the existence of critical points we consider the following determinants:

$$(44) \qquad \begin{cases} |D_{o1}(\omega)| = 20\omega[(a-10)\omega^4 + 2(4a^2 - 20a + 5)\omega^2 + a^2(a-10)], \\ |D_{o2}(\omega)| = 20\omega[\omega^6 - (2a+5)\omega^4 + (4 - 14a + 5a^2)\omega^2 - 5a^2]. \end{cases}$$

Easy computations show that the equations $|D_{o1}(\omega)| = 0$, $|D_{o2}(\omega)| = 0$ admit a common positive root (of multiplicity 2) at $\omega^* = \sqrt{a}$. Since this root is multiple, it also solves the second equation of (25) and hence it satisfies condition C_1. The straight line $5(a-1)p_1 + 20p_2 - 2(11-a) = 0$, a subset of Γ, is found to be a set of critical points. The solution of the distance problem (3) according to the l_∞ norm in parameter space is attained at the nonseparating point $p^* = [(7-a)/5, (7-a)/5]^T \approx [0.234, 0.234]^T$ of this critical line (see Fig. 2(a)). The corresponding stability margin is $\rho \approx 0.234$. Since p^* is nonseparating, we deduce by Theorem 3 that there exists an infinitesimal data perturbation such that the stability margin changes are of a finite quantity. Figure 2(b) reports the perturbed boundary manifolds (1) and (2) corresponding to perturbations given according to the technique used to prove necessity of Theorem 3, with $\eta = 0.001$ and $\eta = 0.01$, respectively. The obtained stability margins are $\rho \approx 0.422$ and $\rho \approx 0.456$, respectively. The corresponding maximal stability balls are also shown in Fig. 2(b).

Example 2. Consider the polynomial

$$(45) \quad \Delta(s;p) = s^4 + (p_2+3)s^3 + (p_1+5.5)s^2 + (p_1+p_2+4.5)s + 3p_1 - p_2 + 5.5$$

with T, t, and p^o given by

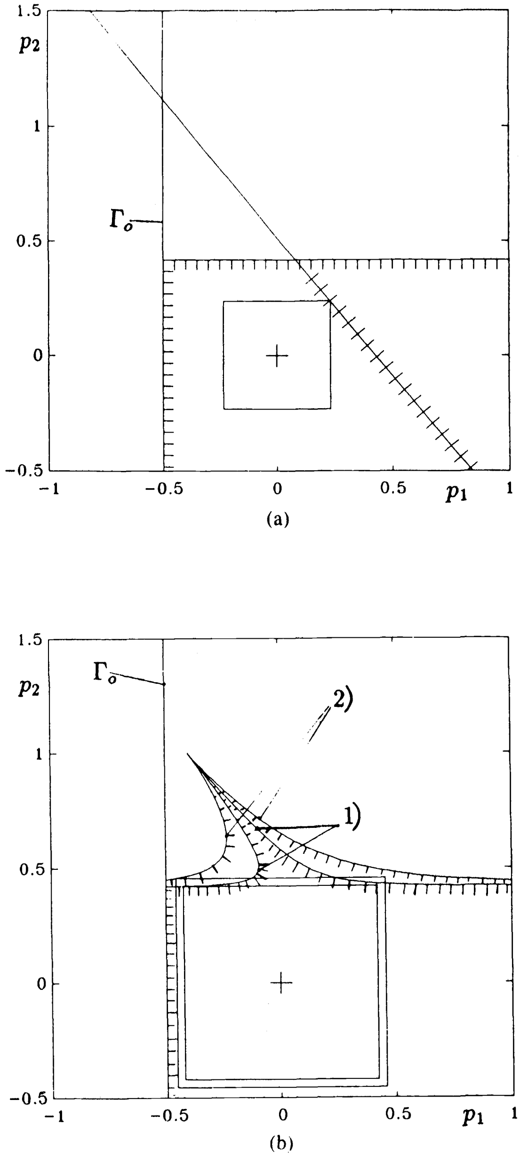
$$(46) \qquad T = \begin{bmatrix} 3 & 1 & 1 & 0 \\ -1 & 1 & 0 & 1 \end{bmatrix}^T; \quad t = [5.5 \quad 4.5 \quad 5.5 \quad 3]^T; \quad p^o = [0 \quad 0]^T$$

and data perturbation

$$(47) \qquad \delta T = \begin{bmatrix} 0 & r & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}^T; \quad \delta t = [0 \quad 1.5r \quad 0 \quad 0]^T; \quad \delta p^o = [0 \quad 0]^T.$$

This problem is obtained from the study of robust stability of the feedback control system of Fig. 1 with a plant family given by

$$(48) \qquad G(s;p) = \frac{(p_1 + 1.5)}{s^4 + (3+p_2)s^3 + (5.5+p_1)s^2 + (3+p_2)s + 5.5 + 3p_1}$$



(a)

(b)

FIG. 2. (a) *Stability domain for nominal data of example* 1. (b) *Stability domains with data perturbations of example* 1.

and a proportional controller, whose unity gain represents the problem data subject to perturbation

$$C(s;\lambda) = 1 + \lambda, \tag{49}$$

where $\lambda = r$. The nominal manifolds Γ_o and Γ are given by

$$\Gamma_o = \{p : 3p_1 - p_2 + 5.5 = 0\}, \tag{50}$$

$$\Gamma = \{p : p_2 + 2 = 0\} \cup \{p : p_1 - p_2 - 1.5 = 0\}. \tag{51}$$

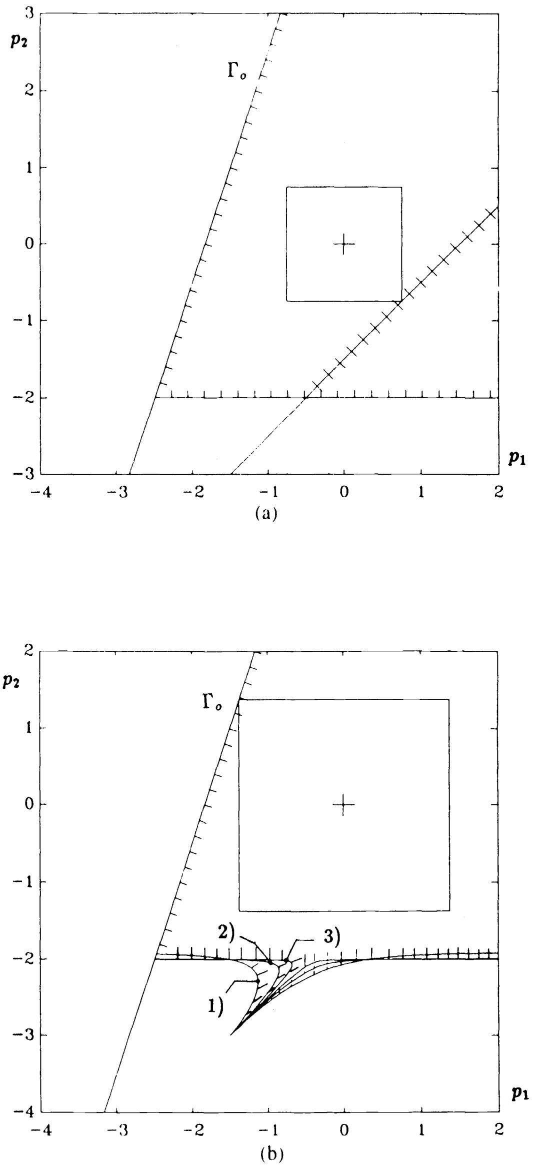


FIG. 3. (a) *Stability domain for nominal data of example* 2. (b) *Stability domains with data perturbations of example* 2.

The dashed area in Fig. 3(a) represents the stability domain in the parameter plane. Note that the straight line of equation $p_1 - p_2 - 1.5 = 0$ separates two subsets of the domain of stability.

The determinants $|D_{o1}|$ and $|D_{o2}|$ have the following expressions:

$$(52) \qquad \begin{cases} |D_{o1}(\omega)| = -2\omega(\omega^2 - 2)^2, \\ |D_{o2}(\omega)| = -\omega(\omega^2 - 2)^2(\omega^2 - 2.5). \end{cases}$$

It can be checked from (52) that $\omega^* = \sqrt{2}$ satisfies condition C_1; i.e., there exists a straight line of critical points described by the equation $p_1 - p_2 - 1.5 = 0$.

The solution of the distance problem (3) according to the l_∞ norm in parameter space is attained at the nonseparating point $p^* = (0.75, -0.75)^T$ of this critical line (see Fig. 3(a)). The corresponding stability margin is $\rho = 0.75$.

Since the point $p^* = (0.75, -0.75)^T$ is nonseparating, Theorem 3 ensures that a discontinuity in the stability margin can be generated by a suitable perturbation. Actually, it turns out that arbitrarily small values of r destroy the critical manifold, inducing a discontinuity in ρ as a function of r. Fig. 3(b) reports the perturbed boundary manifolds (1), (2), and (3) obtained for $r = 0.1, 0.01$, and 0.001, respectively, and the maximal stability ball corresponding to $\rho = 1.375$ obtained for the given data perturbations.

Example 3. Consider the feedback control system of Fig. 1, where $G(s;p)$ is an interval plant

$$G(s;p) = \frac{3+p_3}{s^2[s^2+(3+p_2)s+10+p_1]} \tag{53}$$

and the controller $C(s;\lambda)$ is given by

$$C(s;\lambda) = \frac{1+4s}{1+\lambda s}. \tag{54}$$

The closed-loop characteristic polynomial is

$$\Delta(s;p;\lambda) = \lambda s^5+[1+\lambda(3+p_2)]s^4+[3+p_2+\lambda(10+p_1)]s^3+(10+p_1)s^2+(4p_3+12)s+3+p_3. \tag{55}$$

One can compute that the stability domain in the controller parameter space is the segment $\Lambda \approx (0, 2.286)$. For any given $\lambda \in \Lambda$, the equations defining the boundary manifold Γ_b are $(\omega \geq 0)$

$$\begin{cases} R(p;\omega) = -\omega^2 p_1 + \lambda\omega^4 p_2 + p_3 + 3 - 10\omega^2 + \omega^4(1+3\lambda) = 0, \\ I(p;\omega) = \omega[-\lambda\omega^2 p_1 - \omega^2 p_2 + 4p_3 + 12 - (3+10\lambda)\omega^2 + \lambda\omega^4] = 0. \end{cases} \tag{56}$$

The determinants $|D_{o1}|$, $|D_{o2}|$, and $|D_{o3}|$ are given by

$$\begin{cases} |D_{o1}(\omega)| = -3\omega^3[\lambda^2\omega^4 + \omega^2 + \lambda - 4], \\ |D_{o2}(\omega)| = -\omega^3[\lambda^2\omega^6 + (1-10\lambda^2)\omega^4 + 2(6\lambda - 5)\omega^2 + 3], \\ |D_{o3}(\omega)| = \omega^3[(11\lambda + 4)\omega^2 + \lambda - 37]. \end{cases} \tag{57}$$

Since for any $\lambda \in \Lambda$, it can be readily verified that equations $|D_{o1}(\omega)| = 0$, $|D_{o2}(\omega)| = 0$, and $|D_{o3}(\omega)| = 0$ do not admit any positive common root, neither condition C_1 nor C_2 is satisfied so that $\Omega_c(\lambda) = \emptyset$, $\forall\, \lambda \in \Lambda$. Hence, as predicted by Theorem 4, the stability margin is continuous on λ.

6. Conclusions.

In this paper the problem of continuity of the stability margin on problem data has been addressed. The case in which the coefficients of the polynomials of an uncertain family are affine in a vector of physical parameters has been considered. It has been shown how the continuity problem is related to the presence of critical points in the manifold bounding the domain of stability in parameter space. Sufficient and necessary and sufficient conditions which are easily computable in practice have been given. By use of these conditions, it has been shown that the important and widely studied class of interval feedback systems enjoys the regularity property. Numerical examples have been illustrated to show applications of the conditions obtained.

Appendix.

Proof of Lemma 2. Suppose that for a given x^* such that $G(x^*;0) = 0$, $\nabla G_x(x^*;0)$ has row rank 2. By the implicit function theorem, there exist at least two variables, x_k and x_l, $k \neq l$, such that the equation defining $\Gamma^\Omega(y)$, i.e., $G(x;y) = 0$, is uniquely solvable with respect to x_k and x_l, for any $y \in U_\varepsilon(0)$, a sufficiently small neighborhood of $y = 0$, and any $x \in U_\eta(x^*)$, a sufficiently small neighborhood of x^*. This means that there exist functions $x_k = x_k(x_1, \ldots, x_{k-1}, x_{k+1}, \ldots, x_{l-1}, x_{l+1}, \ldots, x_{q+1}; y)$ and $x_l = x_l(x_1, \ldots, x_{k-1}, x_{k+1}, \ldots, x_{l-1}, x_{l+1}, \ldots, x_{q+1}; y)$, $\forall y \in U_\varepsilon(0)$ and $\forall x \in U_\eta(x^*)$. In addition, the same theorem ensures that these functions $x_k(\cdot;\cdot)$, $x_l(\cdot;\cdot)$ are continuous at $(x;y) = (x^*;0)$. Thus, the statement of the lemma follows immediately from the rank assumption on $\nabla G(x;0)$ in all the domain of definition of the function $z = G(x;0)$.

Proof of Theorem 2. Assume that for $y = 0$ there exists at least one regular point solving the minimal distance problem (3). Consider a sufficiently small neighborhood of the origin $U_\varepsilon(0)$ in the space of data. By the implicit function theorem (see also proof of Lemma 2), $\forall y \in U_\varepsilon(0)$ all regular points of $\Gamma^\Omega(y)$, and hence also the solution point(s) of the stability margin problem (3), are infinitesimally close to points of $\Gamma^\Omega(0)$, implying continuity of $\rho(y)$ at $y = 0$.

Suppose now that at least one solution point, say $\bar{x} = (\bar{p}^T, \bar{\omega})$, satisfies condition C_2 in (26). Of course, if a regular solution point also exists, the theorem follows from the preceding argument. In the other case, the critical set corresponding to $\omega = \bar{\omega}$ is a $q-2$ dimensional linear manifold in parameter space (see the second case in the discussion preceding Lemma 1). Since $\bar{x}$ is a solution of $G(x;0) = 0$, which is a system of equations linear in the parameter p and polynomial in ω, any neighborhood $U_\varepsilon(\bar{x})$ contains a $q-1$ dimensional set of solutions of $G(x;0) = 0$. This implies that even if the critical solution point $\bar{x}$ is destroyed by a small data perturbation, regular points of Γ^Ω in arbitrarily small neighborhoods of $\bar{x}$ ensure continuity of the distance of the origin of parameter space from Γ.

Proof of Corollary 1. Consider a third-order polynomial with q uncertain parameters. It is easy to check that a necessary condition for satisfying C_1 is that there exist values of ω such that

$$(58)\qquad \begin{cases} |D_{ij}(\omega)| = 0, \\ |D_{ij}(\omega)|' = 0, \qquad \forall i,j = 1, \ldots, q, \quad i \neq j. \end{cases}$$

Computation of determinants $|D_{ij}(\omega)|$ and their derivatives gives

$$(59)\qquad \begin{cases} |D_{ij}(\omega)| = \omega[(T_{1i}T_{2j} - T_{2i}T_{1j}) + (T_{2i}T_{3j} - T_{3i}T_{2j})\omega^2], \\ |D_{ij}(\omega)|' = (T_{1i}T_{2j} - T_{2i}T_{1j}) + 3(T_{2i}T_{3j} - T_{3i}T_{2j})\omega^2, \end{cases}$$

where T_{ij} is the generic entry of the matrix T. Therefore, solutions ω of (58) different from zero exist if and only if

$$(60)\qquad \begin{cases} T_{1i}T_{2j} - T_{2i}T_{1j} = 0, \\ T_{2i}T_{3j} - T_{3i}T_{2j} = 0 \quad \forall i,j = 1, \ldots, q, \quad i \neq j. \end{cases}$$

It can easily be checked that if there exists i such that $T_{2i} \neq 0$, then verification of (60) necessarily implies that matrix T has column (or row) rank equal to 1, which contradicts the assumption on T. On the other hand, consider any of the infinite solutions of (60) with $T_{2i} = 0$, $\forall i = 1, \ldots, q$. In this case, the manifold Γ^Ω is given by a unique $q-1$ dimensional hyperplane defined by $R(p;\bar{\omega}) = 0$, where $\bar{\omega}$ is the only strictly positive solution of $b_o(\omega) = 0$. This means that Γ^Ω is a regular manifold, excluding the existence of critical manifolds.

Since any other value of T_{ij} different from those satisfying (60) cannot verify condition C_1, it follows from Theorem 2 that the stability margin $\rho(y)$ is continuous in the problem data y.

Proof of Lemma 3. Assume that $x^* = (p^{*T}, \omega^*) \in \Gamma_b^\Omega$ is a regular solution point of the stability margin problem.

If $x^* \in \Gamma_o^\Omega$, the separating property follows from the fact that the hyperplane Γ_o separates points corresponding to polynomials such that $c_0(p) > 0$ and $c_o(p) < 0$ and the fact that any polynomial with a negative constant term is strictly unstable.

Suppose now that $x^* \in \Gamma^\Omega$. Consider the Hurwitz determinant of order $n-1$ of the polynomial $\Delta(s;p)$ and denote it by $H_{n-1}(p)$. Since $p^* \in \Gamma$ represents a polynomial which has at least one pure imaginary root and all the remaining ones in the left half-plane, we know by Orlando's formula [6] that $p^* \in \Gamma^H$, where

$$\Gamma^H = \{p \in R^q : H_{n-1}(p) = 0\}. \tag{61}$$

Moreover, continuity of roots of a polynomial on its coefficients implies that Γ coincides with Γ^H locally; that is,

$$\exists\, \varepsilon > 0 : \Gamma^H \cap U_\varepsilon(p^*) \equiv \Gamma \cap U_\varepsilon(p^*). \tag{62}$$

Thus, p^* is a regular point of Γ^H. As a consequence, Γ^H admits at $p = p^*$ a tangent hyperplane. In turn, this implies that any neighborhood $U_\varepsilon(p^*)$ of p^* contains points p where $H_{n-1}(p) > 0$ and points where $H_{n-1}(p) < 0$. Thus, $U_\varepsilon(p^*)$ contains both stable and strictly unstable points, implying that p^* is a separating point.

Proof of Theorem 4. i) Let us assume that $C(s;\lambda_o)$ does not have multiple imaginary roots and that $\lambda_o \in \Lambda$. First, we show by contradiction that the set $\Omega_c(\lambda_o)$ is empty. Suppose that $\Omega_c(\lambda_o)$ is nonempty, and take an element $\omega_o \in \Omega_c(\lambda_o)$. Since $\Delta(s;0;\lambda_o)$ is Hurwitz, the determinants $|D_{oi}(\omega_o, \lambda_o)|$ and $|D_{oi}(\omega_o, \lambda_o)|'$, $i = 1, 2$, obtained from (37) are null only if

$$\begin{aligned} D_e^{c^2}(\omega_o^2;\lambda_o) + \omega_o^2 D_o^{c^2}(\omega_o^2;\lambda_o) &= 0,\\ D_e^{c'^2}(\omega_o^2;\lambda_o) + \omega_o^2 D_o^{c'^2}(\omega_o^2;\lambda_o) &= 0. \end{aligned} \tag{63}$$

As a consequence, the denominator of $C(s;\lambda)$ can be factorized as

$$D^c(s;\lambda_o) = (s^2 + \omega_o^2)^2 \hat{D}_c(s;\lambda_o), \tag{64}$$

where $\hat{D}_c(s;\lambda)$ is a suitable polynomial. This contradicts the hypothesis that $C(s;\lambda_o)$ does not have multiple imaginary poles.

ii) First, we prove by contradiction that the set $\Omega_c(\lambda)$ is empty for $\lambda \in \Lambda$. Let us assume that $\lambda_o \in \Lambda$ and that there exists a positive value $\omega_o \in \Omega_c(\lambda_o)$. Since $m \geq 1$, we obtain from (37) and condition C_1 in (25)

$$\left\{ \begin{aligned} &\Delta_e(\omega_o^2;\lambda_o) D_o^c(\omega_o^2;\lambda_o) - \Delta_o(\omega_o^2;\lambda_o) D_e^c(\omega_o^2;\lambda_o) = 0,\\ &\Delta_e(\omega_o^2;\lambda_o) D_e^c(\omega_o^2;\lambda_o) + \omega_o^2 \Delta_o(\omega_o^2;\lambda_o) D_o^c(\omega_o^2;\lambda_o) = 0, \end{aligned} \right. \tag{65}$$

$$\left\{ \begin{aligned} &\Delta_e(\omega_o^2;\lambda_o) N_o^c(\omega_o^2;\lambda_o) - \Delta_o(\omega_o^2;\lambda_o) N_e^c(\omega_o^2;\lambda_o) = 0,\\ &\Delta_e(\omega_o^2;\lambda_o) N_e^c(\omega_o^2;\lambda_o) + \omega_o^2 \Delta_o(\omega_o^2;\lambda_o) N_o^c(\omega_o^2;\lambda_o) = 0. \end{aligned} \right. \tag{66}$$

Since $\Delta(s;0;\lambda_o)$ is Hurwitz, $\nexists\, \omega_o : \Delta_e(\omega_o^2;\lambda_o) = \Delta_o(\omega_o^2;\lambda_o) = 0$. Hence, the two equation systems (65) and (66) admit solutions only if

$$\begin{aligned} D_e^{c^2}(\omega_o^2;\lambda_o) + \omega_o^2 D_o^{c^2}(\omega_o^2;\lambda_o) &= 0,\\ N_e^{c^2}(\omega_o^2;\lambda_o) + \omega_o^2 N_o^{c^2}(\omega_o^2;\lambda_o) &= 0, \end{aligned} \tag{67}$$

which means that $N^c(j\omega_o;\lambda_o) = D^c(j\omega_o;\lambda_o) = 0$. This contradicts the hypothesis that $\Delta(s;0;\lambda_o)$ is Hurwitz. Thus, $\Omega_c(\lambda)$ is empty $\forall\,\lambda \in \Lambda$. The theorem statement follows from Theorem 1.

REFERENCES

[1] B. R. Barmish, P. P. Khargonekar, Z. C. Shi, and R. Tempo, *Robustness margin need not be a continuous function of the problem data*, Systems Control Lett., 15 (1990), pp. 91–98.

[2] S. P. Bhattacharyya, *Robust Stabilization Against Structured Perturbations*, Lecture Notes in Control and Information Sciences 99, Springer-Verlag, New York, 1987.

[3] R. M. Biernacki, H. Hwang, and S. P. Bhattacharyya, *Robust stability with structured real parameter perturbations*, IEEE Trans. Automat. Control, AC-32 (1987), pp. 495–506.

[4] R. R. E. de Gaston and M. G. Safonov, *Exact calculation of the multiloop stability margin*, IEEE Trans. Automat. Control, AC-33 (1988), pp. 156–171.

[5] A. V. Fiacco, *Introduction to Sensitivity and Stability Analysis in Nonlinear Programming*, Academic Press, London, 1983.

[6] A. T. Fuller, *Conditions for a matrix to have only characteristic roots with negative real parts*, J. Math. Anal. Appl., 23 (1968), pp. 71–98.

[7] D. Hinrichsen and A. J. Pritchard, *New robustness results for linear systems under real perturbations*, in Proc. 27th IEEE Conference on Decision and Control, Austin, TX, 1988, pp. 1375–1379.

[8] ———, *A note on some differences between real and complex stability radii*, Systems Control Lett., 14 (1990), pp. 401–408.

[9] B. J. W. Milnor, *Topology from the Differentiable Viewpoint*, The University Press of Virginia, Charlottesville, 1965.

[10] M. P. Polis, W. Olbrot, and M. Fu, *An overview of recent results on the parametric approach to robust stability*, in Proc. 28th IEEE Conference on Decision and Control, Tampa, FL, 1989, pp. 23–29.

[11] L. Qiu and E. J. Davison, *A simple procedure for the exact stability robustness computation of polynomials with affine coefficient perturbations*, Systems Control Lett., 13 (1989), pp. 413–420.

[12] A. Sideris and R. S. Sánchez Peña, *Fast computation of the multivariable stability margin for real interrelated uncertain parameters*, in Proc. ACC, Atlanta, GA, 1988.

[13] D. D. Šiljak, *Nonlinear Systems: The Parameter Analysis and Design*, Wiley, New York, 1969.

[14] ———, *Parameter space methods for robust control design: A guided tour*, IEEE Trans. Automat. Control, AC-34 (1989), pp. 674–688.

[15] A. Tesi and A. Vicino, *Robustness analysis for uncertain dynamical systems with structured perturbations*, IEEE Trans. Automat. Control, 35 (1990), pp. 186–191.

[16] ———, *A new fast algorithm for robust stability analysis of linear control systems with linearly correlated parametric uncertainty*, Systems Control Lett., 13 (1989), pp. 321–329.

[17] A. Vicino, *Maximal polytopic stability domains in parameter space for uncertain systems*, Internat. J. Control, 49 (1989), pp. 351–361.

[18] A. Vicino and A. Tesi, *Regularity conditions for robust stablity problems with linearly structured perturbations*, in Proc. 29th IEEE Conference on Decision and Control, Honolulu, HI, 1990, pp. 46–51.

[19] A. Vicino, A. Tesi, and M. Milanese, *Computation of nonconservative stability perturbation bounds for systems with nonlinearly correlated uncertainties*, IEEE Trans. Automat. Control, AC-35 (1990), pp. 835–841.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 33, No. 4, pp. 1017–1027, July 1995

© 1995 Society for Industrial and Applied Mathematics

003

THE H_∞-PROBLEM FOR INFINITE-DIMENSIONAL SEMILINEAR SYSTEMS*

VIOREL BARBU†

Abstract. We study the H_∞-optimal control problem for semilinear systems in Hilbert spaces in connection with the corresponding H_∞-problem for the linearized system around an equilibrium state. The result of this paper, which expands some recent results of A. J. van der Schaft, is that if the linearized H_∞-problem has a suboptimal solution, then a suboptimal solution to the H_∞-problem for the semilinear system can be constructed in terms of a Hamilton–Jacobi equation.

Key words. Hamilton–Jacobi equation, Hamiltonian system, sup inf problem, exponentially stable semigroup

AMS subject classifications. 93B50, 93C35, 93C05

1. Introduction. Consider the input-output system

$$\begin{aligned} x' &= Ax + Fx + B_2u + B_1w, \quad t \geq 0; \quad x(0) = 0, \\ z &= C_1x + D_{12}u, \end{aligned} \tag{1.1}$$

in a real Hilbert space X. Here $x' = \frac{dx}{dt}$; A is the infinitesimal generator of a C_0-semigroup e^{At} on X; $F \in C^3(X)$; $F0 = 0$; $B_2 \in L(U, X)$; $B_1 \in L(W, X)$; $C_1 \in L(X, Z)$; $D_{12} \in L(U, Z)$; and U, W, and Z are real Hilbert spaces with the norms denoted $|\cdot|_U$, $|\cdot|_W$, and $|\cdot|_Z$, and scalar products $(\cdot,\cdot)_U$, $(\cdot,\cdot)_W$, and $(\cdot,\cdot)_Z$. The norm and the scalar product of X are denoted by $|\cdot|$ and $(\cdot,\cdot)$, respectively. In system (1.1) x is the state, u is the control input, w is an exogenous variable which includes disturbances, and z is the controlled input. We shall denote by $\nabla F : X \to L(X, X)$ the gradient of F and by $\nabla^2 F$ the second-order differential.

Consider the linearized system around 0

$$\begin{aligned} y' &= (A + \nabla F(0))y + B_2u + B_1w, \qquad t \geq 0, \\ y(0) &= 0. \end{aligned} \tag{1.2}$$

Given $\gamma > 0$ we say that $L \in L(X, U)$ is a suboptimal solution to the H_∞-problem associated with system (1.2) if the operator $A + \nabla F(0) + B_2L$ generates an exponentially stable semigroup and

$$\int_0^\infty |C_1y + D_{12}Ly|_Z^2\,dt < (\gamma^2 - \epsilon)\int_0^\infty |w(t)|_W^2\,dt, \quad \forall\, w \in L^2(R^+; W), \tag{1.3}$$

where y is the solution to (1.2) where $u = Ly$. Throughout the following we shall assume the following standard hypotheses on linearized system (1.2).

(i) The pair $(A + \nabla F(0), B_2)$ is exponentially stabilizable; i.e., there is $G \in L(X, U)$ such that $A + \nabla F(0) + B_2G$ generates an exponentially stable semigroup.

(ii) The pair $(A + \nabla F(0), C_1)$ is exponentially detectable, i.e., there is $K \in L(Z, X)$ such that $A + \nabla F(0) + KC_1$ generates an exponentially stable semigroup.

(iii) $D_{12}^*[C_1, D_{12}] = [0, I]$.

Assumption (iii) implies that

$$|C_1x + D_{12}u|_Z^2 = |C_1x|^2 + |u|^2, \quad \forall\,(x, u) \in X \times U.$$

* Received by the editors February 8, 1993; accepted for publication (in revised form) January 19, 1994.

† Department of Mathematics, University of Iaşi, 6600 Iaşi, România.

It is well known [2] – [4], [6] that under assumptions (i), (ii), and (iii) if the H_∞-problem for the linearized system (1.2) has a suboptimal solution, then the algebraic Riccati equation

$$(1.4)\quad (A+\nabla F(0))^*P + P(A+\nabla F(0)) - P(B_2B_2^* - \gamma^{-2}B_1B_1^*)P + C_1^*C_1 = 0$$

has a unique solution $P \in L(X,X)$, $P = P^* \geq 0$ such that $A + \nabla F(0) - (B_2B_2^* - \gamma^{-2}B_1B_1^*)P$ is exponentially stable.

The main result of this work, Theorem 1 below, relates the H_∞-control problem for the linearized system (1.2) to a certain Hamilton–Jacobi equation associated with the control system (1.1).

THEOREM 1. *Let $\gamma > 0$ and assume that the H_∞-problem for the system* (1.2) *has a suboptimal solution L. Then the Hamilton–Jacobi equation*

$$(1.5)\quad \begin{aligned} 2(Ax+Fx, \nabla\varphi(x)) + \gamma^{-2}|B_1^*\nabla\varphi(x)|_W^2 - |B_2^*\nabla\varphi(x)|_U^2 \\ + |C_1x|_Z^2 = 0, \quad \forall x \in D(A)\cap X_0 \end{aligned}$$

has a solution $\varphi \in C^2(X_0)$ in a neighborhood X_0 of the origin, satisfying

$$(1.6)\quad \varphi(0) = 0, \quad \nabla\varphi(0) = 0, \quad \nabla^2\varphi(0) = P.$$

Moreover, the solutions x to the system

$$(1.7)\quad \begin{aligned} x' = Ax + Fx - (B_2B_2^* - \gamma^{-2}B_1B_1^*)\nabla\varphi(x), \\ x(0) = x_0 \in X_0, \end{aligned}$$

which remain in X_0 have the property that $x \in L^2(R^+;X)$, $\lim_{t\to\infty}$; $x(t) = 0$, and the feedback control

$$(1.8)\quad u = -B_2^*\nabla\varphi(x)$$

asymptotically stabilizes system (1.1) *on X_0 and guarantees the closed-loop inequality*

$$(1.9)\quad \int_0^\infty (|C_1x(t)|_Z^2 + |u(t)|_U^2)dt < \gamma^2 \int_0^\infty |w(t)|_W^2\, dt,$$

$\forall w \in L^2(R^+;W)$, $w \not\equiv 0$ *and all solutions x to system* (1.1) *which remain in X_0 for all $t > 0$.*

The solution to (1.5) *is unique among the functions $\varphi \in C^2(X_0)$ satisfying* (1.6) *and having the property that every solution $x \in L^2(R^+;X)$ to the corresponding closed-loop system* (1.7) *with $x(0) \in X_0$ remains in X_0 for all $t \geq 0$.*

The approach we use here is quite different, and it relies on an existence result for the Hamiltonian system (2.1) corresponding with the given H_∞-problem (Proposition 1). As a matter of fact, as we shall see below, the solution φ to (1.5) can be characterized by the property that $\{(x,p) \in X_0 \times X;\ p + \nabla\varphi(x) = 0\}$ is a positively invariant manifold for this Hamiltonian system and that the corresponding closed-loop system is asymptotically stable. In [7], [8] the existence of such an invariant manifold follows via the stable manifold theorem. It is likely, however, that the methods of [7] can be extended to the present situation by using some recent results of center manifold theory in infinite dimensions. By quite different methods this problem has also been studied by Isidori and Astolfi [5] and in the author's work [1] where the given H_∞-problem is related to generalized solutions to the Hamilton–Jacobi equation (1.5).

Theorem 1 applies to semilinear distributed control systems, but we omit examples.

2. The Hamiltonian system. We shall study the existence of the Hamiltonian system

$$
\begin{aligned}
&x' = Ax + Fx + (B_2B_2^* - \gamma^{-2}B_1B_1^*)p,\\
&p' = -(A + \nabla F(x))^*p + C_1^*C_1x,\\
&x(0) = x_0, \qquad p(\infty) = 0,
\end{aligned}
\tag{2.1}
$$

in a neighborhood $K_\varepsilon = \{(x,p) \in X \times X; |x| < \varepsilon, |p| < \varepsilon\}$ of the origin.

PROPOSITION 1. *Under assumptions* (i)–(iii) *there are* $\varepsilon > 0$ *and* $0 < \delta < \varepsilon$ *such that for* $|x_0| < \delta$, *system* (2.1) *has a unique mild solution* $(x,p) \in L^2(R^+;X) \times L^2(R^+;X)$, $(x(t),p(t)) \in K_\varepsilon$, $\forall t \geq 0$.

By mild solution to (2.1) we mean a pair of continuous functions $(x,p) : R^+ \to X \times X$ which satisfy the equations

$$x(t) = e^{At}x_0 + \int_0^t e^{A(t-s)}(Fx(s) + (B_2B_2^* - \gamma^{-2}B_1B_1^*)p(s))ds, \quad \forall t \geq 0,$$

$$p(t) = U^*(T,t)p(T) - \int_t^T U^*(s,t)C_1^*C_1x(s)ds, \qquad 0 < t \leq T < \infty$$

for all $T > 0$. Here $\{U(s,t);\ 0 \leq t \leq s < \infty\}$ is the evolution operator associated with $A + \nabla F(x(t))$.

Proof of Proposition 1. We may rewrite system (2.1) as

$$
\begin{aligned}
&x' = (A + \nabla F(0))x + (B_2B_2^* - \gamma^{-2}B_1B_1^*)p + \mu(x),\\
&p' = -(A + \nabla F(0))^*p + C_1^*C_1x + \nu(x)p,\\
&x(0) + x_0, \qquad p(\infty) = 0,
\end{aligned}
\tag{2.2}
$$

where $\mu(x) = F(x) - \nabla F(0)x$ and $\nu(x) = (\nabla F(x))^* - (\nabla F(0))^*$. Clearly, we have

$$
\begin{aligned}
&|\mu(x)| \leq C_r|x|^2, \qquad \|\nu(x)\|_{L(X,X)} \leq C_r|x|, \quad \forall x \in \textstyle\sum_r\\
&\|\mu\|_{\mathrm{Lip}(\sum_r)} + \|\nu\|_{\mathrm{Lip}(\sum_r)} \leq C_r^1 r, \quad \forall r > 0
\end{aligned}
\tag{2.3}
$$

where $\sum_r = \{x \in X;\ |x| \leq r\}$ and $\|\cdot\|_{\mathrm{Lip}(\sum_r)}$ is the Lipschitz norm on $\sum_r$. In the space $Y = L^2(R^+;X) \times L^2(R^+;X)$ define the operator $\mathcal{A}$

(a) $D(\mathcal{A})$ is the set of all $(x,p) \in Y$ such that

$$
\begin{aligned}
&x' = (A + \nabla F(0))x + (B_2B_2^* - \gamma^{-2}B_1B_1^*)p + f,\ t \in R^+,\\
&p' = -(A + \nabla F(0))^*p + C_1^*C_1x + g,\ t \in R^+,\\
&x(0) = x_0, \qquad p(\infty) = 0
\end{aligned}
\tag{2.4}
$$

for some $(f,g) \in Y$.

The solution (x,p) to system (2.4) is considered in the mild sense, i.e.,

$$
\begin{aligned}
x(t) &= e^{(A+\nabla F(0))t}x_0 + \int_0^t e^{(A+\nabla F(0))(t-s)}((B_2B_2^* - \gamma^{-2}B_1B_1^*)p(s)\\
&\qquad + f(s))ds,\\
p(t) &= e^{(A+\nabla F(0))^*(T-t)}p(T) - \int_t^T e^{(A+\nabla F(0))^*(s-t)}(C_1^*C_1x(s) + g(s))ds,
\end{aligned}
\tag{2.5}
$$

for all $0 \le t \le T < \infty$.

(b) For $(x,p) \in D(\mathcal{A})$, $\mathcal{A}(x,p) = (f,g)$.

Obviously $\mathcal{A}$ is well defined and single valued. Formally we may define $\mathcal{A}$ as

$$\begin{aligned}\mathcal{A}(x,p) = \{&x' - (A + \nabla F(0))x - (B_2B_2^* - \gamma^{-2}B_1B_1^*)p,\\ &p' + (A+\nabla F(0))^*p - C^*C_1x\}, \quad \forall\,(x,p) \in D(\mathcal{A}),\end{aligned}$$

where

$$D(\mathcal{A}) = \{(x,p) \in Y;\ x(0) = x_0,\ p(\infty) = 0;\ \mathcal{A}(x,p) \in Y\}.$$

Let us denote by $\mathcal{A}_0$ the operator $\mathcal{A}$ in the case $x_0 = 0$. We have the following lemma.

LEMMA 1. *The operator $\mathcal{A}_0^{-1}$ is continuous from Y to $Y \cap (C(R^+;X) \times C(R^+;X))$.*

Proof. One must prove that for every $(f,g) \in Y$ the Hamiltonian system

$$\begin{aligned}x' &= (A + \nabla F(0))x + (B_2B_2^* - \gamma^{-2}B_1B_1^*)p + f, \qquad t \ge 0,\\ p' &= -(A+\nabla F(0))^*p + C_1^*C_1x + g, \qquad t \ge 0,\\ x(0) &= 0, \qquad p(\infty) = 0,\end{aligned} \tag{2.6}$$

has a unique solution $(x,p) \in Y \cap (C(R^+;X) \times C(R^+;X))$ and the map $(f,g) \to (x,p)$ is continuous. To this purpose consider the sup inf problem

$$\sup_{w \in \mathcal{W}} \inf_{u \in \mathcal{U}} \left\{ \int_0^\infty (|C_1x|_Z^2 + |u|_U^2 + 2(g,x) - \gamma^2|w|_W^2)dt;\right.\\ \left. x' = (A+\nabla F(0))x + B_2u + B_1w + f,\ x(0) = 0 \right\}, \tag{2.7}$$

where $\mathcal{U} = L^2(R^+;U)$ and $\mathcal{W} = L^2(R^+;W)$. We set

$$\psi(w) = \inf_{u\in\mathcal{U}} \left\{ \int_0^\infty (|C_1x|_Z^2 + |u|_U^2 + 2(g,x))dt;\right.\\ \left. x' = (A+\nabla F(0))x + B_2u + B_1w + f,\ x(0) = 0 \right\}. \tag{2.8}$$

Then by assumption (see (1.3)) the function $\gamma^2\|w\|_W^2 - \psi(w)$ is convex and coercive, and so the problem

$$\sup\{\psi(w) - \gamma^2\|w\|_W^2\} \tag{2.9}$$

has a unique solution w^*. Hence problem (2.7) has a unique solution $(u^*, w^*) \in \mathcal{U} \times \mathcal{W}$. For every $w \in \mathcal{W}$ we shall denote by $\bar{u} = \Gamma w$ the solution to problem (2.8), i.e.,

$$\bar{u} = \Gamma w = \arg\inf \left\{ \int_0^\infty (|C_1x|_Z^2 + |u|_U^2 + 2(g,x))dt;\right.\\ \left. x' = (A+\nabla F(0))x + B_2u + B_1w;\ x(0) = 0,\ u \in \mathcal{U} \right\}.$$

Then by standard arguments in linear quadratic control theory it follows that there is $\bar{p} \in L^2(R^+; X) \cap C(R^+; X)$ such that

$$\begin{aligned} \bar{p}' &= -(A + \nabla F(0))^* \bar{p} + C_1^* C_1 \bar{x} + g \quad \text{in}\, R^+, \\ \bar{p}(\infty) &= 0. \end{aligned} \tag{2.10}$$

$$\bar{u} = B_2^* \bar{p}, \quad \forall\, t \geq 0. \tag{2.11}$$

Moreover, by assumption (i) we may write (2.10) as

$$\begin{aligned} \bar{p}' &= -(A + \nabla F(0) + B_2 G)^* \bar{p} + G^* \bar{u} + C_1^* C_1 \bar{x} + g, \\ \bar{p}(\infty) &= 0. \end{aligned} \tag{2.12}$$

Since $A + \nabla F(0) + B_2 G$ is exponentially stable, we infer that the solution $\bar{p}$ to (2.10) is unique and

$$|\bar{p}(t)| \leq C \int_t^\infty e^{-\omega(s-t)} (|\bar{u}(s)|_U + |C_1 \bar{x}(s)|_Z + |g(s)|) ds, \, \forall\, t \geq 0, \tag{2.13}$$

where $\omega > 0$. Hence $\bar{p} \in L^2(R^+; X)$.

Note also that

$$\nabla \psi(w) = -2 B_1^* \bar{p}, \quad \text{a.e.}\, t > 0, \tag{2.14}$$

where $\bar{p}$ is the solution to (2.10). Indeed if x^w is the solution to $x' = (A + \nabla F(0))x + B_2 \Gamma w + B_1 w$, $x(0) = 0$, then we have for all $\bar{w} \in \mathcal{W}$

$$\begin{aligned} \psi(w) - \psi(\bar{w}) = \int_0^\infty (|C_1 x^w|_Z^2 - |C_1 x^{\bar{w}}|_Z^2) dt + \int_0^\infty (|\Gamma w|_U^2 - |\Gamma \bar{w}|_U^2) dt \\ + 2 \int_0^\infty (g, x^w - x^{\bar{w}}) dt \leq 2 \int_0^\infty ((C_1^* C_1 x^w, x^w - x^{\bar{w}}) \\ + (\Gamma w, \Gamma(w - \bar{w}))_U + (g, x^w - x^{\bar{w}})) dt. \end{aligned}$$

Then using system (2.10) we get

$$\begin{aligned} \psi(w) - \psi(\bar{w}) \leq -2 \int_0^\infty ((p, B_1(w - \bar{w})) + (B_2^* p, \Gamma(w - \bar{w}))_U \\ + (\Gamma w, \Gamma(w - \bar{w}))_U) dt = -2 \int_0^\infty (B_1^* p, w - \bar{w})_W \end{aligned}$$

as claimed. Since w^* is the solution to (2.9), we see by (2.10) and (2.14) that system (2.6) has a solution (x, p) where $x = x^*$ is just the solution to

$$(x^*)' = (A + \nabla F(0)) x^* + B_2 u^* + B_1 w^* + f; \; x^*(0) = 0$$

and $(u^*, w^*) \in \mathcal{U} \times \mathcal{W}$ is the solution to problem (2.7). Since $C_1 x^* \in L^2(R^+; Z)$, it follows by detectability assumption (ii) that $x^* \in L^2(R^+; X)$. Similarly by (2.13) we see that $p \in L^2(R^+; X)$ and $\lim_{t \to \infty} p(t) = p(\infty) = 0$.

Now if $(x, p) \in Y$ is any solution to (2.6), we have

$$\int_0^\infty (|C_1 x|_Z^2 + |B_2^* p|_U^2 + 2(g, x)) dt \leq \int_0^\infty (|C_1 y|_Z^2 + |v|_U^2 + 2(g, y)) dt, \tag{2.15}$$

for all (y, v) satisfying the system

$$y' = (A + \nabla F(0))y + B_2 v + B_1\tilde{w} + f, \qquad y(0) = 0, \tag{2.16}$$

where $\tilde{w} = -\gamma^{-2}B^*p$. On the other hand, after some calculation involving (2.6) we see that

$$\int_0^\infty (|B_2^*p|_U^2 + |C_1x|_Z^2 + (g,x) + (f,p))dt = \gamma^{-2}\int_0^\infty |B_1^*p|_W^2\, dt. \tag{2.17}$$

Then substituting into (2.15) we get

$$\gamma^{-2}\int_0^\infty |B_1^*p|_W^2\, dt \le \int_0^\infty (|C_1y|_Z^2 + |v|_U^2 + (g, 2y - x) + (f,p))dt.$$

In system (2.16) we take $v = Ly$, provided that (1.3) holds. This yields

$$\gamma^{-2}\int_0^\infty |B_1^*p|_W^2\, dt \le (\gamma^{-2} - \rho)\int_0^\infty |B_1^*p|_W^2\, dt$$
$$+C\int_0^\infty (|f|^2 + |g|^2)dt + \int_0^\infty (|f||p| + |g||x|)dt,$$

where C and $\rho > 0$ are independent of f and g. Hence

$$\int_0^\infty |B_1^*p|_W^2\, dt \le C\rho^{-1}\left(\int_0^\infty (|f|^2 + |g|^2)dt + \int_0^\infty (|f||p| + |g||x|)dt\right)$$

and by (2.17) we have

$$\begin{aligned}&\int_0^\infty (|B_2^*p|_U^2 + |C_1x|_Z^2 + |B_1^*p|_W^2)dt\\ &\quad\le C\int_0^\infty (|f|^2 + |g|^2 + |f||p| + |g||x|)dt\end{aligned} \tag{2.18}$$

for some positive constant C independent of $(f,g) \in Y$. Then by assumptions (i) and (ii) and system (2.6) it follows that

$$|x(t)|^2 + \int_0^\infty |x(t)|^2\, dt \le C\int_0^\infty (|C_1x|_Z^2 + |B_2^*p|_U^2 + |B_1^*p|_W^2 + |f|^2)dt,$$

$$|p(t)|^2 + \int_0^\infty |p(t)|^2\, dt \le C\int_0^\infty (|C_1x|_Z^2 + |B_2^*p|_U^2 + |g|^2)dt$$

and by (2.18) we see that

$$\begin{aligned}|x(t)|^2 + |p(t)|^2 + \int_0^\infty (|x(t)|^2 + |p(t)|^2)dt &\le C\int_0^\infty (|f|^2 + |g|^2)dt,\\ &\forall\,(f,g) \in Y, \qquad t \ge 0,\end{aligned} \tag{2.19}$$

as claimed.

Proof of Proposition 1 (*continued*). We note that for all $(f,g) \in Y$,

$$\mathcal{A}^{-1}(f,g) = \mathcal{A}_0^{-1}(f,g) + (\tilde{x}, \tilde{p}), \tag{2.20}$$

where $(\tilde{x},\tilde{p})$ is the solution to the system

$$\begin{aligned} &\tilde{x}' = (A+\nabla F(0))\tilde{x} + (B_2B_2^* - \gamma^{-2}B_1B_1^*)\tilde{p}, \qquad t \geq 0,\\ &\tilde{p}' = -(A+\nabla F(0))^*\tilde{p} + C_1^*C_1\tilde{x},\\ &\tilde{x}(0) = x_0, \qquad \tilde{p}(\infty) = 0. \end{aligned} \tag{2.21}$$

We have

$$\int_0^\infty (|C_1\tilde{x}|_Z^2 + |B_2^*\tilde{p}|_U^2)dt \leq \int_0^\infty (|C_1y|_Z^2 + |v|_U^2)dt$$

and

$$\int_0^\infty (|B_2^*\tilde{p}|_U^2 + |C_1\tilde{x}|_Z^2)dt = \gamma^{-2}\int_0^\infty (|B_1^*\tilde{p}|_W^2\,dt - (\tilde{p}(0),x_0))dt.$$

Here (y,v) is the solution to system (2.16) where $y(0) = x_0$ and $f \equiv 0$. Then arguing as in the proof of Lemma 1 we get the estimate

$$\int_0^\infty (|C_1\tilde{x}|_Z^2 + |B_2^*\tilde{p}|_U^2 + |B_1^*\tilde{p}|_W^2)dt \leq C|\tilde{p}(0)||x_0|, \quad \forall x_0 \in X.$$

Once again using assumptions (i) and (ii) in system (2.21) we see that

$$|\tilde{p}(t)|^2 + |\tilde{x}(t)|^2 + \int_0^\infty (|\tilde{x}|^2 + |\tilde{p}|^2)dt \leq C|\tilde{p}(0)||x_0|, \quad \forall t \geq 0$$

and therefore

$$|\tilde{p}(t)|^2 + |\tilde{x}(t)|^2 + \int_0^\infty (|\tilde{x}|^2 + |\tilde{p}|^2)dt \leq C|x_0|^2, \quad \forall x_0 \in X. \tag{2.22}$$

Now by virtue of (2.2) we may write system (2.1) as

$$(x,p) = \mathcal{A}_0^{-1}(\mu,\nu p) + (\tilde{x},\tilde{p}). \tag{2.23}$$

Consider the subset of Y,

$$\begin{aligned} Y_\varepsilon =\{&(x,p) \in Y; \|x\|_{L^2(R^+;X)}, \|p\|_{L^2(R^+;X)} \leq \varepsilon;\\ &\|x\|_{L^\infty(R^+;X)}, \|p\|_{L^\infty(R^+;X)} \leq \varepsilon\}. \end{aligned}$$

By estimate (2.22) we see that for $|x_0| \leq \delta(\varepsilon)$, $(\tilde{x},\tilde{p}) \in Y_{\varepsilon/2}$. Moreover, by Lemma 1 and estimates (2.3) we have

$$\mathcal{A}_0^{-1}(\mu,\nu p) \in Y_{\varepsilon/2}, \quad \forall\,(x,p) \in Y_\varepsilon$$

if ε is sufficiently small. Finally, by (2.3) and (2.19) we see that

$$\|\mathcal{A}_0^{-1}(\mu,\nu p) - \mathcal{A}_0^{-1}(\bar{\mu},\bar{\nu}\bar{p})\|_Y \leq C\varepsilon(\|x-\bar{x}\|_Y + \|p-\bar{p}\|_Y)$$

for all (x,p), $(\bar{x},\bar{p}) \in Y_\varepsilon$. Here $\bar{\mu} = \mu(\bar{x})$, $\bar{\nu} = \nu(\bar{x})$, and C is some positive constant. Then by the contraction principle, (2.23) (equivalently, system (2.1)) has a unique solution $(x,p) \in Y_\varepsilon \subset X_0$ for $|x_0| < \delta$ and $0 < \delta < \varepsilon$ sufficiently small. This completes the proof.

3. Proof of Theorem 1. Let $X_0 = \{x_0 \in X;\ |x_0| \le \delta\}$ and let $\Phi : X_0 \to X$ be the map

$$\Phi(x_0) = -p(0),$$

where p is the solution to system (2.1). As seen in Proposition 1, $\Phi(X_0) \subset \{x \in X;\ |x| \le \varepsilon\}$.

LEMMA 2. *$\Phi \in C^1(X_0)$ and $\Phi(x_0) = \nabla\varphi(x_0), \forall\, x_0 \in X_0$ where*

$$\varphi(x_0) = \frac{1}{2}\int_0^\infty (|C_1x(t)|_Z^2 + |B_2^*p(t)|_U^2 - \gamma^{-2}|B_1^*p(t)|_W^2)dt, \quad \forall\, x_0 \in X_0. \tag{3.1}$$

Proof. Let $h \in X$ be arbitrary but fixed. Consider the solution $(y, q) \in Y$ to system

$$\begin{aligned} y' &= (A + \nabla F(x))y + (B_2B_2^* - \gamma^{-2}B_1B_1^*)q, \\ q' &= -(A + \nabla F(x))^*q + C_1^*C_1y - (\nabla^2F(x)y)^*q, \\ y(0) &= h, \qquad q(\infty) = 0. \end{aligned} \tag{3.2}$$

We may write (3.2) as

$$\begin{aligned} y' &= (A + \nabla F(0))y + (B_2B_2^* - \gamma^{-2}B_1B_1^*)q + \mu_1(y), \\ q' &= -(A + \nabla F(0))^*q + C_1^*C_1y + \mu_2(y, q), \\ y(0) &= h, \qquad q(\infty) = 0, \end{aligned}$$

where $|\mu_1(y)| \le C\varepsilon|y|$, $|\mu_2(y, q)| \le C\varepsilon(|y| + |q|)$. Moreover, μ_1 and μ_2 are Lipschitzian with the Lipschitz constant $C\varepsilon$. Then arguing as in Lemma 1 we infer that system (3.2) has a unique solution $(y, q) \in Y \cap (C(R^+; X) \times C(R^+; X))$.

Now let (x, p) and (x_1, p_1) be two solutions to system (2.1), corresponding to x_0 and $x_0 + h$, respectively. We set $\bar{x} = x_1 - x - y$ and $\bar{p} = p_1 - p - q$ and notice that

$$\begin{aligned} \bar{x}' &= (A + \nabla F(0))\bar{x} + (B_2B_2^* - \gamma^{-2}B_1B_1^*)\bar{q} + (\nabla F(x) - \nabla F(0))\bar{x} + \nu_1(t), \\ \bar{p}' &= -(A + \nabla F(0))^*\bar{p} + C_1^*C_1\bar{x} - (\nabla F(x)\bar{x})^*p + (\nabla F(0) - \nabla F(x))^*p + \nu_2(t), \\ \bar{x}(0) &= 0, \qquad \bar{p}(\infty) = 0, \end{aligned}$$

where $|\nu_1(t)| + |\nu_2(t)| \le C|x_1(t) - x(t)|^2$, $\forall\, t \ge 0$. Recall that by Lemma 1,

$$\|x_1 - x\|_{L^\infty(R^+;X)\cap L^2(R^+;X)} \le C|h|.$$

Then using the invertibility of $\mathcal{A}_0$ we see that for ε sufficiently small

$$\|\bar{x}\|_{L^\infty(R^+;X)\cap L^2(R^+;X)} + \|\bar{p}\|_{L^\infty(R^+;X)\cap L^2(R^+;X)} \le C|h|^2. \tag{3.3}$$

The latter implies that

$$\|\Phi(x_0 + h) - \Phi(x_0) + q(0)\| \le C|h|^2.$$

Hence $\Phi \in C^1(X_0)$ and $\nabla\Phi(x_0)h = -q(0)$. In particular it follows that $\nabla\Phi(0) = P$ where $P \in L(X, X^*)$, $P = P^* \ge 0$ is defined by $Ph = -q(0)$, q being the solution to (3.2) where $x = 0$. Hence P is the unique solution to (1.4), having the property that $A + \nabla F(0) - (B_2B_2^* - \gamma^{-2}B_1B_1^*)P$ is exponentially stable.

Let us denote by φ the function defined by the right-hand side of (3.1). Taking into account (3.3) it is readily seen that φ is differentiable on X_0 and

$$(\nabla\varphi(x_0), h) = \int_0^\infty ((C_1^*C_1x, y) + (B_2B_2^*p, q) - \gamma^{-2}(B_1B_1^*p, q))dt,$$

where (y,q) is the solution to (3.2). Then by a little calculation involving (2.1) and (3.2) we see that

$$(\nabla\varphi(x_0),h) = -(p(0),h), \qquad |h| \le \delta.$$

Hence $\nabla\varphi(x_0) = \Phi(x_0)$, $\forall\, x_0 \in X_0$ as claimed. We will now prove that φ satisfies the Hamilton–Jacobi equation (1.4) on X_0. To this purpose we note that by uniqueness in system (2.1) we have

$$p(t) = -\Phi(x(t)), \quad \forall\, t \in [0,\mu), \tag{3.4}$$

where μ is such that $x(t) \in X_0$, $\forall\, t \in [0,\mu)$. Hence on the interval $[0,\mu)$, x is the solution to the closed-loop differential system

$$\begin{gathered} x'(t) = Ax(t) + Fx(t) - (B_2B_2^* - \gamma^{-2}B_1B_1^*)\Phi(x(t)), \\ x(0) = x_0. \end{gathered} \tag{3.5}$$

Let $x_0 \in D(A) \cap X_0$. Since F and Φ are smooth, we conclude that x is continuously differentiable on $[0,\mu)$ and so is p. Then multiplying (3.5) by $\Phi(x(t))$ we get

$$\begin{aligned} \frac{d}{dt}\varphi(x(t)) &= (Ax(t) + Fx(t), \Phi(x(t))) \\ &\quad - |B_2^*\Phi(x(t))|_U^2 + \gamma^{-2}|B_1^*\Phi(x(t))|_W^2, \ \forall\, t \in [0,\mu). \end{aligned} \tag{3.6}$$

On the other hand, we have

$$\varphi(x(t)) = \frac{1}{2}\int_t^\infty (|C_1x|_Z^2 + |B_2^*p|_U^2 - \gamma^{-2}|B_1^*p|_W^2)dt, \quad \forall\, t \ge 0.$$

Along with (3.6) the latter yields

$$\begin{gathered} |C_1x_0|_Z^2 + |B_2^*\Phi(x_0)|_U^2 - \gamma^{-2}|B_1^*\Phi(x_0)|_W^2 + 2(Ax_0 + Fx_0, \Phi(x_0)) = 0, \\ \forall\, x_0 \in X_0 \cap D(A), \end{gathered}$$

as claimed.

Consider now the closed-loop system

$$\begin{gathered} x' = Ax + Fx - B_2B_2^*\nabla\varphi(x) + B_1w, \\ x(0) = x_0, \end{gathered} \tag{3.7}$$

where $x_0 \in X_0$ and $w \in L^2(R^+;W)$. We shall assume that $x_0 \in D(A)$ and $w \in C^1(R^+;W)$. Then problem (3.7) has a unique local smooth solution $x = x(t)$. We restrict ourselves to w, x_0 for which $x(t) \in X_0$ for all $t \ge 0$. (This happens, for instance, if $|x_0|$ and $\|w\|_{L^2(R^+;W)}$ are sufficiently small.) For such a solution we have

$$\begin{aligned} \frac{d}{dt}\varphi(x(t)) &= (Ax(t) + Fx(t), \nabla\varphi(x(t)) - |B_2^*\nabla\varphi(x(t))|_U^2 \\ &\quad + (w(t), B_1^*\nabla\varphi(x(t)))_W \\ &= -2^{-1}(|B_2^*\nabla\varphi(x(t))|_U^2 + \gamma^{-2}|B_1^*\nabla\varphi(x(t))|_W^2 \\ &\quad + |C_1x(t)|_Z^2) + (w(t), B_1^*\nabla\varphi(x(t)))_W, \quad \forall\, t \ge 0. \end{aligned} \tag{3.8}$$

Since by virtue of (1.6) φ is positive definite in a neighborhood of the origin (let it be X_0) the latter yields

$$\begin{aligned}\int_0^\infty (|C_1x(t)|_Z^2 + |B_2^*\nabla\varphi(x(t))|_U^2 - \gamma^{-2}|w(t)|_W^2)dt \\ = \varphi(x_0) - \gamma^{-2}\int_0^\infty |B_1^*\nabla\varphi(x(t)) - \gamma^2 w(t)|_W^2\, dt.\end{aligned} \tag{3.9}$$

This equality clearly extends to all $w \in L^2(R^+; W)$ and $x_0 \in X_0$ such that $x(t)$ remain in X_0 for all $t \geq 0$. Let $K \in L(Z, X)$ be such that $A + \nabla F(0) + KC_1$ is exponentially stable. Then writing system (3.7) with $w = 0$ as

$$x' = (A + \nabla F(0) + KC_1)x - KC_1x - B_2B_2^*\nabla\varphi(x) + F(x) - \nabla F(0)x,$$

it follows by (3.9) that

$$|x(t)| \leq C(e^{-\omega t}|x_0| + \varepsilon\int_0^t e^{-\omega(t-s)}|x(s)|ds + f_0(t)), \qquad t \geq 0,$$

where $\omega > 0$ and $f_0 \in L^2(R^+)$. Then for ε sufficiently small the previous inequality implies that $x \in L^2(R^+; X)$ and $\lim_{t\to\infty}|x(t)| = 0$. We have proved, therefore, that the feedback control (1.8) asymptotically stabilizes system (1.1) in X_0. On the other hand, inequality (3.9) for $x_0 = 0$ implies (1.9) as claimed.

Now let $x = x(t)$ be a solution to system (1.7) such that $x(t) \in X_0$, $\forall t \geq 0$. We may write (1.7) as

$$\begin{gathered}x' = (A + \nabla F(0) - B_2B_2^*P + \gamma^{-2}B_1B_1^*P)x + \mu(x),\\ x(0) = x_0,\end{gathered}$$

where $|\mu(x)| \leq C|x|^2 \leq \delta|x|$. Recalling that $A + \nabla F(0) - B_2B_2^*P + \gamma^{-2}B_1B_1^*P$ is exponentially asymptotically stable [6], we infer that if X_0 is sufficiently small, then $x \in L^2(R^+; X)$ as claimed.

The uniqueness of solution φ to (1.5) satisfying (1.6) and having the property that $A + F - (B_2B_2^* - \gamma^{-2}B_1B_1^*)\nabla\varphi$ is asymptotically stable in X_0 remains to be proven. If $\varphi \in C^2(X_0)$ is such a function, consider the solution x to system (1.7) where $x_0 \in X_0 \cap D(A)$. Then $x(t) \in X_0 \cap D(A)$, $\forall t \geq 0$. We set $p(t) = -\nabla\varphi(x(t))$ and notice that by virtue of (1.5), (x, p) is the solution to the Hamiltonian system (2.1). Indeed we have

$$p'(t) = -\nabla^2\varphi(x(t))(Ax(t) + Fx(t) - (B_2B_2^* - \gamma^{-2}B_1B_1^*)\nabla\varphi(x(t))) \quad \forall t \geq 0$$

while by (1.5),

$$\begin{aligned}&\nabla^2\varphi(x)(Ax + Fx - (B_2B_2^* - \gamma^{-2}B_1B_1^*)\nabla\varphi(x))\\ &\quad = -(A + \nabla F(x))^*\nabla\varphi(x) - C_1^*C_1x, \quad \forall x \in D(A) \cap X_0.\end{aligned}$$

Hence

$$p'(t) = -(A + \nabla F(x(t)))^*p(t) + C_1^*C_1x(t), \quad \forall t \geq 0.$$

By the uniqueness in system (2.1) (Proposition 1) we infer that $\nabla\varphi$ is uniquely defined on X_0, thereby completing the proof.

Remark. By the previous proof it follows that in a neighborhood of the origin $\{(x, p); p + \nabla\varphi(x) = 0\}$ is a positively invariant manifold of the Hamiltonian system (2.1) and the closed-loop system (1.7) is asymptotically stable. Moreover, the function φ is unique with these properties. In fact, we may reformulate Theorem 1 in these terms.

REFERENCES

[1] V. BARBU, *H_∞-control for semilinear systems in Hilbert spaces*, Systems Control Lett., 21 (1993), pp. 65–72.

[2] A. BENSOUSSAN AND P. BERNHARD, *Remarks on the theory of robust control*, Optimization, Optimal Control, and Partial Differential Equations, V. Barbu, F. Bonans, and D. Tiba, eds., Birkhäuser Verlag, Basel, 1992, pp. 149–166.

[3] J. DOYLE, K. GLOVER, P. KHARGONEKAR, AND B. FRANCIS, *State space solution to standard H_2 and H_∞-control problems*, IEEE Trans. Automat. Control, AC 34 (1989), pp. 831–847.

[4] A. ICHIKAWA, *H_∞-control and min-max problems in Hilbert space*, to appear.

[5] A. ISIDORI, A. ASTOLFI, *Nonlinear H_∞-control via measurement feedback*, J. Math. Systems Estim. Control, 2 (1992), pp. 31–44.

[6] B. VAN KEULEN, M. PETERS, AND R. CURTAIN, *H_∞-control with state feedback: The infinite dimensional case*, J. Math. Systems Estim. Control, 3 (1993), pp. 1–39.

[7] A. J. VAN DER SCHAFT, *A state space approach to nonlinear H_∞-control*, Systems Control Lett. 16 (1991), pp. 1–8.

[8] ———, *L_2-gain analysis of nonlinear systems and nonlinear state feedback H_∞-control*, IEEE Trans. Automat. Control, 37 (1992), pp. 770–784.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 33, No. 4, pp. 1028–1039, July 1995

© 1995 Society for Industrial and Applied Mathematics

004

RELAXED MINIMAX CONTROL*

E. N. BARRON[†] AND R. JENSEN[‡]

Abstract. The relaxation of the optimal control problem with cost functional which is the supremum in time of some function $h(t, x, z)$ is determined. The trajectory is convexified in the usual way but the cost functional is convexified in a nonobvious manner. Thus, if the original value function is $V(t,x) = \inf_{\zeta \in \mathcal{Z}} \|h(s, \xi(s), \zeta(s))\|_{L^\infty[t,T]}$, then the relaxed value function is

$$\hat{V}(t,x) = \inf_{\mu \in \hat{\mathcal{Z}}[t,T]} \| \|h(s, \hat{\xi}(s), z)\|_{L^\infty(Z;\mu(s))} \|_{L^\infty[t,T]}$$

where the inner norm is the essential sup of h over $z \in Z$ with respect to the measure $\mu(s)$. We prove that V and $\hat{V}$ coincide.

Key words. L^∞ optimal control, relaxed problem

AMS subject classifications. 49A10, 49A40, 93C60

1. Introduction.

Nonconvex optimal control problems with integral or terminal cost functionals, the problems of Mayer, Bolza, and Lagrange, are generally not guaranteed to possess optimal controls. The usual method of getting around this problem is to enlarge the class of controls to include probability measures, the class of relaxed controls. Convexifying the trajectory dynamics and the cost functional leads in classical problems to a weakly lower semicontinuous and convex problem for which an optimal relaxed control is guaranteed to exist. In general the optimal relaxed control can be approximated by ordinary controls. The books by Warga [11] and Cesari [6] are good references (see also [9], [10]).

Considering the importance of relaxation, the question arises as to what is the relaxed control problem when the cost functional is not the standard type mentioned above but is instead a problem of minimax control. The minimax, or L^∞ problem, consists of minimizing, using controls, the supremum in time of some given function h of time, the trajectory, and the control. The value function is given by $V(t,x) = \inf_{\zeta \in \mathcal{Z}} \|h(s, \xi(s), \zeta(s))\|_{L^\infty[t,T]}$, where $\xi(\cdot)$ is the trajectory given as the solution of $d\xi/d\tau = f(\tau, \xi(\tau), \zeta(\tau))$ and $\zeta(\cdot)$ is the control. Now, there is more than one method for convexifying this problem. One way, which is probably the most obvious, is to simply convexify $f(\cdot,\cdot,z)$ and $h(\cdot,\cdot,z)$ by replacing them by $\int_Z f(t,x,z)\mu(dz)$ and $\int_Z h(t,x,z)\mu(dz)$, respectively, where μ is a probability measure on the control set Z. We present an example below where it is clear that this is *not* the appropriate relaxation.

The question is what properties does one want the relaxed problem to have. The answer, by analogy with classical problems, is that the relaxed problem should (i) possess an optimal control and (ii) the value function of the relaxed problem should coincide with the value function of the original problem for reasonably behaved functions h. In the example below, the proposed relaxation on a first attempt is incorrect because (ii) is not satisfied, even though the original problem has convex f and linear h in the z variable.

The way we approach determining the correct relaxation is to look at L^p approximations to the L^∞ norm. The L^p problem is a problem of Lagrange and can be relaxed in the usual way. When we take the limit as $p \to \infty$, we see that the L^∞ relaxation of h should not be

* Received by the editors June 16, 1993; accepted for publication (in revised form) January 28, 1994.

[†] Department of Mathematical Sciences, Loyola University of Chicago, 6525 North Sheridan Road, Chicago, Illinois 60626 (enb@math.luc.edu). The research of this author was supported by National Science Foundation grants DMS-9102967 and DMS-9300805.

[‡] Department of Mathematical Sciences, Loyola University of Chicago, 6525 North Sheridan Road, Chicago, Illinois 60626 (rrj@math.luc.edu). The research of this author was supported by National Science Foundation grant DMS-9101799.

$\int_Z h(t,x,z)\mu(dz)$ but instead $\|h(t,x,z)\|_{L^\infty(Z;\mu)}$, i.e., the essential supremum of $h(\cdot,\cdot,z)$ with respect to the probability measure μ. However, we see that $\mu \mapsto \|h(t,x,z)\|_{L^\infty(Z;\mu)}$ is *not* convex but is rather *quasiconvex*, i.e., the level sets are all convex. Further, this map is lower semicontinuous with respect to weak convergence.

We prove that this relaxation will guarantee the existence of an optimal relaxed control. Therefore (i) above will hold. We then prove that the relaxed value function and the original value function coincide by establishing that the relaxed value function is a viscosity solution of the same Hamilton–Jacobi–Bellman equation as the original value function. Uniqueness of viscosity solutions yields the fact that they must be identical. Thus, both desirable properties of a relaxed control problem will hold.

2. Statement of the problem. Consider the controlled system of ordinary differential equations

$$d\xi(\tau)/d\tau = f(\tau,\xi(\tau),\zeta(\tau)), \qquad 0 \le t < \tau \le T, \tag{2.1}$$

$$\xi(t) = x \in R^n. \tag{2.2}$$

The control functions $\zeta(\cdot)$ are chosen from the class of functions

$$\mathcal{Z}[t,T] = \{\zeta : [t,T] \to Z | \zeta \text{ is Lebesgue measurable}\},$$

where Z, the control set, is a fixed compact subset of some euclidean space. We assume for simplicity that

(A) $f : [0,T] \times R^n \times Z \to R^n$ is jointly continuous and is Lipschitz in x. That is, there is a generic constant K such that

$$|f(t,x,z) - f(t,x',z)| \le K|x-x'|, \quad \forall\, x,x' \in R^n. \tag{2.3}$$

In addition, $|f(t,x,z)| \le K(1+|x|)$.

We are also given a function $h : [0,T] \times R^n \times Z \to R^1$ assumed to be jointly continuous and also uniformly Lipschitz in x. Without loss of generality we assume that h is a nonnegative function since otherwise we relace h by $|h|$.

Under (A), (2.1)–(2.2) has a unique solution for any $h \in \mathcal{Z}\,[t,T]$ (see [1]). The goal is to choose a control function $\zeta \in \mathcal{Z}$ that minimizes the largest that $h(r,\xi(r),\zeta(r))$ can be on the time interval $[t,T]$. We approach this problem by studying the value function $V : [0,T] \times R^n \to R^1$:

$$V(t,x) = \inf_{\zeta\in\mathcal{Z}} \operatorname*{ess\,sup}_{t\le r\le T} h(r,\xi(r),\zeta(r)). \tag{2.4}$$

It was proved in [4] (see also [2] for a direct proof) that the ess $\sup_{t\le r\le T}$ in (2.4) can be replaced by $\sup_{t\le r\le T}$. Intuitively, jumps of ζ on time intervals of Lebesgue measure zero are not necessary since they cannot change the dynamics.

The main result of [4] is that the function V is the unique continuous viscosity solution of the problem

$$\max\left\{V_t + \min_{\{z\in Z|h(t,x,z)\le V(t,x)\}} D_xV\cdot f(t,x,z),\ \min_{z\in Z} h(t,x,z) - V(t,x)\right\} = 0, \tag{2.5}$$

$$V(T,x) = \min_{z\in Z} h(T,x,z). \tag{2.6}$$

Remark 2.1. By examining the proofs of [4] one can weaken the continuity assumption of $h(t, x, z)$ in $z \in Z$ to lower semicontinuity uniformly in (t, x). This is important because we see that the corresponding relaxation involves a lower semicontinuous rather than a continuous function.

Define the hamiltonian function

$$H(t,x,r,p) = \min_{\{z \in Z | h(t,x,z) \leq r\}} p \cdot f(t,x,z) \tag{2.7}$$

for $r \in R^1$ and $p \in R^n$.

Remark 2.2. Observe that the minimum is over the set $\{z \in Z | h(t,x,z) \leq r\}$. If this set is empty, H is defined as $+\infty$. The form of (2.5) implies that the set is nonempty. Furthermore, the minimum in the hamiltonian is guaranteed since Z is compact and (A) holds.

We have that (see [4])

$$H^*(t,x,r,p) = H(t,x,r-0,p)$$

and

$$H_*(t,x,r,p) = H(t,x,r+0,p)$$

are the upper and lower semicontinuous envelopes of H, respectively. The function V then satisfies Eq. (2.5) as in the following definition [7], [8].

DEFINITION 2.1. *V is a viscosity solution of* (2.5) *if*

(i) *V is a viscosity subsolution, i.e., for any (t_0, x_0) for which $V - \varphi$ has a maximum, for a smooth function φ, it follows that*

$$\max\{\varphi_t + H^*(t_0, x_0, V(t_0, x_0), D_x\varphi(t_0, x_0)), \min_{z \in Z} h(t_0, x_0, z) - V(t_0, x_0)\} \geq 0$$

and

(ii) *V is a viscosity supersolution, i.e., for any (t_0, x_0) for which $V - \varphi$ has a minimum, for a smooth function φ, it follows that*

$$\max\{\varphi_t + H_*(t_0, x_0, V(t_0, x_0), D_x\varphi(t_0, x_0)), \min_{z \in Z} h(t_0, x_0, z) - V(t_0, x_0)\} \leq 0.$$

[4] does not consider the important problem of the existence of optimal controls. In the classical theory of optimal control one requires some sort of convexity of f and h in the controls in order to guarantee existence. Without convexity assumptions there may not exist an optimal control. In that event, one convexifies the problem by convexifying f and h. This is accomplished through the use of relaxed controls. Thus, we seek the corresponding relaxed control problem for minimax optimal control.

To motivate what follows we first take the naive approach that we convexify f and h in exactly the same way as in the problem of Lagrange. We introduce the space of relaxed controls.

Let $M(Z)$ denote the space of bounded measures on Z. Viewing $M(Z)$ as the dual space of $C(Z)$ = continuous functions on Z, we endow $M(Z)$ with the weak star topology of $C(Z)^*$. Let the space of relaxed controls be given by

$$\widehat{\mathcal{Z}}[t,T] = \{\mu \in L^\infty([t,T]; M(Z)) | \mu(\tau) \text{ is a probability measure a.e. } \tau \in [t,T]\}.$$

Let $\mathcal{M}(Z)$ be the set of probability measures on Z. Then, we may write that $\widehat{\mathcal{Z}}[t,T] = L^\infty([t,T]; \mathcal{M}(Z))$, the space of essentially bounded, Lebesgue measurable maps $\mu : [t,T] \to \mathcal{M}(Z)$. For any relaxed control $\mu \in \widehat{\mathcal{Z}}[t,T]$ there is a unique relaxed trajectory given by

$$\hat{\xi}(\tau) = x + \int_t^\tau \int_Z f(s, \hat{\xi}(s), z)\mu(s, dz)ds. \tag{2.8}$$

For any $\mu \in \mathcal{M}(Z)$ define the functions

$$\hat{f}(t,x,\mu) = \int_Z f(t,x,z)\mu(dz) \tag{2.9}$$

and

$$\hat{h}(t,x,\mu) = \mu - \operatorname*{ess\,sup}_{z\in Z} h(t,x,z) \equiv \|h(t,x,z)\|_{L^\infty(Z;\mu)}, \tag{2.10}$$

where $L^\infty(Z;\mu)$ is the space of essentially bounded, with respect to the measure $\mu \in \mathcal{M}(Z)$, real valued functions on Z. We show that the correct relaxation of the L^∞ problem involves the functions $\hat{f}$ and $\hat{h}$. The relaxed dynamic is the usual $\hat{f}$, but the classical relaxation of the cost functional uses $\int_Z h(t,x,z)\mu(dz)$ rather than $\hat{h}$. Now we show that replacing h by $\int_Z h(t,x,z)\,d\mu$ is *not* the correct relaxation.

Define the relaxed value function $\widehat{W} : [0,T] \times R^n \to R^1$ as

$$\widehat{W}(t,x) = \inf_{\mu\in\widehat{\mathcal{Z}}[t,T]} \operatorname*{ess\,sup}_{t\le r\le T} \int_Z h(r,\hat{\xi}(r),z)\mu(r,dz). \tag{2.11}$$

The function $\widehat{W}$ is easily shown to be the unique viscosity solution of

$$\begin{aligned}\max\Bigg\{\widehat{W}_t + \min_{\{\mu\in\mathcal{M}(Z)\mid \int_Z h(t,x,z)\mu(dz)\le\widehat{W}(t,x)\}} D_x\widehat{W}\cdot\int_Z f(t,x,z)\mu(dz),\\ \min_{\mu\in\mathcal{M}(Z)}\int_Z h(t,x,z)\mu(dz) - \widehat{W}(t,x)\Bigg\} = 0,\end{aligned} \tag{2.12}$$

$$\widehat{W}(T,x) = \min_{\mu\in\mathcal{M}(Z)} \int_Z h(T,x,z)\mu(dz). \tag{2.13}$$

Since it is possible that

$$\min_{\{\mu\in\mathcal{M}(Z)\mid \int_Z h(t,x,z)\mu(dz)\le r\}} p\cdot\int_Z f(t,x,z)\mu(dz) < \min_{\{z\in Z\mid h(t,x,z)\le r\}} p\cdot f(t,x,z)$$

it will not be the case, in general, that $\widehat{W} = V$. Consider the following.

Example. On the time interval $[-1,0]$, take $f(t,x,z) = z^2 - 10$, $Z = [0,10]$, and $h(t,x,z) = |x| + z$. First, we calculate an upper bound for the relaxed value for the initial conditions $x = 0$, $t = -1$. Choose the constant (in time) relaxed control

$$\mu(\tau) = 0.9\,\delta(z-0) + 0.1\,\delta(z-10).$$

We are using the notation that for any $z_0 \in Z$, $\delta(z-z_0)$ is the measure, all of whose mass is concentrated at z_0. The relaxed trajectory is given by

$$\hat{\xi}(\tau) = \int_{-1}^{\tau}\int_{[0,10]} z^2 - 10\,\mu(s,dz)ds = 0$$

for all $-1 \le \tau \le 0$ since $\int_{[0,10]} z^2\,\mu(s,dz) = 10$. Since $\int_{[0,10]} z\,\mu(s,dz) = 1$, the relaxed value function $\widehat{W}$ then satisfies

$$\widehat{W}(-1,0) \le \left\|\int_{[0,10]} z\,\mu(s,dz)\right\|_{L^\infty[-1,0]} = 1.$$

On the other hand, if $V(-1,0) \leq 1$, then it is clear that we may consider only controls for which $0 \leq \zeta(\tau) \leq 1$ for all τ. Clearly the optimal control must then be $\zeta(\tau) \equiv 1$ since that is the control that keeps ξ as close to zero as possible. But then $\xi(\tau) = -9(\tau + 1)$ and

$$V(-1,0) = \sup_{-1\leq\tau\leq 0} (|\xi(\tau)| + \zeta(\tau)) = 10 > 1.$$

We conclude that the relaxed value $\widehat{W}(-1,0) < V(-1,0)$ even though the function f is quadratic and h is linear in z in this example. Therefore, this relaxation of the L^∞ problem cannot be the correct version.

To motivate the relaxation using the function $\hat{h}$ in (2.10), we look at $L^p, p > 1$, approximations to the L^∞ problem as in [4]. Define the L^p value function

$$V_p(t,x) = \inf_{\zeta\in\mathcal{Z}[t,T]} \left(\int_t^T h^p(s,\xi(s),\zeta(s))ds\right)^{\frac{1}{p}}.$$

Then V_p is the unique continuous viscosity solution of

$$(V_p)_t + \min_{z\in Z} \left(D_x V_p \cdot f(t,x,z) + \frac{1}{p}\left(\frac{h(t,x,z)}{V_p(t,x)}\right)^p V_p(t,x)\right) = 0.$$

If we relax the L^p problem using classical theory, we obtain the value function

$$\widehat{V}_p(t,x) = \inf_{\mu\in\widehat{\mathcal{Z}}} \left(\int_t^T \int_Z h(s,\hat{\xi}(s),z)^p \, \mu(s,dz)ds\right)^{\frac{1}{p}}. \tag{2.14}$$

Remark 2.3. Observe that the function $\widehat{W}$ in (2.11) is obtained formally as a limit as $p \to \infty$ of

$$\widehat{W}_p(t,x) = \inf_{\mu\in\widehat{\mathcal{Z}}[t,T]} \left(\int_t^T \left(\int_Z h(s,\hat{\xi}(s),z)\,\mu(s,dz)\right)^p ds\right)^{\frac{1}{p}}. \tag{2.15}$$

This corresponds to relaxing h, not h^p. But the integrals on the right side of (2.15) are *not* weakly lower semicontinuous in μ for all $p > 1$ because this is a nonlinear nonconvex for odd p, functional of μ. Therefore, we are guaranteed the existence of an optimal relaxed control for the L^p problem when we relax h^p, but this is *not* the case when we relax h.

The function $\widehat{V}_p$ satisfies for each p, the problem

$$\widehat{V} + \min_{\mu\in\mathcal{M}} \left(D_x\widehat{V} \cdot \hat{f}(t,x,\mu) + \frac{1}{p}\left(\frac{\int_Z h(t,x,z)^p\,\mu(dz)}{\widehat{V}(t,x)^p}\right)\widehat{V}(t,x)\right) = 0.$$

If a limit is going to exist as $p \to \infty$, it is clear that we must have

$$(\widehat{V}_p(t,x))^p \geq \int_Z h(t,x,z)^p\,\mu(dz)$$

or, in other words, for large p,

$$\widehat{V}_p(t,x) \geq \left(\int_Z h(t,x,z)^p \mu(dz)\right)^{\frac{1}{p}}.$$

If $\widehat{V}_p$ converges to, say, $\widehat{V}$, then this puts the requirement on the measures μ that

$$\widehat{V}(t,x) \geq \|h(t,x,z)\|_{L^\infty(Z;\mu)} = \hat{h}(t,x,\mu)$$

since $(\int_Z h^p(t,x,z)\,\mu(dz))^{1/p} \to \|h(t,x,z)\|_{L^\infty(Z;\mu)}$ as $p \to \infty$. Thus, we arrive at the following statement, whose proof is similar to the proof of Proposition 2.6 and Theorem 4.2 in [4] and is therefore omitted.

THEOREM 2.1. $\lim_{p\to\infty} \widehat{V}_p(t,x) = \widehat{V}(t,x)$ *exists on* $[0,T] \times R^n$ *and* $\widehat{V}$ *is the unique, continuous viscosity solution of*

(2.16)

$$\max\left\{ \widehat{V}_t + \min_{\{\mu\in\mathcal{M}(Z)|\hat{h}(t,x,\mu)\leq\widehat{V}\}} (D_x\widehat{V}\cdot\hat{f}(t,x,\mu)),\ \min_{\mu\in\mathcal{M}(Z)} \hat{h}(t,x,\mu) - \widehat{V}(t,x) \right\} = 0$$

and terminal condition

$$\widehat{V}(T,x) = \min_{\mu\in M(Z)} \hat{h}(T,x,\mu). \tag{2.17}$$

Furthermore,

$$\widehat{V}(t,x) = \inf_{\mu(\cdot)\in\widehat{\mathcal{Z}}[t,T]} \|\hat{h}(s,\hat{\xi}(s),\mu(s))\|_{L^\infty[t,T]}.$$

With this theorem as motivation, we make the following definition.

DEFINITION 2.2. *The relaxed value function associated with the* L^∞ *problem is*

$$\widehat{V}(t,x) = \inf_{\mu(\cdot)\in\widehat{\mathcal{Z}}[t,T]} \|\hat{h}(s,\hat{\xi}(s),\mu(s))\|_{L^\infty[t,T]}. \tag{2.18}$$

It is also possible to prove directly that $\widehat{V}$ defined by (2.18) is the viscosity solution of (2.16) and (2.17) by using the dynamic programming principle

$$\widehat{V}(t,x) = \inf_{\mu\in\widehat{\mathcal{Z}}[t,s]} \max\{\|\hat{h}(r,\hat{\xi}(r),\mu(r))\|_{L^\infty[t,s]},\ \widehat{V}(s,\hat{\xi}(s))\}. \tag{2.19}$$

The proof is similar to that of Theorem 3.2 in [4]. However, one now uses the weakened assumption that $\hat{h}(\cdot,\cdot,\mu)$ is lower semicontinuous (with respect to weak convergence), which is proved in Lemma 3.2 below. Uniqueness of viscosity solutions is proved as in Theorem 4.2 in [4].

3. The relaxed L^∞ value: Existence of an optimal control. Having motivated the definition of the relaxed value function for the L^∞ optimal control problem, we must now establish that it is the appropriate relaxation in the sense that (i) an optimal relaxed control will exist and (ii) the relaxed value is the same as the original value function. To begin, we need some preliminary results.

LEMMA 3.1. *For each* (t,x) *fixed, the mapping*

$$\mu \in \mathcal{M}(Z) \mapsto \hat{h}(t,x,\mu)$$

is a quasiconvex function. In fact,

$$\hat{h}(t,x,\lambda\mu_1 + (1-\lambda)\mu_2) = \max\{\hat{h}(t,x,\mu_1),\hat{h}(t,x,\mu_2)\} \tag{3.1}$$

for $0 < \lambda < 1$.

Remark 3.1. A quasiconvex function is a function, say $g : X \to R^1$, X a convex set, satisfying

$$g(\lambda x + (1-\lambda)y) \leq \max\{g(x), g(y)\}, \quad 0 \leq \lambda \leq 1, \quad x, y \in X.$$

Equivalently, g is quasiconvex if the level set of g, $\{x \in X | g(x) \leq r\}$ is convex for all $r \in R^1$. This definition is not to be confused with quasiconvexity for functionals defined in the calculus of variations.

Proof. Fix $0 < \lambda < 1$, $\mu_1, \mu_2 \in \mathcal{M}(Z)$ and set $\nu = \lambda\mu_1 + (1-\lambda)\mu_2$. Then both μ_1 and μ_2 are absolutely continuous with respect to ν. Therefore,

$$\begin{aligned} \hat{h}(t,x,\nu) &= \inf\{\sup_{z \in E} h(t,x,z) : E \text{ such that } \nu(Z-E) = 0\} \\ &\geq \inf\{\sup_{z \in E} h(t,x,z) : E \text{ such that } \mu_i(Z-E) = 0\} \\ &= \hat{h}(t,x,\mu_i), \qquad i = 1,2. \end{aligned}$$

Recall that we are assuming that $h \geq 0$ to simplify matters.

On the other hand, if $\hat{h}(t,x,\nu) > \max\{\hat{h}(t,x,\mu_1), \hat{h}(t,x,\mu_2)\} \equiv a$, then there is a set $E \subset Z$ such that $\nu(E) > 0$ and $h(t,x,z) > a, \forall\, z \in E$. Therefore, either $\mu_1(E) > 0$ or $\mu_2(E) > 0$, or both, and so either $\hat{h}(t,x,\mu_1) > a$ or $\hat{h}(t,x,\mu_1) > a$, which contradicts the definition of a. □

To prove that $\hat{h}(\cdot,\cdot,\mu)$ is not actually convex in general, consider the following example.

Example. Suppose $h(z)$ is a continuous function that has the value 1 at z_1 and the value 0 at z_2. Let $0 < \lambda < 1$ and let $\mu_i = \delta(z - z_i), i = 1,2$, and $\nu = \lambda\mu_1 + (1-\lambda)\mu_2$. Then $\hat{h}(\nu) = 1$, $\hat{h}(\mu_1) = 1$, $\hat{h}(\mu_2) = 0$. Thus, $\hat{h}(\nu) > \lambda\hat{h}(\mu_1) + (1-\lambda)\hat{h}(\mu_2)$.

We will next prove the following lemma.

LEMMA 3.2. *For each (t,x) fixed,*

$$\mu \in \mathcal{M}(Z) \mapsto \hat{h}(t,x,\mu)$$

is weakly sequentially lower semicontinuous.

Proof. If this is not true, then, given $\varepsilon > 0$, there exists a $\mu^* \in \mathcal{M}(Z)$ and a sequence of measures $\{\mu_i\} \subset \mathcal{M}(Z)$ with $\mu_i \rightharpoonup \mu^*$, and sets $\{E_i\}$ such that $\mu_i(Z - E_i) = 0$ and

$$\sup_{z \in E_i} h(s,x,z) < \|h(t,x,z)\|_{L^\infty(Z;\mu^*)} - \varepsilon$$

for all large i. Set

$$F = \{z \in Z | h(s,x,z) \leq \|h(s,x,z)\|_{L^\infty(Z;\mu^*)} - \varepsilon\}.$$

F is a compact set. Then $\mu_i(Z-F) = 0$ for all i. The fact that $\mu_i \rightharpoonup \mu^*$ implies that $\liminf_{i\to\infty} \mu_i(Z-F) \geq \mu^*(Z-F)$ since $Z - F$ is open. We conclude that $\mu^*(Z-F) = 0$. Therefore,

$$\|h(t,x,z)\|_{L^\infty(Z;\mu^*)} \leq \sup_{z \in F} h(s,x,z) \leq \|h(s,x,z)\|_{L^\infty(Z;\mu^*)} - \varepsilon,$$

which is a contradiction. □

We are now ready to state the main theorem in this section.

THEOREM 3.3. *For each fixed* $(t,x) \in [0,T] \times R^n$, *there exists an optimal relaxed control* $\mu^*(\cdot) \in \widehat{\mathcal{Z}}[t,T]$.

This theorem is actually a corollary of the following more general statement, which is what we shall prove.

THEOREM 3.4. *For each* $(t,x,a) \in \Omega \equiv [0,T] \times R^n \times R^1$, *consider the set valued map*

$$\mathcal{L}(t,x,a) = \{\eta \in R^n | \eta = f(t,x,z), a \geq h(t,x,z), \exists z \in Z\}.$$

Assume condition (A) *weakened to assume that* $h(\cdot,\cdot,z)$ *is merely lower semicontinuous in* z. *If* $\mathcal{L}(t,x,a)$ *is a convex set for each* $(t,x,a) \in \Omega$, *then for any initial position* $(t_0,x_0) \in [0,T] \times R^n$, *there exists an optimal control* $\zeta^* \in \mathcal{Z}[t_0,T]$ *and an associated optimal trajectory* ξ^*, *with* $\xi^*(t_0) = x_0$, *for the* $L^\infty[t_0,T]$ *problem* (2.4).

Proof. Under our assumptions on f and h, $\mathcal{L}$ is upper semicontinuous and upper semicontinuous with respect to set inclusion.

To verify the last statement, pick a point $(t_0,x_0,a_0) \in \Omega$. We have to show that given $\varepsilon > 0$, there is $\delta > 0$ such that $\mathcal{L}(t,x,a) \subset [\mathcal{L}(t_0,x_0,a_0)]_\varepsilon$ for all $(t,x,a) \in B_\delta(t_0,x_0,a_0)$. Here $[A]_\varepsilon$, for any set A, is the set of points whose distance from A is $\leq \varepsilon$.

Suppose this is not the case. Then there exists $\varepsilon > 0$, a sequence of points $(t_k,x_k,a_k) \to (t_0,x_0,a_0)$, $k \to \infty$, and $\eta_k \in \mathcal{L}(t_k,x_k,a_k)$ so that $d(\eta_k, \mathcal{L}(t_0,x_0,a_0)) \geq \varepsilon$. We use the notation that $d(x,A)$ is the distance from the point x to the set A. Now, $\eta_k = f(t_k,x_k,z_k)$ and $a_k \geq h(t_k,x_k,z_k)$ for each $k = 1,2,\ldots$, for some $z_k \in Z$. Since Z is compact, $z_k \to z^*$ on a subsequence, still denoted $\{k\}$. Passing to limits and using the lower semicontinuity of $h(\cdot,\cdot,z)$, we obtain $\eta_0 \equiv f(t_0,x_0,z^*)$ and $a_0 \geq h(t_0,x_0,z^*)$, which implies that $\eta_0 \in \mathcal{L}(t_0,x_0,a_0)$, which is a contradiction.

Now, let (ζ_k, ξ_k) be a minimizing sequence, with $\xi_k(t_0) = x_0$, $k = 1,2,\ldots$. Then,

$$\|h(s,\xi_k(s),\zeta_k(s))\|_{L^\infty[t_0,T]} \to V \equiv V(t_0,x_0),$$

and, on a subsequence $\{k\}$, $\xi_k \to \xi^*$ uniformly and $d\xi_k/d\tau \rightharpoonup d\xi^*/d\tau$ weakly on $[t_0,T]$.

We claim that there is a measurable function $\lambda(\tau)$ satisfying $\|\lambda(\tau)\|_{L^\infty[t_0,T]} \leq V$ and $d\xi^*/d\tau \in \mathcal{L}(\tau,\xi^*(\tau),V)$ for almost all $\tau \in [t_0,T]$.

To see this, since $d\xi_k/d\tau \rightharpoonup d\xi^*/d\tau$ by Mazur's lemma, for every j there exists an integer n_j, a set of integers $i = 1,2,\ldots,k$, a set of nonnegative numbers $\alpha_{1j},\ldots,\alpha_{kj}$, with $\sum_{i=1}^k \alpha_{ij} = 1$, such that, $n_{j+1} > n_j + k$, and, if we define

$$y_j(\tau) = \sum_{i=1}^k \alpha_{ij} \frac{d\xi_{n_j+i}}{d\tau},$$

then $y_j(\tau) \to d\xi^*/d\tau$ for a.e. $\tau \in [t_0,T]$. We are not distinguishing between convergence on a subsequence. Now define

$$\lambda_j(\tau) = \max_{1 \leq i \leq k} h(\tau, \xi_{n_j+i}(\tau), \zeta_{n_j+i}(\tau)),$$

and

$$\lambda(\tau) \equiv \liminf_{j \to \infty} \lambda_j(\tau).$$

By lower semicontinuity,

$$\|\lambda(\tau)\|_{L^\infty[t_0,T]} \leq \liminf_{j\to\infty} \max_{1 \leq i \leq k} \|h(\tau,\xi_{n_j+i}(\tau),\zeta_{n_j+i}(\tau))\|_{L^\infty[t_0,T]} = V.$$

Set

$$\sigma_j(\tau) = \sum_{i=1}^{k} \alpha_{ij} f(\tau, \xi^*(\tau), \zeta_{n_j+i}(\tau)), \qquad \vartheta_j(\tau) = \max_{1\le i\le k} h(\tau, \xi^*(\tau), \zeta_{n_j+i}(\tau)).$$

Then, since $\xi_k \to \xi^*$ uniformly in τ, by assumption (A) it follows that on a subsequence, $\vartheta_j(\tau) \to \lambda(\tau)$, and $\sigma_j(\tau) \to d\xi^*/d\tau$ for a.e. $\tau \in [t_0, T]$. Therefore, we have that

$$f(\tau, \xi^*(\tau), \zeta_{n_j+i}(\tau)) \in \mathcal{L}(\tau, \xi^*(\tau), \vartheta_j(\tau)) \text{ a.e. } \tau \in [t_0, T].$$

Since, for τ fixed, $\mathcal{L}(\tau, \xi^*(\tau), \vartheta_j(\tau))$ is a convex set, $\sigma_j(\tau) \in \mathcal{L}(\tau, \xi^*(\tau), \vartheta_j(\tau))$. By upper semicontinuity, since $\sigma_j \to d\xi^*/d\tau$ and $\vartheta_j \to \lambda$, $d\xi^*/d\tau \in \mathcal{L}(\tau, \xi^*(\tau), \lambda(\tau))$, for a.e. $\tau \in [t_0, T]$. Hence ([5], p. 94), there is a measurable map $\zeta^* \in \mathcal{Z}[t_0, T]$, such that $d\xi^*/d\tau = f(\tau, \xi^*(\tau), \zeta^*(\tau))$ and $\lambda(\tau) \ge h(\tau, \xi^*(\tau), \zeta^*(\tau))$, a.e. $\tau \in [t_0, T]$. But $V(t_0, x_0) \ge \|\lambda(\tau)\|_{L^\infty[t_0,T]}$, and we conclude that (ζ^*, ξ^*) is optimal. □

Remark 3.2. The proof of this theorem is adapted from that in [5, Thms. 4.1 and 8.1].

Remark 3.3. The proof of Theorem 3.3 is immediate by noting that $\hat{f}(\cdot, \cdot, \mu)$ is linear and $\hat{h}(\cdot, \cdot, \mu)$ is quasiconvex.

We conclude this section by giving a result that shows that quasiconvexity of h in z is a necessary condition for weak lower semicontinuity of L^∞ functionals in the calculus of variations. The proof is adapted from [6, pp. 105–107]. The analogous result for classical variational problems is that *convexity* is necessary for weak lower semicontinuity of integral functionals [6, p. 104]. Under coercivity assumptions on $h(\cdot, \cdot, z)$, the converse is also true [6, p. 112].

THEOREM 3.5. *Let Z be R^n and assume that $h(t, x, z)$ is continuous and that $f(t, x, z) = z$. Then, considering the functional $I[\xi] \equiv \|h(s, \xi(s), \xi'(s))\|_{L^\infty[a,b]}$ in the class of absolutely continuous functions, if $I[\xi]$ is weakly lower semicontinuous for each fixed $0 \le a < b \le T$, then $h(t, x, z)$ is quasiconvex in z for each (t, x).*

Proof. Suppose that $h(t_0, x_0, z)$ is not quasiconvex in z with $t_0 \in (0, T)$. Then, there is a $\sigma > 0$ and a convex combination $z_0 = \sum_{i=1}^m \lambda_i z_i, z_i \in R^n, i = 1, \ldots, m, \lambda_i \ge 0, \sum_{i=1}^m \lambda_i = 1$, such that

$$h(t_0, x_0, z_0) \ge \max_{1\le i\le m} h(t_0, x_0, z_i) + \sigma. \tag{3.2}$$

Pick $\delta > 0$ such that $t_0 + \delta < T$ and $|h(t, x, z_i) - h(t_0, x_0, z_i)| < \sigma/3$ for $i = 0, 1, \ldots, m$, if $(t, x) \in B_\delta(t_0, x_0)$. Set $\gamma = \delta$ if $|z_0| \le 1/2$ and $\gamma = \delta/(2|z_0|)$ if $|z_0| > 1/2$.

Consider the trajectory on $[t_0, t_0 + \gamma]$ corresponding to the constant control $\zeta_0(\tau) \equiv z_0$ given by $\xi_0(\tau) = x_0 + z_0(\tau - t_0)$. Now, construct a sequence of trajectories $\xi_k(\cdot)$ on $[t_0, t_0 + \gamma]$ emanating from x_0 such that $d\xi_k(\tau)/d\tau = z_r$ if

$$\tau \in \left[t_0 + \left(j + \sum_{i=1}^{r-1} \lambda_i\right) \frac{\gamma}{k}, t_0 + \left(j + \sum_{i=1}^{r} \lambda_i\right) \frac{\gamma}{k}\right], \qquad r = 1, 2, \ldots, m$$

for each $j = 0, 1, \ldots, k$. Then [6, p. 106], $\xi_k(\tau) \to \xi_0(\tau)$ uniformly and at least on a subsequence, $d\xi_k(\tau)/d\tau \rightharpoonup d\xi_0(\tau)/d\tau$ a.e. as $k \to \infty$. Set $\tau_{jr} = t_0 + (j + \sum_{i=1}^r \lambda_i)\frac{\gamma}{k}$, $r = 1, \ldots, m, j = 0, \ldots, k$. For all k sufficiently large, $\xi_k(\tau) \in B_\delta(\xi_0(\tau))$. Now we estimate

$$\begin{aligned}
I[\xi_k] &= \|h(\tau, \xi_k(\tau), \xi_k'(\tau))\|_{L^\infty[t_0, t_0+\gamma]} \\
&= \max_{1 \le j \le k} \max_{1 \le r \le m} \|h(\tau, \xi_k(\tau), z_r)\|_{L^\infty[t_{j-1,r-1}, t_{j-1,r-1} + \gamma \frac{\lambda r}{k}]} \\
&\le \max_{1 \le j \le k} \max_{1 \le r \le m} \|h(t_0, x_0, z_r)\| + \frac{\sigma}{3} \\
&\le \max_{1 \le j \le k} h(t_0, x_0, z_0) - \sigma + \frac{\sigma}{3} \quad \text{(by (3.2))} \\
&= h(t_0, x_0, z_0) - \frac{2\sigma}{3} \\
&\le \max_{1 \le j \le k} \|h(\tau, \xi_0(\tau), z_0)\|_{L^\infty\left[t_0 + \frac{(j-1)\gamma}{k}, t_0 + \frac{j\gamma}{k}\right]} + \frac{\sigma}{3} - \frac{2\sigma}{3} \\
&= \|h(\tau, \xi_0(\tau), \xi_0'(\tau))\|_{L^\infty[t_0, t_0+\gamma]} - \frac{\sigma}{3} \\
&= I[\xi_0] - \frac{\sigma}{3}.
\end{aligned}$$

This contradicts the assumed weak lower semicontinuity of I. □

4. Relaxed value = original value: $\widehat{V} = V$. The final goal is to establish that the relaxed value and original value coincide. To do that, we prove that $\widehat{V}$, which is a viscosity solution of the relaxed Bellman equation (2.16), is also a viscosity solution of the original Bellman equation (2.5). We must also show that $\widehat{V}$ satisfies the terminal condition (2.6). Then uniqueness of viscosity solutions will yield the result.

Define the relaxed hamiltonian

$$\hat{H}(t, x, r, p) = \min_{\{\mu \in \mathcal{M}(Z) | \hat{h}(t,x,\mu) \le r\}} p \cdot \hat{f}(t, x, \mu). \tag{4.1}$$

THEOREM 4.1. *$V(t, x) \equiv \widehat{V}(t, x)$ on $[0, T] \times R^n$.*

Proof. Consider first $t = T$. We have that

$$\begin{aligned}
\widehat{V}(T, x) &= \min_{\mu \in \mathcal{M}(Z)} \hat{h}(T, x, \mu) \\
&= \min_{\mu \in \mathcal{M}(Z)} \|h(T, x, z)\|_{L^\infty(Z; \mu)} \\
&\le \min_{z \in Z} h(T, x, z) = V(T, x).
\end{aligned}$$

The last inequality is true by choosing $z = z_0 \in Z$ arbitrarily and letting $\mu = \delta(z - z_0)$. The inequality $\widehat{V}(T, x) \ge V(T, x)$ follows from

$$\widehat{V}(T, x) = \min_{\mu \in \mathcal{M}(Z)} \hat{h}(T, x, \mu) = \min_{\mu \in \mathcal{M}(Z)} \|h(T, x, z)\|_{L^\infty(Z; \mu)} \ge \min_{z \in Z} h(T, x, z).$$

Therefore, $\widehat{V}(T, x) = V(T, x)$.

Similarly,

$$\min_{\mu \in \mathcal{M}(Z)} \hat{h}(t, x, \mu) = \min_{z \in Z} h(t, x, z). \tag{4.2}$$

We show next that $\hat{H}(t, x, r, p) = H(t, x, r, p)$, i.e.,

$$\min_{\{\mu \in \mathcal{M}(Z) | \hat{h}(t,x,\mu) \le r\}} p \cdot \hat{f}(t, x, \mu) = \min_{\{z \in Z | h(t,x,z) \le r\}} p \cdot f(t, x, z). \tag{4.3}$$

First, it is clear that the set $\{\mu \in \mathcal{M}(Z) | \hat{h}(t,x,\mu) \le r\}$ is nonempty if and only if the set $\{z \in Z | h(t,x,z) \le r\}$ is nonempty. Without loss of generality, we may assume the sets are nonempty since otherwise the hamiltonians are both $+\infty$. Now, $\hat{H}(t,x,r,p) \le H(t,x,r,p)$ since, for each z_0 such that $h(t,x,z_0) \le r$, we can let $\mu = \delta(z - z_0)$. To see the opposite inequality, let $\varepsilon > 0$ be given and $\mu^* \in \mathcal{M}(Z)$ satisfy

$$\hat{H}(t,x,r,p) \ge p \cdot \hat{f}(t,x,\mu^*) - \varepsilon = \int_Z p \cdot f(t,x,z)\mu^*(dz) - \varepsilon$$

and $\|h(t,x,z)\|_{L^\infty(Z;\mu^*)} \le r$. Let $B = \{z \in Z | h(t,x,z) \le r\}$. Then $\mu^*(B) = 1, \mu^*(B^c) = 0$ and

$$\begin{aligned} \hat{H}(t,x,r,p) &\ge \int_Z p \cdot f(t,x,z)\mu^*(dz) - \varepsilon \\ &= \int_B p \cdot f(t,x,z)\mu^*(dz) + \int_{B^c} p \cdot f(t,x,z)\mu^*(dz) - \varepsilon \\ &= \int_B p \cdot f(t,x,z)\mu^*(dz) - \varepsilon \\ &\ge \min_{z\in B} p \cdot f(t,x,z) - \varepsilon = H(t,x,r,p) - \varepsilon. \end{aligned}$$

Therefore, (4.3) is true. Combining (4.3) and (4.2), we conclude that $\widehat{V}$ and V both satisfy the Bellman equation (2.5) (in the viscosity sense) and the terminal condition (2.6). Uniqueness of viscosity solutions then yields that $\widehat{V}$ and V are the same function. □

Remark 4.1. A terminal cost can also be included in the results of this paper by considering the value function

$$V(t,x) = \inf_{\zeta \in \mathcal{Z}[t,T]} \max\{\|h(s,\xi(s),\zeta(s))\|_{L^\infty[t,T]}, g(\xi(T))\}.$$

In this case, V satisfies the same Bellman equation but now satisfies the terminal condition

$$V(T,x) = \max\{\min_{z\in Z} h(T,x,z), g(x)\}.$$

This problem is relaxed exactly as before.

We conjecture that when we allow lower semicontinuous terminal data g, the corresponding relaxed value function coincides with the lower semicontinuous envelope of the unrelaxed value function. Furthermore, the relaxed value function will be the unique lower semicontinuous solution of Eq. (2.5) achieving the lower semicontinuous data g. This is analogous to the classical problem considered in [3].

Remark 4.2. A formulation of the relaxed problem using chattering controls is the following:

$$\widehat{V}(t,x) = \inf_{\alpha(\cdot),\zeta(\cdot)} \|\max_i h(\tau,\xi(\tau),\zeta(\tau))\|_{L^\infty[t,T]},$$

where, $\alpha = \{\alpha_i\}_{i=1}^{n+2}$, $\alpha_i = \alpha_i(\tau) \ge 0$, $\sum_{i=1}^{n+2} \alpha_i = 1$, $\zeta = \{\zeta_i\}_{i=1}^{n+2}$, with $\zeta_i \in \mathcal{Z}[t,T]$. The max is taken over all $1 \le i \le n+2$ for which $\alpha_i(\tau) \ne 0$, and

$$\frac{d\xi}{d\tau} = \sum_{i=1}^{n+2} \alpha_i(\tau) f(\tau,\xi(\tau),\zeta_i(\tau)), \qquad \xi(t) = x.$$

Refer to [5] for chattering controls for classical problems.

REFERENCES

[1] H. Amann, *Ordinary Differential Equations*, Walter de Gruyter & Co., Berlin, Germany, 1990.

[2] E. N. Barron, *Optimal control and calculus of variations in L^∞*, in Optimal Control of Differential Equations, N. H. Pavel, ed., Marcel Dekker, New York, 1994.

[3] E. N. Barron and R. Jensen, *Optimal control and semicontinuous viscosity solutions*, Proc. Amer. Math. Soc., 113 (1992), pp. 397–402.

[4] E. N. Barron and H. Ishii, *The Bellman equation for minimizing the maximum cost*, Nonlinear Anal., TMA, 13 (1989), pp. 1067–1090.

[5] L. D. Berkovitz, *Optimal Control Theory*, Springer-Verlag, New York, 1974.

[6] L. Cesari, *Optimization Theory and Application*, Springer-Verlag, New York, 1983.

[7] M. G. Crandall and P.-L. Lions, *Viscosity solutions of Hamilton Jacobi equations*, Trans. Amer. Math. Soc., 277 (1983), pp. 1–42.

[8] M. G. Crandall, L. C. Evans, and P.-L. Lions, *Some properties of viscosity solutions of Hamilton Jacobi equations*, Trans. Amer. Math. Soc., 282 (1984), pp. 487–502.

[9] I. Ekeland and R. Temam, *Convex Analysis and Variational Problems*, North Holland, New York, 1976.

[10] E. B. Lee and L. Markus, *Foundations of Optimal Control Theory*, Wiley, New York, 1967.

[11] J. Warga, *Optimal Control of Differential and Functional Equations*, Academic Press, New York, 1972.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 33, No. 4, pp. 1040–1060, July 1995

© 1995 Society for Industrial and Applied Mathematics
005

SENSITIVITY ANALYSIS IN NONLINEAR PROGRAMS AND VARIATIONAL INEQUALITIES VIA CONTINUOUS SELECTIONS*

JIMING LIU[†]

Abstract. Sufficient conditions for solution Lipschitz continuity, piecewise differentiability, and directional differentiability are presented for parametric nonlinear programs and variational inequalities using the idea of continuous selections. The gaps between the sufficient conditions obtained here and the weakest possible conditions for the corresponding conclusions are discussed and measured by known regularity conditions.

Key words. nonlinear program, variational inequality problem, sensitivity analysis, continuous selection, Lipschitz continuity, piecewise differentiability, directional differentiability, perturbed solution

AMS subject classifications. Primary, 90C30; Secondary, 49A29, 49B50

1. Introduction.

This paper deals with sensitivity analysis for parametric nonlinear programming (NLP) problems of the form

$$\text{NLP}(\varepsilon) \qquad \underset{x}{\text{minimize}}\ f(x,\varepsilon), \quad \text{s.t. } x \in K(\varepsilon),$$

as well as for parametric variational inequality (VI) problems of the form

$$\text{VI}(\varepsilon) \qquad \text{find } x \in K(\varepsilon) \quad \text{such that } F(x,\varepsilon)^{\mathrm{T}}(x'-x) \geq 0 \text{ for any } x' \in K(\varepsilon),$$

where $K(\varepsilon) = \{x \in R^n : g(x,\varepsilon) \in R^p_- \times \{0\}^q\}$, f, F, and g are functions from $R^n \times R^t$ to R, R^n, and R^{p+q}, respectively, and $\varepsilon \in R^t$ is a perturbation parameter. We always consider that $\varepsilon_0 \in R^t$ is a fixed parameter and that NLP(ε_0) or VI(ε_0) is the original nonlinear program or variational inequality problem, while NLP(ε) or VI(ε) is the perturbed version of NLP(ε_0) or VI(ε_0).

The basic theory and computational aspects of sensitivity analysis in nonlinear programming can be found in the pioneering work of Fiacco [2]. For recent developments in this field the reader may consult Fiacco and Liu [3]. In his seminal paper on the generalized equation, Robinson [29] introduced a key notion of strong regularity and proved that if the linear independence condition and strong second-order sufficient condition hold at a stationary point of a nonlinear program, then the perturbed stationary point and its associated multiplier are single-valued locally Lipschitz functions of the perturbation parameter. At nearly the same time a similar result, without the Lipschitz continuity, was established by Kojima [13] under the Mangasarian–Fromovitz constraint qualification and general strong second-order sufficient condition as part of his investigation of strong stability. Shortly afterward, Robinson [30] obtained the upper Lipschitz continuity of the perturbed stationary point and multiplier, assuming the Mangasarian–Fromovitz constraint qualification and general second-order sufficient condition. In view of these results, Robinson [30] proposed a very natural question: can one find an appropriate combination of a constraint qualification (weaker than linear independence) and a second-order condition, under which the perturbed stationary point or local solution will be a single-valued locally Lipschitz function? He also gave an example to show that under a constraint qualification stronger than the Mangasarian–Fromovitz constraint qualification and the strongest possible second-order condition, the answer to the question is no. The first purpose of the present research is to attempt to answer this question; we show that under the assumptions of the Mangasarian–Fromovitz constraint qualification and constant

* Received by the editors May 17, 1993; accepted for publication (in revised form) February 4, 1994.

[†] Department of Operations Research, George Washington University, Washington D.C. 20052 (e-mail: jiming@seas.gwu.edu).

rank constraint qualification as well as the general strong second-order sufficient condition, the answer is yes.

The second purpose of this research stems from a recent study Liu and Falk [24] that is concerned with nonlinear two-level optimization problems. An approach in which ascent methods are used in the upper-level problem with gradient information from the lower-level problem is proposed and investigated. In the present paper we sharpen the classical results of Jittorntrum [7] and Robinson [31] concerning the directional differentiability of the perturbed local solution of a nonlinear program and the obtained results play an important role in [24].

The third motivation for this paper, along another closely related direction, comes from a most recent work of Kyparisis [19] in which he extended the solution directional differentiability result of Jittorntrum [7] in nonlinear programming to nonlinear programs and variational inequality problems with nonunique multipliers under the assumption of constant rank constraint qualification and some other standard assumptions. We observed that in terms of solution directional differentiability, the rank assumption imposed on those inequality constraints with zero multipliers can be dropped without losing the solution directional differentiability so that the assumptions used in [19] can be weakened.

Parallel to the stability and sensitivity developments in nonlinear programs, the stability and sensitivity theory for variational inequalities has been expanding rapidly after Robinson's work [29], [30] on the generalized equation. The results obtained before 1990 have been surveyed in Kyparisis [18]. More recent advances can be found in Robinson [32], [33]; Kyparisis [20]; King and Rockafellar [11]; Gowda and Pang [4]; Pang [27]; Mordukhovich [26] and Liu [21]–[23]. As with nonlinear programs, we investigate the sensitivity issue in variational inequalities and parallel conclusions are obtained.

The overall approach of the paper is via continuous selections to obtain sensitivity results. The idea of continuous selections in sensitivity analysis study has appeared in many papers, including Jittorntrum [7]; Dontchev and Jongen [1]; Jongen, Moebert, and Tammer [8]; Jongen, Wetterling, and Zwier [9]; Jongen, Twilt, and Weber [10]; Kyparisis [19]; Kummer [15]; and Klatte [12].[1] This idea is elaborated on and systematized to some extent here. Some results concerning Lipschitz continuity, piecewise differentiability, and directional differentiability of a function that is continuously selected from a finite number of other "better" functions are provided in §2.

In §3 we concentrate on sensitivity analysis of nonlinear programs. Continuous selection results for the perturbed local solution under the Mangasarian–Fromovitz constraint qualification and general strong second-order sufficient condition are established first. Then various sensitivity results can be obtained that not only sharpen some classical sensitivity results but provide new conditions for the Lipschitz continuity, piecewise differentiability, and directional differentiability of the perturbed local solution. Parallel development for variational inequality problems is made in §4. Similar to those in §3 we give continuous selection theorems for the perturbed stationary point under the Mangasarian–Fromovitz constraint qualification and general strong second-order condition and conclude then with several sensitivity results concerning Lipschitz continuity, piecewise differentiability, and directional differentiability of the perturbed stationary point in variational inequalities.

2. Lipschitz continuity, piecewise differentiability, and directional differentiability of a continuous selection.

In order to obtain the main results of the paper, we provide in this section some results concerning Lipschitz continuity, piecewise differentiability, and directional differentiability of a continuous selection.

[1] Several references cited here were brought to our attention by a referee. Also, this referee pointed out that the whole theory of implicit function theorems for nonlinear programs essentially uses the idea of continuous selections.

We shall first define precisely what we mean by piecewise differentiability. Let K be a collection of a finite number of closed convex polyhedral cones σ. K is said to be a subdivision of R^t if the union of all σ in K is R^t and for every pair of σ_1 and σ_2 in K with $\sigma_1 \cap \sigma_2 \neq \emptyset$, $\sigma_1 \cap \sigma_2$ is a common face of σ_1 and σ_2. Each $\sigma \in K$ is called a piece of K.

Let $\varepsilon_0 \in R^t$ and $N \subset R^t$ be a neighborhood of ε_0. We say that a continuous function $x(\cdot) : N \rightarrow R^n$ is piecewise differentiable at ε_0 if there exists a subdivision K of R^t such that for each piece σ of K, there are a neighborhood $U \subset N$ of ε_0 and a function $y(\cdot) \in C^1(U) : U \rightarrow R^n$ satisfying

$$x(\varepsilon_0 + d) = y(\varepsilon_0 + d) + o(d) \quad \text{for all } d \in \sigma.$$

Intuitively, a piecewise differentiable function $x(\cdot)$ at ε_0 is a result of locally bending a C^1 function at ε_0 along some faces of a subdivision K.

By definition it is obvious that if $x(\cdot)$ is piecewise differentiable at ε_0, then it is directionally differentiable at ε_0 along any direction d. It is also true that piecewise differentiability is a stronger property than the B-differentiability introduced in Robinson [31], which in turn is stronger than directional differentiability. To see this, let $Dx(\varepsilon_0; d)$ denote the directional derivative of $x(\cdot)$ at ε_0 in any direction d. According to [31] $x(\cdot)$ is said to be B-differentiable at ε_0 if one has

$$x(\varepsilon_0 + d) = x(\varepsilon_0) + Dx(\varepsilon_0; d) + o(d).$$

It is easy to see that if in addition the homogeneous function $Dx(\varepsilon_0; d)$ is piecewise linear (see Kojima [13]), i.e., there is a subdivision K such that $Dx(\varepsilon; d)$ is a linear function of d when restricted on a piece $\sigma \in K$, then $x(\cdot)$ is piecewise differentiable at ε_0. Hence B-differentiability does not imply piecewise differentiability.

An abstract and elegant Lipschitz continuous selection result was obtained in Hager [5] which we state first. Suppose that Z is a Banach space, D is a convex subset of a Banach space, and $x : D \rightarrow Z$ is a continuous function. Furthermore, suppose $I : D \rightarrow 2^{\{1,\ldots,m\}}$ (the power set of $\{1, \ldots, m\}$) has the following property:

(2.1) If $\{z_i\} \subset D$ with $z_i \rightarrow z \in D$ as $i \rightarrow \infty$ and $J \subset I(z_i)$ for all i, then $J \subset I(z)$.

The points z', $z^* \in D$ are called compatible if $I(z') = I(z^*)$ and $I(z) \subset I(z')$ for all $z \in [z', z^*]$ (the line segment joining z' and z^*).

PROPOSITION 2.1 (Hager [5]). *If λ satisfies*

$$\|x(z') - x(z^*)\|_Z \leq \lambda \|z' - z^*\|_D \tag{2.2}$$

for all compatible z', $z^* \in D$, then λ satisfies (2.2) *for all z', $z^* \in D$.*

Specifying it to our case, we obtain the following proposition that is a simple extension of a result in Dontchev and Jongen [1].

PROPOSITION 2.2. *Let $\varepsilon_0 \in R^t$ and $B(\varepsilon_0; r) = \{\varepsilon \in R^t : \|\varepsilon - \varepsilon_0\| < r\}$, where r is a positive number, and let $x_i : B(\varepsilon_0; r) \rightarrow R^n$ be a Lipschitz continuous function with Lipschitz constant λ_i, $i = 1, \ldots, m$. Suppose $x : B(\varepsilon_0; r) \rightarrow R^n$ is a continuous function such that for all $\varepsilon \in B(\varepsilon_0; r)$, $x(\varepsilon) = x_i(\varepsilon)$ for some $i \in \{1, \ldots, m\}$. Then x is Lipschitz continuous with Lipschitz constant $\lambda :=$* $\max\{\lambda_i : i = 1, \ldots, m\}$.

Proof. See the proof of Theorem 2.2 in Dontchev and Jongen [1]. □

Similarly, we have the following result for piecewise differentiability and directional differentiability of a continuous selection. It was pointed out by a referee that this result is closely related to Kummer's theorem [14] on the representation of the Clarke subdifferential.

We say that a continuous function $x(\cdot)$ is directionally differentiable at ε_0 of order p along a direction d if there are vectors $Dx(\varepsilon_0; d)$, $D^2x(\varepsilon_0; d), \ldots, D^px(\varepsilon_0; d)$ such that for scalar $s \geq 0$

$$x(\varepsilon_0 + sd) = x(\varepsilon_0) + sDx(\varepsilon_0; d) + \frac{1}{2!}s^2D^2x(\varepsilon_0; d) + \cdots + \frac{1}{p!}s^pD^px(\varepsilon_0; d) + o(s^p).$$

THEOREM 2.3. *Let $\varepsilon_0 \in R^t$ and $B(\varepsilon_0; r) = \{\varepsilon \in R^t : \|\varepsilon - \varepsilon_0\| < r\}$, where r is a positive number, and let $x_i \in C^p : B(\varepsilon_0; r) \to R^n$, $i = 1, \ldots, m$, with integer $p \geq 1$. Suppose $x : B(\varepsilon_0; r) \to R^n$ is a continuous function selected from $\{x_i\}$, i.e., for all $\varepsilon \in B(\varepsilon_0; r)$, $x(\varepsilon) = x_i(\varepsilon)$ for some $i \in \{1, \ldots, m\}$. Then x is piecewise differentiable at ε_0 and directionally differentiable at ε_0 of order p along any direction d. Furthermore, the directional derivatives $D^qx(\varepsilon_0; d)$ of order $q \leq p$ as functions of d are Lipschitzian and $D^qx(\varepsilon_0; d) = D^qx_i(\varepsilon_0; d)$ for some i.*

Proof. We show first that x is directionally differentiable at ε_0 of order p along any given direction d. Let $I(d) := \{i : i \in \{1, \ldots, m\}$ and there is a sequence $\{s_j\}$ with $s_j \downarrow 0$ such that $x(\varepsilon_0 + s_jd) = x_i(\varepsilon_0 + s_jd)\}$. Obviously $I(d) \neq \emptyset$. We claim that for any k, $q \in I(d)$, we have that $Dx_k(\varepsilon_0; d) = Dx_q(\varepsilon_0; d)$, $D^2x_k(\varepsilon_0; d) = D^2x_q(\varepsilon_0; d), \ldots, D^px_k(\varepsilon_0; d) = D^px_q(\varepsilon_0; d)$. Then it follows that x is directionally differentiable at ε_0 of order p and $D^qx(\varepsilon_0; d) = D^qx_i(\varepsilon_0; d)$ for any $i \in I(d)$ and $1 \leq q \leq p$. Without loss of generality we may assume that when restricted on $[0, r']$ for some small $r' > 0$, $x(s) = x(\varepsilon_0 + sd)$ is continuously selected from $\{x_i(s) = x_i(\varepsilon_0 + sd) : i \in I(d)\}$. To prove the assertion claimed above, we define $R_i = \{s : 0 \leq s \leq r'$ and $x(s) = x_i(s)\}$ for any $i \in I(d)$. From the continuity of x and x_i it follows that each R_i is a closed set. Define a relation $\sim$ in $I(d)$ such that $\nu \sim \ell$ if there is a sequence $\{s_j\}$ with $s_j \downarrow 0$ satisfying both $\{s_j\} \subset R_\nu$ and $\{s_j\} \subset R_\ell$.

Suppose now $\nu \sim \ell$. Then by definition there is a sequence $\{s_j\}$ with $s_j \downarrow 0$ such that $\{s_j\} \subset R_\nu$ and $\{s_j\} \subset R_\ell$. Since x_ν and x_ℓ are in C^p, we find from $\{s_j\} \subset R_\nu$ and $\{s_j\} \subset R_\ell$ that

$$\begin{aligned} x(\varepsilon_0 + s_jd) &= x(\varepsilon_0) + s_jDx_\nu(\varepsilon_0; d) + \frac{1}{2!}s_j^2D^2x_\nu(\varepsilon_0; d) + \cdots + \frac{1}{p!}s_j^pD^px_\nu(\varepsilon_0; d) + o(s^p) \\ &= x(\varepsilon_0) + s_jDx_\ell(\varepsilon_0; d) + \frac{1}{2!}s_j^2D^2x_\ell(\varepsilon_0; d) + \cdots + \frac{1}{p!}s_j^pD^px_\ell(\varepsilon_0; d) + o(s^p). \end{aligned}$$

Then it follows from $s_j \downarrow 0$ and the above relation that

$$\begin{gathered} Dx_\nu(\varepsilon_0; d) = Dx_\ell(\varepsilon_0; d), \\ D^2x_\nu(\varepsilon_0; d) = D^2x_\ell(\varepsilon_0; d), \ldots, D^px_\nu(\varepsilon_0; d) = D^px_\ell(\varepsilon_0; d). \end{gathered} \tag{2.3}$$

Furthermore, note that for any pair k, $q \in I(d)$, since $[0, r'] = \cup_{i \in I(d)} R_i$, there is a chain such that $k \sim \ell_1$, $\ell_1 \sim \ell_2, \ldots, \ell_{h-1} \sim \ell_h$, $\ell_h \sim q$. So the desired assertion follows from (2.3).

We have proved that x is directionally differentiable at ε_0 of order p along any direction d. Note that using (2.3) and the continuity of x it is not hard to show that for $1 \leq q \leq p$ the q times directional derivative $D^qx(\varepsilon_0; d)$ is continuously selected from the directional derivatives $\{D^qx_i(\varepsilon_0; d) : i = 1, \ldots, m\}$. Since each x_i is continuously differentiable at ε_0 of order p, we find from calculus (see Marlow [25, p. 202]) that

$$D^qx_i(\varepsilon_0; d) = \nabla_\varepsilon^q x_i(\varepsilon_0)d \ldots d,$$

which is Lipschitzian as a function of d. Then from Proposition 2.2 we conclude that $D^qx(\varepsilon_0; d)$ is Lipschitzian for each $1 \leq q \leq p$.

Now what we have left is to show that x is piecewise differentiable at ε_0, i.e., the directional derivative $Dx(\varepsilon_0; d)$ as a function of d is actually piecewise linear. Let $A_i = \nabla_\varepsilon x_i(\varepsilon_0)$ for $i = 1, \dots, m$. Each A_i is an n-by-t matrix. Denote the ℓth row of A_i by A_i^ℓ. For each direction d define $J(d) = \{i : i \in \{1, \dots, m\}$ and $Dx(\varepsilon_0; d) = A_i d\}$. We shall construct a subdivision K of R^t such that for each $\sigma \in K$ there exists some $i \in \{1, \dots, m\}$, $Dx(\varepsilon_0; d) = A_i d$ for all $d \in \sigma$. Note that for each direction d the relations between $Dx(\varepsilon_0; d)$ and $A_i d$, $i = 1, \dots, m$, are such that for any $i \in J(d)$, any $j \in J(d)\backslash\{i\}$, and any $1 \le \ell \le n$, one has

$$A_i^\ell d = A_j^\ell d.$$

For any $i \in J(d)$, any $j \in \{1, \dots, m\}\backslash J(d)$, there exists some ℓ, $1 \le \ell \le n$, and one has

$$\text{either } A_i^\ell d < A_j^\ell d \quad \text{or} \quad A_i^\ell d > A_j^\ell d.$$

According to the above relationship we can construct a set $\text{Rel}(d)$ for each direction d in the following way. For any $i \in J(d)$, any $j \in J(d)\backslash\{i\}$, and any $1 \le \ell \le n$,

$$\text{put } (i, j, \ell, =) \text{ into } \text{Rel}(d).$$

For any $i \in J(d)$, any $j \in \{1, \dots, m\}\backslash J(d)$, and any $1 \le \ell \le n$,

$$\begin{aligned} &\text{put } (i, j, \ell, =) \text{ into } \text{Rel}(d) \text{ if } A_i^\ell d = A_j^\ell d; \\ &\text{put } (i, j, \ell, \le) \text{ into } \text{Rel}(d) \text{ if } A_i^\ell d \le A_j^\ell d; \\ &\text{put } (i, j, \ell, \ge) \text{ into } \text{Rel}(d) \text{ if } A_i^\ell d \ge A_j^\ell d. \end{aligned}$$

In this way we have obtained finitely many different sets, say, $\text{Rel}_1, \text{Rel}_2, \dots, \text{Rel}_L$. For each Rel_k, $1 \le k \le L$, define a convex polyhedral cone as follows:

$$\begin{aligned} \sigma(k) = \{z \in R^t : \quad & A_i^\ell z = A_j^\ell z \text{ if } (i, j, \ell, =) \in \text{Rel}_k; \\ & A_i^\ell z \le A_j^\ell z \text{ if } (i, j, \ell, \le) \in \text{Rel}_k; \\ & A_i^\ell z \ge A_j^\ell z \text{ if } (i, j, \ell, \ge) \in \text{Rel}_k\}. \end{aligned}$$

Then it is not difficult to see that $K = \cup_{k \in \{1, \dots, L\}} \sigma(k)$ forms a subdivision of R^t.

At this point what remains is to show that for each $\sigma(k) \in K$ we have that $Dx(\varepsilon_0; d) = A_i d$ for all $d \in \sigma(k)$ with some $i \in \{1, \dots, m\}$. Given a $\sigma(k) \in K$ that is defined by Rel_k, from the construction of K we know that there exists at least one direction d whose $\text{Rel}(d)$ is equal to Rel_k. Take any $i \in J(d)$. Obviously $Dx(\varepsilon_0; d) = A_i d$. We shall show that $Dx(\varepsilon_0; d') = A_i d'$ for all $d' \in \sigma(k)$ by contraposition. Suppose the contrary, i.e., there is a $d' \in \sigma(k)$ such that $Dx(\varepsilon_0; d') \ne A_i d'$. Let $d(\lambda) = (1 - \lambda)d + \lambda d'$. Define

$$\delta' = \max\{\delta : \delta \in [0, 1] \quad \text{such that } Dx(\varepsilon_0; d(\lambda)) = A_i d(\lambda) \text{ for } \lambda \in [0, \delta]\}.$$

The intuitive explanation of δ' is that $Dx(\varepsilon_0; d(\lambda))$ takes the value of $A_i d(\lambda)$ for λ from 0 to δ' and then switches to some other $A_j d(\lambda)$ such that $j \notin J(d)$ right after δ', i.e., there is a sequence $\{\lambda_h\} \downarrow \delta'$ such that $Dx(\varepsilon_0; d(\lambda_h)) = A_j d(\lambda_h)$. The existence of such j and $\{\lambda_h\}$ can be proved because otherwise we would have $Dx(\varepsilon_0; d') = A_i d'$, which contradicts our assumption. By continuity we find that

$$A_i d(\delta') = A_j d(\delta'). \tag{2.4}$$

Since $j \notin J(d)$, this means that there exists at least one ℓ, $1 \le \ell \le n$, such that either $A_i^\ell d < A_j^\ell d$ or $A_i^\ell d > A_j^\ell d$. Without loss of generality we may suppose

$$A_i^\ell d < A_j^\ell d. \tag{2.5}$$

Then from the definition of Rel(d) we know that we have put $(i, j, \ell, \leq)$ into Rel(d), and consequently it follows by the definition of $\sigma(k)$ that

$$A_i^\ell d' \leq A_j^\ell d'. \tag{2.6}$$

But (2.5) and (2.6) contradict (2.4) since $\delta' \in (0, 1)$. □

Finally we give a result concerning the directional differentiability of a continuous selection. We omit its proof because it can be shown easily by reasoning similar to that used to prove Theorem 2.3.

THEOREM 2.4. *Let $\varepsilon_0 \in R^t$ and $B(\varepsilon_0; r) = \{\varepsilon \in R^t : \|\varepsilon - \varepsilon_0\| < r\}$, where r is a positive number. Suppose $x : B(\varepsilon_0; r) \to R^n$ is a continuous function selected from continuous functions $\{x_i\}$, i.e., for all $\varepsilon \in B(\varepsilon_0; r)$, $x(\varepsilon) = x_i(\varepsilon)$ for some $i \in \{1, \ldots, m\}$. If each x_i is directionally differentiable at ε_0 in a direction d, then x is directionally differentiable at ε_0 in d and its directional derivative $Dx(\varepsilon_0; d) = Dx_i(\varepsilon_0; d)$ for some i.*

3. Lipschitz continuity, piecewise differentiability, and directional differentiability of solutions of NLP. The basic conditions we shall impose on the local solution in question of a nonlinear program are the Mangasarian–Fromovitz constraint qualification and general strong second-order sufficient condition. It is well known (see [13]) that these conditions are the weakest ones to ensure the existence and local uniqueness of the perturbed local solution. We shall show that under these basic conditions the perturbed local solution can be continuously selected from the perturbed local solutions of related binding subprograms and relaxed subprograms. These subprograms are derived based on the so-called limiting index sets. Each such limiting index set determines a binding subprogram and a relaxed subprogram of NLP(ε) that have locally unique local solutions. Then using the fruitful results established in §2, various sensitivity conclusions can be drawn that not only sharpen some classical sensitivity results but provide new conditions for the Lipschitz continuity, piecewise differentiability, and directional differentiability of the perturbed local solution of the original problem.

We briefly review the optimality conditions for NLP(ε). It is well known that if certain regularity conditions, called constraint qualifications, hold at a local minimizer x to NLP(ε), then the Karush–Kuhn–Tucker (KKT) conditions or stationary conditions hold at x. There exist multipliers $u \in R_+^p \times R^q$ such that (x, u) satisfies the following generalized equation (equivalently, a variational inequality, see Robinson [30]):

$$0 \in \begin{bmatrix} \nabla_x L(x, u, \varepsilon) \\ -g(x, \varepsilon) \end{bmatrix} + N_{R^n \times R_+^p \times R^q}(x, u), \tag{3.1}$$

where $L(x, u, \varepsilon) := f(x, \varepsilon) + u^\top g(x, \varepsilon)$, and the notation N denotes the normal cone operator. For a convex set $C \subset R^t$ and $z \in R^t$

$$N_C(z) := \begin{cases} \{y \in R^t : y^\top(z' - z) \leq 0 \text{ for all } z' \in C\} & \text{if } z \in C, \\ \emptyset & \text{if } z \notin C. \end{cases}$$

A point x that satisfies (3.1) with some u is said to be a stationary point of NLP(ε) and the pair (x, u) is said to be a KKT point of NLP(ε). We shall denote the set of multipliers associated with a stationary point x of NLP(ε) by $M(x, \varepsilon)$, i.e., $M(x, \varepsilon) = \{u : (x, u)$ is a KKT point of NLP(ε)$\}$, and denote the set of extreme points of $M(x, \varepsilon)$ by $E(x, \varepsilon)$. If $u \in R_+^p \times R^q$, define the index set $PM(u)$ that represents the inequality constraints with positive multipliers and the equation constraints, i.e., $PM(u) = \{i \in \{1, \ldots, p\} : u_i > 0\} \cup \{p + 1, \ldots, p + q\}$. The stationary point set and local minimizer set of NLP(ε) will be denoted by $SP(\varepsilon)$ and $LM(\varepsilon)$, respectively. Let x_0 be a stationary point or a local minimizer of

NLP(ε_0). For any $\delta > 0$ the localized stationary point set and local minimizer set are defined by $SP_\delta(\varepsilon) := SP(\varepsilon) \cap B(x_0;\delta)$ and $LM_\delta(\varepsilon) := LM(\varepsilon) \cap B(x_0;\delta)$, respectively, where $B(x_0;\delta) = \{x \in R^n : \|x - x_0\| < \delta\}$.

Let L be a subset of $\{1,\dots,p,p+1,\dots,p+q\}$. We define the so-called binding subprogram

$$\text{BNLP}_L(\varepsilon) \qquad \underset{x}{\text{minimize}}\ f(x,\varepsilon), \quad \text{s.t. } x \in BK_L(\varepsilon),$$

where $BK_L(\varepsilon) = \{x \in R^n : g_L(x,\varepsilon) \in \{0\}^\ell\}$, $g_L(x,\varepsilon)$ consists of functions $g_i(x,\varepsilon)$, $i \in L$, and ℓ is the appropriate corresponding dimensionality, and the relaxed subprogram

$$\text{RNLP}_L(\varepsilon) \qquad \underset{x}{\text{minimize}}\ f(x,\varepsilon), \quad \text{s.t. } x \in RK_L(\varepsilon),$$

where $RK_L(\varepsilon) = \{x \in R^n : g_L(x,\varepsilon) \in R_-^{p'} \times \{0\}^{q'}\}$, and p', q' are the appropriate corresponding dimensionalities. Loosely speaking, the binding subprogram forces some inequality constraints of the original program to be binding but relaxes some inequality constraints. Hence its solution may be or may not be a solution of the original program. The relaxed subprogram is obtained by relaxing some constraints of the original program. Therefore, any solution of the original program must be a solution of the relaxed subprogram, but not vice versa.

Suppose x_0 is a stationary point of NLP(ε_0). Two important index sets of inequality constraints are defined as follows. The active set of inequality constraints is defined by

$$I(x_0,\varepsilon_0) = \{i \in \{1,\dots,p\} : g_i(x_0,\varepsilon_0) = 0\}$$

and the Lagrange active set of inequality constraints

$$J(x_0,\varepsilon_0) = \{i \in I(x_0,\varepsilon_0) : \text{there exist some } u \in M(x_0,\varepsilon_0) \text{ such that } u_i > 0\}.$$

It should be noted that the roles of inequality constraints with different characters differ substantially. If an inequality constraint $g_i(x,\varepsilon_0)$ of NLP(ε_0) is inactive at x_0, i.e., $i \notin I(x_0,\varepsilon_0)$, then we can simply eliminate it without losing any stability information near x_0. If its associated multipliers are always positive, then it can be regarded as an equation constraint locally. Finally, if it is Lagrange inactive, i.e., $i \notin J(x_0,\varepsilon_0)$, then the solution directional differentiability will not be affected by its behavior, as we shall see later. In view of the above arguments, to simplify the presentation we shall always assume that the inactive inequality constraints have been eliminated, i.e., $I(x_0,\varepsilon_0) = \{1,\dots,p\}$.

Several regularity conditions at x_0 for the unperturbed problem NLP(ε_0) will be used in what follows.

(a) The linear independence (LI) condition holds at x_0 if

$$\{\nabla_x g_i(x_0,\varepsilon_0) : i \in I(x_0,\varepsilon_0) \cup \{p+1,\dots,p+q\}\} \quad \text{are linearly independent.}$$

(b) The Mangasarian–Fromovitz constraint qualification (MFCQ) holds at x_0 if
  (i) $\{\nabla_x g_i(x_0,\varepsilon_0) : i \in \{p+1,\dots,p+q\}\}$ are linearly independent;
  (ii) There exists a vector $z \in R^n$ such that

$$\nabla_x g_i(x_0,\varepsilon_0)^{\mathsf{T}} z < 0, \qquad i \in I(x_0,\varepsilon_0),$$
$$\nabla_x g_i(x_0,\varepsilon_0)^{\mathsf{T}} z = 0, \qquad i = p+1,\dots,p+q.$$

(c) The strict Mangasarian–Fromovitz constraint qualification (SMFCQ) [16] holds at x_0 if

(i) $\{\nabla_x g_i(x_0, \varepsilon_0) : i \in J(x_0, \varepsilon_0) \cup \{p+1, \dots, p+q\}\}$ are linearly independent;

(ii) There exists a vector $z \in R^n$ such that

$$\nabla_x g_i(x_0, \varepsilon_0)^{\mathsf{T}} z < 0, \quad i \in I(x_0, \varepsilon_0) \backslash J(x_0, \varepsilon_0),$$

$$\nabla_x g_i(x_0, \varepsilon_0)^{\mathsf{T}} z = 0, \quad i \in J(x_0, \varepsilon_0) \cup \{p+1, \dots, p+q\}.$$

(d) The strong second-order sufficient condition (SSOSC) holds at x_0 with $u \in M(x_0, \varepsilon_0)$ if

$$z^{\mathsf{T}} \nabla_x^2 L(x_0, u, \varepsilon_0) z > 0 \text{ for all } z \neq 0$$

such that $z \in Z(x_0, u)$, which is defined by

$$Z(x_0, u) := \{z \in R^n : \nabla_x g_i(x_0, \varepsilon_0)^{\mathsf{T}} z = 0 \text{ if } u_i > 0,\ i \in I(x_0, \varepsilon_0),$$
$$\nabla_x g_i(x_0, \varepsilon_0)^{\mathsf{T}} z = 0,\ i \in \{p+1, \dots, p+q\}\}.$$

(e) The general strong second-order sufficient condition (GSSOSC) holds at x_0 if

$$\text{SSOSC holds at } x_0 \text{ with } u \text{ for all } u \in M(x_0, \varepsilon_0).$$

(f) The constant rank (CR) condition holds at x_0 if for any subset $L \subset I(x_0, \varepsilon_0)$ the family $\{\nabla_x g_i(x, \varepsilon) : i \in L \cup \{p+1, \dots, p+q\}\}$ remains of constant rank near the point (x_0, ε_0).

(g) The weak constant rank (WCR) condition holds at x_0 if for any subset $L \subset J(x_0, \varepsilon_0)$ the family $\{\nabla_x g_i(x, \varepsilon) : i \in L \cup \{p+1, \dots, p+q\}\}$ remains of constant rank near the point (x_0, ε_0).

Conditions (a) – (e) are well known in the literature. The CR condition introduced in Janin [6] is a constraint qualification. The WCR is new and will be discussed later. At this point let us summarize the relations between these conditions. The notation $\Rightarrow\Rightarrow$ means "strictly stronger than" in logical sense.

PROPOSITION 3.1. (i) *LI* $\Rightarrow\Rightarrow$ *SMFCQ* $\Leftrightarrow$ *uniqueness of multipliers* $\Rightarrow\Rightarrow$ *MFCQ* + *WCR*;

(ii) *LI* $\Rightarrow\Rightarrow$ *MFCQ* + *CR* $\Rightarrow\Rightarrow$ *MFCQ* + *WCR*;

(iii) *CR* $\Rightarrow\Rightarrow$ *WCR*;

(iv) *SMFCQ* $\not\Leftarrow$ *CR*, *CR* $\not\Leftarrow$ *MFCQ*.

Remark 3.2. (1) LI $\Rightarrow\Rightarrow$ SMFCQ $\Leftrightarrow$ uniqueness of multipliers $\Rightarrow\Rightarrow$ MFCQ can be found in Kyparisis [19]. (2) LI $\Rightarrow\Rightarrow$ MFCQ + CR and SMFCQ $\not\Leftarrow$ CR were shown by Kyparisis [19]. CR $\not\Leftarrow$ MFCQ was proved in Janin [6]. (3) The implications SMFCO $\Rightarrow$ WCR and CR $\Rightarrow$ WCR are evident. The counterexamples showing MFCQ + WCR $\not\Rightarrow$ SMFCQ and WCR $\not\Rightarrow$ CR are provided later.

The differentiability assumptions about the problem functions used in the paper are described in a unified way. We say that the parametric program NLP(ε) is $C^{k,\ell}$ ($k \geq 1$, $\ell \geq 0$) near a point (x_0, ε_0) if the problem functions f and g are k times continuously differentiable with respect to x for x near x_0 with every ε near ε_0, their gradients with respect to x are ℓ times continuously differentiable in (x, ε) near (x_0, ε_0).

We shall first state below Fiacco's sensitivity result for equation constrained nonlinear programs and Kojima's strong stability theorem.

PROPOSITION 3.3 (Fiacco [2]). *Let L be a subset of* $\{1, \dots, p, p+1, \dots, p+q\}$. *Suppose KKT, SSOSC, and LI hold at* x_0 *with multipliers* u_L *for BNLP*$_L(\varepsilon_0)$ *and that BNLP*$_L(\varepsilon)$ *is* $C^{k+1,k}$ ($k \geq 1$) *near* (x_0, ε_0). *Then* (a) *for* ε *in a neighborhood of* ε_0, *there exists a unique* *function* $y_L(\varepsilon) = [x_L(\varepsilon), u_L(\varepsilon)]^{\mathrm{T}} \in C^k$ *satisfying that the KKT conditions hold at* $x_L(\varepsilon)$ *with* $u_L(\varepsilon)$ *for BNLP*$_L(\varepsilon)$ *and that* $y(\varepsilon_0) = (x_0, u_L)$.

(b) *The Jacobian* $Q_L(\varepsilon)$ *of the system*

$$\begin{aligned} \nabla_x L(x_L, u_L, \varepsilon) &= 0, \\ g_i(x_L, \varepsilon) &= 0, \quad i \in L, \end{aligned} \tag{3.2}$$

with respect to (x_L, u_L) *is locally nonsingular and*

$$\nabla_\varepsilon y(\varepsilon) = Q(\varepsilon)^{-1} N(\varepsilon),$$

where $-N(\varepsilon)$ *is the Jacobian of the system* (3.2) *with respect to* ε.

PROPOSITION 3.4 (Kojima [13]). *Let* x_0 *be a stationary point of NLP*(ε_0). *Suppose that MFCQ and GSSOSC hold at* x_0 *for NLP*(ε_0) *and that NLP*(ε) *is* $C^{2,0}$ *near* (x_0, ε_0). *Then there exist* δ, $r > 0$ *and a unique continuous function* $x : B(\varepsilon_0; r) \longrightarrow R^n$ *such that* $x(\varepsilon) = SP_\delta(\varepsilon) = LM_\delta(\varepsilon)$ *and MFCQ holds at* $x(\varepsilon)$ *for all* $\varepsilon \in B(\varepsilon_0; r)$.

Under the hypotheses of the above proposition, $M(x(\varepsilon), \varepsilon)$ is a polytope for all $\varepsilon \in B(\varepsilon_0; r)$ since MFCQ holds at $x(\varepsilon)$. For any $\varepsilon \in B(\varepsilon_0; r)$, the KKT conditions hold at $(x(\varepsilon), u)$, i.e.,

$$0 \in \begin{bmatrix} \nabla_x L(x(\varepsilon), u, \varepsilon) \\ -g(x(\varepsilon), \varepsilon) \end{bmatrix} + N_{R^n \times R^p_+ \times R^q}(x(\varepsilon), u) \tag{3.3}$$

for each $u \in E(x(\varepsilon), \varepsilon)$. For any $\varepsilon \in B(\varepsilon_0; r)$ choose one $u(\varepsilon) \in E(x(\varepsilon), \varepsilon)$ and let $L(\varepsilon) = PM(u(\varepsilon))$. It follows from (3.3) that $[x(\varepsilon), u(\varepsilon)]$ is a KKT point of the binding subprogram BNLP$_{L(\varepsilon)}(\varepsilon)$. Obviously, the number of elements in the set $\{L(\varepsilon) : \varepsilon \in B(\varepsilon_0; r)\}$ is finite. We shall call some index set L belonging to this set a limiting index set in the sense that if there exists a sequence of $\{\varepsilon_j\}$ with $\varepsilon_j \longrightarrow \varepsilon_0$ such that for all j

$$L = L(\varepsilon_j).$$

Denote by LIS(x_0, ε_0) the set of all such limiting index sets; each is a subset of $\{1, \ldots, p, p+1, \ldots, p+q\}$. Without loss of generality we may assume that LIS$(x_0, \varepsilon_0) = \{L(\varepsilon) : \varepsilon \in B(\varepsilon_0; r)\}$. Now denote the elements in LIS(x_0, ε_0) by $L(1), \ldots, L(s)$. Consider the following binding subprogram

BNLP$_{L(i)}(\varepsilon)$ $\quad$ $\underset{x}{\text{minimize}}\ f(x, \varepsilon), \quad \text{s.t. } x \in BK_{L(i)}(\varepsilon)$

for $i = 1, \ldots, s$. In addition, we assume that CR holds at x_0 for NLP(ε_0). We observe that (1) for each $L(i)$ there exist $u_i \in E(x_0, \varepsilon_0)$ such that $PM(u_i) \subset L(i)$, since by definition there exists a sequence of $\{\varepsilon_j\}$ with $\varepsilon_j \longrightarrow \varepsilon_0$ such that for all j, $L(i) = L(\varepsilon_j) = PM(u(\varepsilon_j))$. So any accumulation point of $\{u(\varepsilon_j)\}$ can be chosen as such a u_i. Consequently, the KKT conditions hold at x_0 with u_i for BNLP$_{L(i)}(\varepsilon_0)$. (2) Since we have assumed that CR holds at x_0 for NLP(ε_0), this implies that for each $L(i)$ LI condition holds at x_0 for BNLP$_{L(i)}(\varepsilon_0)$. To see this, note that for each ε_j the vectors $\{\nabla_x g_k(x(\varepsilon_j), \varepsilon_j) : k \in L(i) = PM(u(\varepsilon_j))\}$ are linearly independent since $u(\varepsilon_j)$ is an extreme point of $M(x(\varepsilon_j), \varepsilon_j)$. Therefore, LI holds at x_0 for BNLP$_{L(i)}(\varepsilon_0)$ as $\{\nabla_x g_k(x(\varepsilon), \varepsilon) : k \in L(i)\}$ remains of constant rank. Furthermore, LI condition indicates that u_i is the unique multiplier at x_0 for BNLP$_{L(i)}(\varepsilon_0)$. So the fact that GSSOSC holds at x_0 for NLP(ε_0) implies that SSOSC holds at x_0 for BNLP$_{L(i)}(\varepsilon_0)$. (3) Now applying Proposition 3.4 to BNLP$_{L(i)}(\varepsilon)$ for each $L(i)$, we find that there exist δ_i, $r_i > 0$, and a continuous function $x_i : B(\varepsilon_0; r_i) \longrightarrow R^n$ such that for any $\varepsilon \in B(\varepsilon_0; r_i)$, $x_i(\varepsilon)$ is the unique stationary point of BNLP$_{L(i)}(\varepsilon)$ in $B(x_0; \delta_i)$. Also, Proposition 3.4 says that for any $\varepsilon \in$

$B(\varepsilon_0; r)$, $x(\varepsilon)$ is the unique stationary point of NLP(ε) in $B(x_0; \delta)$. (4) We may make r small enough that a common δ can be found such that for each $\varepsilon \in B(\varepsilon_0; r)$, $x(\varepsilon)$, $x_1(\varepsilon), \ldots, x_s(\varepsilon)$ are the unique stationary points of NLP(ε), BNLP$_{L(1)}, \ldots$, BNLP$_{L(s)}$ in $B(x_0; \delta)$, respectively. For any fixed $\varepsilon \in B(\varepsilon_0; r)$, let $L(i) = PM(u(\varepsilon))$. Note that the stationary point $x(\varepsilon)$ of NLP(ε) is also a stationary point of BNLP$_{L(i)}(\varepsilon)$ since $[x(\varepsilon), u(\varepsilon)]$ satisfies the KKT conditions for BNLP$_{L(i)}(\varepsilon)$. Then by the uniqueness property of stationary points, we deduce that $x(\varepsilon) = x_i(\varepsilon)$. Therefore, we have established an interesting connection between $x(\varepsilon)$ and $x_i(\varepsilon)$ that for each $\varepsilon \in B(\varepsilon_0; r)$ $x(\varepsilon)$ can be identified to some $x_i(\varepsilon)$. We summarize this fact in the following continuous selection theorem.

THEOREM 3.5. *In the above setting, i.e., under the hypotheses of Proposition* 3.4 *and CR condition, there exist* δ, $r > 0$ *and continuous functions* x, $x_i : B(\varepsilon_0; r) \to R^n$, $i = 1, \ldots, s$, *such that for any* $\varepsilon \in B(\varepsilon_0; r)$, $x(\varepsilon)$ *and* $x_i(\varepsilon)$, $i = 1, \ldots, s$, *are the unique stationary points and local minimizers of NLP*(ε) *and BNLP*$_{L(i)}(\varepsilon)$, $i = 1, \ldots, s$, *on* $B(x_0; \delta)$, *respectively. Moreover,* $x(\cdot)$ *is a continuous selection from* $\{x_i(\cdot) : i = 1, \ldots, s\}$ *on* $B(\varepsilon_0; r)$, *i.e., for each* $\varepsilon \in B(\varepsilon_0; r)$, $x(\varepsilon) = x_i(\varepsilon)$ *for some* $i \in \{1, \ldots, s\}$.

Having established the continuous selection theorem, we give below the two main results of the section on Lipschitz continuity, piecewise differentiability, and directional differentiability. Although the first one partially answers Robinson's question, we still do not know whether the imposed conditions are necessary for Lipschitz continuity of the perturbed stationary point and local solution.

THEOREM 3.6. *Let* x_0 *be a stationary point of NLP*(ε_0). *Suppose that MFCQ, CR, and GSSOSC hold at* x_0 *for NLP*(ε_0) *and that NLP*(ε) *is* $C^{k+1,k}$ ($k \geq 1$) *near* (x_0, ε_0). *Then*

(a) *there exist* δ, $r > 0$, *and a unique Lipschitz continuous function* $x : B(\varepsilon_0; r) \to R^n$ *such that* $x(\varepsilon) = SP_\delta(\varepsilon) = LM_\delta(\varepsilon)$ *for all* $\varepsilon \in B(\varepsilon_0; r)$;

(b) x *is piecewise differentiable at* ε_0 *and directionally differentiable at* ε_0 *of order* $h \leq k$ *along any direction d*;

(c) *the directional derivatives* $D^h x(\varepsilon_0; d)$ *of order* $h \leq k$ *as functions of d are Lipschitzian.*

Proof. From Theorem 3.5 we know that there exist δ, $r > 0$, and continuous functions $x, x_i : B(\varepsilon_0; r) \to R^n$, $i = 1, \ldots, s$, such that for any $\varepsilon \in B(\varepsilon_0; r)$, $x(\varepsilon)$ and $x_i(\varepsilon)$, $i = 1, \ldots, s$, are the unique stationary points and local solutions of NLP(ε) and BNLP$_{L(i)}(\varepsilon)$, $i = 1, \ldots, s$, on $B(x_0; \delta)$, respectively and that $x(\cdot)$ is continuously selected from $\{x_i(\cdot) : i = 1, \ldots, s\}$ on $B(\varepsilon_0; r)$. Since under the hypotheses of the theorem each BNLP$_{L(i)}(\varepsilon)$ is $C^{k+1,k}$ near (x_0, ε_0), after applying Proposition 3.3 to BNLP$_{L(i)}(\varepsilon)$ we conclude that each x_i is in C^k in a neighborhood of ε_0. Then the desired conclusions readily follow from Theorem 2.3. □

The same reasoning yields the following result that sharpens some well-known classical results.

THEOREM 3.7. *Let* x_0 *be a stationary point of NLP*(ε_0) *with associated multipliers* u. *Suppose that LI and SSOSC hold at* x_0 *for NLP*(ε_0) *and that NLP*(ε) *is* $C^{k+1,k}$ ($k \geq 1$) *near* (x_0, ε_0). *Then*

(a) *for* ε *in a neighborhood of* ε_0, *there exists a unique Lipschitz continuous function* $y(\varepsilon) = [x(\varepsilon), u(\varepsilon)]^{\mathrm{T}}$ *such that* $[x(\varepsilon), u(\varepsilon)]$ *is the KKT point of NLP*(ε) *and* $y(\varepsilon_0) = [x_0, u]$;

(b) y *is piecewise differentiable at* ε_0 *and directionally differentiable at* ε_0 *of order* $h \leq k$ *along any direction d*;

(c) *the directional derivatives* $D^h y(\varepsilon_0; d)$ *of order* $h \leq k$ *as functions of d are Lipschitzian.*

Remark 3.8. (1) Conclusion (a) above was first obtained in Robinson [29] using the generalized equation approach. Here the same conclusion readily follows from a simple idea of continuous selections. (2) The directional differentiability of order 1 of the perturbed KKT points was first proved by Jittorntrum [7]. Subsequently, Robinson [31] showed that the directional differentiability of order 1 of the perturbed KKT point can be sharpened to B-differentiability. Conclusion (b) above further improves the B-differentiability to piecewise differentiability. (3) To our knowledge the high-order directional differentiability of the perturbed KKT point is obtained here for the first time in the literature.

It is interesting to compare the logical relations between LI, SMFCQ, CR, and MFCQ. From Proposition 3.1 we know that

(i) LI $\Rightarrow\!\!\!\Rightarrow$ MFCQ + CR $\Rightarrow\!\!\!\Rightarrow$ MFCQ;

(ii) LI $\Rightarrow\!\!\!\Rightarrow$ MFCQ + SMFCQ (uniqueness of multipliers) $\Rightarrow\!\!\!\Rightarrow$ MFCQ.

Bearing Theorem 3.6 in mind, the reader may wonder if KKT + MFCQ + uniqueness of muliplierss + SSOSC are sufficient for part of the conclusions derived in Theorem 3.6. The next example, modified from an example of Robinson [30], shows that the Lipschitz property (a) in Theorem 3.6 does not necessarily hold in this case. This finding was inspired by Robinson [34]. However, some conclusions of Theorem 3.6 concerning the directional derivative are still true in this case. Before proceeding to the example, we shall first give a result concerning the Lipschitz property of the directional derivative of the perturbed stationary point under the assumptions of KKT, SMFCQ, and SSOSC.

PROPOSITION 3.9. *Let x_0 be a stationary point of NLP(ε_0) with its associated multipliers u. Suppose that SMFCQ and SSOSC hold at x_0 for NLP(ε_0) and that NLP(ε) is $C^{2,1}$ near (x_0, ε_0). Then there exist δ, $r > 0$, and a unique continuous function $x : B(\varepsilon_0; r) \to R^n$ such that $x(\varepsilon) = SP_\delta(\varepsilon) = LM_\delta(\varepsilon)$ for all $\varepsilon \in B(\varepsilon_0; r)$ and that $x(\varepsilon)$ is directionally differentiable at ε_0 along any direction d, and its directional derivative $Dx(\varepsilon_0; d)$ uniquely solves the following quadratic program*

$$\text{QP}(u)$$

$$\begin{aligned} \underset{z}{\text{minimize}}\ z^{\mathsf T}\nabla_x^2 Lz + 2z^{\mathsf T}\nabla_{\varepsilon x}^2 Ld, \quad \text{s.t. } &\nabla_x g_i^{\mathsf T} z + \nabla_\varepsilon g_i^{\mathsf T} d = 0, \quad i \in PM(u), \\ &\nabla_x g_i^{\mathsf T} z + \nabla_\varepsilon g_i^{\mathsf T} d \le 0, \quad i \in I(x_0, \varepsilon_0)\backslash PM(u), \end{aligned}$$

where g_i are evaluated at (x_0, ε_0) and L at (x_0, u, ε_0). Furthermore, $Dx(\varepsilon_0; d)$ is Lipschitzian as a function of d.

Proof. The directional differentiability of the perturbed solution and the formula for its directional derivative are direct applications of Corollary 3.1 in Qiu and Magnanti [28] and thus the details are omitted. Note that if we treat the direction d in QP(u) as a parameter, we have that MFCQ, CR, and SSOSC hold at any stationary point of QP(u) for any d. Applying Theorem 3.6 to QP(u), we obtain the Lipschitz property of the directional derivative. □

Remark 3.10. The directional differentiability of the perturbed local solution under assumptions of KKT, SMFCQ, and SSOSC and its formula for directional derivative were first derived in Shapiro [35] and [36], assuming more smoothness, however. His approach explicitly resorts to the second-order derivatives of the problem functions with respect to the parameter and thus is not applicable here.

Example 3.11. Consider the following parametric convex quadratic program

$$\text{QP}(\varepsilon_1, \varepsilon_2, \varepsilon_3) \qquad \begin{aligned} &\underset{x}{\text{minimize}}\ f = \frac{1}{2}\|x - (1, 0^{\mathsf T})\|^2 + \frac{1}{2}\varepsilon_1\|x\|^2, \\ &\text{s.t. } g = A(\varepsilon_2)x + a(\varepsilon_3) \le 0, \end{aligned}$$

where

$$A(\varepsilon_2) = \begin{bmatrix} -1 & 0 \\ -1 & -\varepsilon_2 \end{bmatrix}, \qquad a(\varepsilon_3) = \begin{bmatrix} 1 \\ 1 + \varepsilon_3 \end{bmatrix},$$

$\varepsilon \in R^3$ and $\varepsilon_0 = (0,0,0)^{\mathsf{T}}$. Let $x_0 = (1,0)^{\mathsf{T}}$. Obviously, QP(ε) is $C^{k+1,k}$ for any positive integer k. It is easy to show that the constraints of QP$(0,0,0)$ satisfy MFCQ at x_0. Note that the Hessian of the objective function is always positive definite provided that $\varepsilon_1 > -1$ and the constraints are linear. So the problem satisfies the strongest possible second-order condition for ε near ε_0. It is easy to verify that x_0 is a stationary point of QP$(0,0,0)$ with the unique multipliers $(0,0)^{\mathsf{T}}$. Define for $\varepsilon_1 > 0$, $\varepsilon_2 > 0$, and $0 \le \varepsilon_3 \le \varepsilon_1\varepsilon_2^2/(1+\varepsilon_1)$,

$$x(\varepsilon_1,\varepsilon_2,\varepsilon_3) = (1,0)^{\mathsf{T}} + \varepsilon_3/\varepsilon_2(0,1)^{\mathsf{T}},$$
$$u(\varepsilon_1,\varepsilon_2,\varepsilon_3) = \varepsilon_1(1,0)^{\mathsf{T}} + (1+\varepsilon_1)\varepsilon_3/\varepsilon_2^2(-1,1)^{\mathsf{T}},$$

and for $\varepsilon_1 > 0$, $\varepsilon_2 > 0$, and $\varepsilon_3 > \varepsilon_1\varepsilon_2^2/(1+\varepsilon_1)$,

$$x(\varepsilon_1,\varepsilon_2,\varepsilon_3) = (1,0)^{\mathsf{T}} + 1/((1+\varepsilon_1)(1+\varepsilon_2^2))(\varepsilon_3+\varepsilon_1\varepsilon_3-\varepsilon_1\varepsilon_2^2, \varepsilon_2(\varepsilon_1+\varepsilon_3+\varepsilon_1\varepsilon_3))^{\mathsf{T}},$$
$$u(\varepsilon_1,\varepsilon_2,\varepsilon_3) = (\varepsilon_1+\varepsilon_3+\varepsilon_1\varepsilon_3)/(1+\varepsilon_2^2)(0,1)^{\mathsf{T}}.$$

We find that these vectors satisfy the KKT conditions of QP$(\varepsilon_1,\varepsilon_2,\varepsilon_3)$. Actually, we have

$$(1+\varepsilon_1)x(\varepsilon_1,\varepsilon_2,\varepsilon_3) - (1,0)^{\mathsf{T}} + A(\varepsilon_2)^{\mathsf{T}}u(\varepsilon_1,\varepsilon_2,\varepsilon_3) = 0,$$
$$A(\varepsilon_2)x(\varepsilon_1,\varepsilon_2,\varepsilon_3) + a(\varepsilon_3) = 0,$$
$$u(\varepsilon_1,\varepsilon_2,\varepsilon_3) \ge 0.$$

Therefore, for appropriate ε near ε_0 the above vector $x(\varepsilon)$ is the unique local minimizer of QP(ε). On the other hand, for any $\varepsilon_1 > 0$, $\varepsilon_2 > 0$ and $0 \le \varepsilon_3$, $\varepsilon_3' \le \varepsilon_1\varepsilon_2^2/(1+\varepsilon_1)$

$$\|x(\varepsilon_1,\varepsilon_2,\varepsilon_3) - x(\varepsilon_1,\varepsilon_2,\varepsilon_3')\| = (\varepsilon_2)^{-1}\|\varepsilon_3 - \varepsilon_3'\| = (\varepsilon_2)^{-1}\|(\varepsilon_1,\varepsilon_2,\varepsilon_3) - (\varepsilon_1,\varepsilon_2,\varepsilon_3')\|.$$

This means that $x(\varepsilon)$ cannot be Lipschitzian in any neighborhood of ε_0. □

The reasons why KKT + SMFCQ + SSOSC fail to ensure Lipschitz continuity of the perturbed local solution should be of interest. We think this phenomenon may be explained as follows: (1) locally Lipschitz continuity of the perturbed local solution is a *neighborhood property* rather than a *pointwise property*; (2) unlike LI and MFCQ, SMFCQ is not preservable under small perturbations; it can be destroyed even by small perturbations. Hence, in terms of properties of neighborhood-type for the generally perturbed local solution, what we can expect from SMFCQ is probably as much as that from MFCQ.

As mentioned before, in terms of directional differentiability the hypotheses of Theorem 3.6 can be weakened. Before proceeding to establish this result, we give a property of MFCQ. This property plays a key role in the rest of the section.

LEMMA 3.12. *Suppose that x_0 is a stationary point of NLP(ε_0) and MFCQ holds there and that $u \in E(x_0,\varepsilon_0)$. If an index set $L \subset \{1,\dots,p,p+1,\dots,p+q\}$ with $L \supset PM(u)$ satisfies that $\{\nabla_x g_i(x_0,\varepsilon_0) : i \in L\}$ is linearly independent, then there exist some $z \in R^n$ such that*

$$\begin{aligned} &\nabla_x g_i(x_0,\varepsilon_0)^{\mathsf{T}}z < 0, \quad i \in I(x_0,\varepsilon_0)\backslash J(x_0,\varepsilon_0),\\ &\nabla_x g_i(x_0,\varepsilon_0)^{\mathsf{T}}z < 0, \quad i \in L\backslash PM(u),\\ &\nabla_x g_i(x_0,\varepsilon_0)^{\mathsf{T}}z = 0, \quad i \in PM(u). \end{aligned} \tag{3.4}$$

Proof. By definition we find that for each $i \in J(x_0,\varepsilon_0)$, there exist some $u \in M(x_0,\varepsilon_0)$ such that $u_i > 0$. Then it is not difficult to see that there exist some $u^* \in M(x_0,\varepsilon_0)$ such that

$$u_i^* > 0 \quad \text{for each } i \in J(x_0,\varepsilon_0),$$

and

$$(3.5)\qquad \nabla_x f(x_0,\varepsilon_0)+\sum_{i\in J(x_0,\varepsilon_0)} u_i^*\nabla_x g_i(x_0,\varepsilon_0)+\sum_{i=p+1}^{p+q} u_i^*\nabla_x g_i(x_0,\varepsilon_0)=0.$$

Let A denote the subspace spanned by $\{\nabla_x g_i(x_0,\varepsilon_0),\ i\in PM(u)\}$ and B denote the convex cone $\{\lambda_i\nabla_x g_i(x_0,\varepsilon_0) : i\in (I(x_0,\varepsilon_0)\backslash J(x_0,\varepsilon_0))\cup(L\backslash PM(u)),\ \lambda_i>0\}$, respectively. Then, we claim that $A\cap B=\emptyset$. Otherwise, we have

$$\left[\sum_{i\in PM(u)} k_i\nabla_x g_i(x_0,\varepsilon_0)\right]+\sum_{i\in L_1}\lambda_i\nabla_x g_i(x_0,\varepsilon_0)+\sum_{i\in L_2}\lambda_i\nabla_x g_i(x_0,\varepsilon_0)=0,$$

where $L_1\subset I(x_0,\varepsilon_0)\backslash J(x_0,\varepsilon_0)$, $L_2\subset L\backslash PM(u)$, $k_i\in R$, and $\lambda_i>0$. Notice that $(PM(u)\cap\{1,\dots,p\})\subset J(x_0,\varepsilon_0)$. Therefore, when $L_1\neq\emptyset$, by multiplying the above equation by sufficiently small $s<0$ and adding it to (3.5), the fixed index set $J(x_0,\varepsilon_0)$ can be made larger. This is a contradiction. When $L_1=\emptyset$, the above equation contradicts the linear independence assumption. Having proved that $A\cap B=\emptyset$, by the separation theorem we can find $z\in R^n$ with $z\neq 0$ such that for any $y\in A$, $z^{\mathsf T}y\geq 0$, and that for any $y\in B$, $z^{\mathsf T}y<0$. Note that $z^{\mathsf T}y\geq 0$ for any $y\in A$ implies $z^{\mathsf T}y=0$ for any $y\in A$. This establishes (3.4). □

The conclusion of Lemma 3.12 can be interpreted as follows. For any $M\subset I(x_0,\varepsilon_0)\backslash J(x_0,\varepsilon_0)$, SMFCQ holds at x_0 for RNLP$_{M\cup L}(\varepsilon_0)$, or equivalently, the multiplier u appearing in Lemma 3.12 is the unique multiplier for RNLP$_{M\cup L}(\varepsilon_0)$ at x_0.

The next result provides sufficient conditions for the directional differentiability of the perturbed local solutions of RNLP$_L(s)$ for certain index sets L. It is a refinement of Lemma 2.1 in [19].

LEMMA 3.13. *Let x_0 be a stationary point of NLP(ε_0). Suppose that MFCQ and GSSOSC hold at x_0 for NLP(ε_0) and that NLP(ε) is $C^{2,1}$ near (x_0,ε_0). Then for every index set $L\subset\{1,\dots,p,p+1,\dots,p+q\}$ with $L\supset PM(u)$ for some $u\in E(x_0,\varepsilon_0)$ such that $\{\nabla_x g_i(x_0,\varepsilon_0) : i\in L\cap J'(x_0,\varepsilon_0)\}$ are linearly independent, where $J'(x_0,\varepsilon_0)=J(x_0,\varepsilon_0)\cup\{p+1,\dots,p+q\}$, there is a locally unique continuous local minimizer $x_L(\varepsilon)$ of RNLP$_L(\varepsilon)$. Moreover, $x_L(\varepsilon)$ is directionally differentiable at ε_0 in any direction d and its directional derivative $Dx_L(\varepsilon_0;d)$ uniquely solves the following quadratic program*:

QP$_L(u)$

$$\begin{aligned}\underset{z}{\text{minimize}}\ z^{\mathsf T}\nabla_x^2 Lz+2z^{\mathsf T}\nabla_{\varepsilon x}^2 Ld,\quad &\text{s.t. } \nabla_x g_i^{\mathsf T}z+\nabla_\varepsilon g_i^{\mathsf T}d=0, && i\in PM(u),\\ &\nabla_x g_i^{\mathsf T}z+\nabla_\varepsilon g_i^{\mathsf T}d\leq 0, && i\in L\backslash PM(u),\end{aligned}$$

where u is such that $PM(u)\subset L$, the functions g_i are evaluated at (x_0,ε_0) and the function L at (x_0,u,ε_0).

Proof. We show that the KKT conditions, SMFCQ, and SSOSC hold at x_0 for RNLP$_L(\varepsilon_0)$. Then the conclusions readily follow from Propositions 3.4 and 3.9. Note that $u_i=0$ for $i\in\{1,\dots,p\}$ and $i\notin PM(u)$. Since x_0 is a stationary point of NLP(ε_0), using the assumption that $u\in E(x_0,\varepsilon_0)$ and $L\supset PM(u)$, we find that it is also a stationary point of RNLP$_L(\varepsilon_0)$ with associated multiplier u. Evidently we have $L\cap J'(x_0,\varepsilon_0)\supset PM(u)$ since $J'(x_0,\varepsilon_0)\supset PM(u)$. Applying Lemma 3.12 to the index set $L\cap J'(x_0,\varepsilon_0)$, we can deduce that SMFCQ holds at x_0 for RNLP$_L(\varepsilon_0)$. Finally, SSOSC holds at x_0 with u for RNLP$_L(\varepsilon_0)$ because $L\supset PM(u)$ and GSSOSC holds at x_0 for NLP(ε_0). □

We give below a result that presents sufficient conditions for the directional differentiability of the perturbed local solution of NLP(ε). It partially extends Proposition 3.9 and includes the main result of Kyparisis [19] in nonlinear programs as a special case.

THEOREM 3.14. *Let x_0 be a stationary point of NLP(ε_0). Suppose that MFCQ, WCR, and GSSOSC hold at x_0 for NLP(ε_0) and that NLP(ε) is $C^{2,1}$ near (x_0, ε_0). Then for ε in some neighborhood of ε_0, there exists a locally unique continuous local solution $x(\varepsilon)$ of NLP(ε) such that $x(\varepsilon)$ is directionally differentiable at ε_0 in any direction $d \in R^t$ and its directional derivative $Dx(\varepsilon_0; d)$ uniquely solves the following quadratic program for some $u \in E(x_0, \varepsilon_0)$*

QP(u)

$$\begin{aligned}\underset{z}{\text{minimize}}\ z^{\mathsf{T}}\nabla_x^2 Lz + 2z^{\mathsf{T}}\nabla_{\varepsilon x}^2 Ld, \quad \text{s.t. } &\nabla_x g_i^{\mathsf{T}} z + \nabla_\varepsilon g_i^{\mathsf{T}} d = 0, && i \in PM(u),\\ &\nabla_x g_i^{\mathsf{T}} z + \nabla_\varepsilon g_i^{\mathsf{T}} d \le 0, && i \in I(x_0, \varepsilon_0)\backslash PM(u),\end{aligned}$$

where the functions g_i are evaluated at (x_0, ε_0) and the function L at (x_0, u, ε_0).

Proof. The idea of the proof is similar to that of the proofs for Theorems 3.5 and 3.6. Consider the KKT conditions (3.3) again. Choose any $u(\varepsilon) \in E(x(\varepsilon), \varepsilon)$ and let $L(\varepsilon) = PM(u(\varepsilon))$. Then $[x(\varepsilon), u(\varepsilon)]$ is also a KKT point of the relaxed subprogram RNLP$_{L(\varepsilon)}(\varepsilon)$. The same reasoning that served for Theorem 3.5 yields that there exist δ, $r > 0$, and continuous functions x, $x_i : B(\varepsilon_0; r) \to R^n$, $i = 1, \dots, s$, such that for all $\varepsilon \in B(\varepsilon_0; r)$, $x(\varepsilon)$ and $x_i(\varepsilon)$, $i = 1, \dots, s$, are the unique stationary points of NLP(ε) and RNLP$_{L(i)}(\varepsilon)$, $i = 1, \dots, s$, on $B(x_0; \delta)$, respectively. Furthermore, $x(\cdot)$ is a continuous selection from $\{x_i(\cdot) : i = 1, \dots, s\}$ on $B(\varepsilon_0; r)$.

Consider the relaxed subprogram

$$\text{RNLP}_{L(i)}(\varepsilon) \qquad \underset{x}{\text{minimize}}\ f(x, \varepsilon), \quad \text{s.t. } x \in RK_{L(i)}(\varepsilon)$$

for $i = 1, \dots, s$. We find using arguments similar to those used for Theorems 3.5 and 3.6 that (1) for each $L(i)$, there exist $u_i \in E(x_0, \varepsilon_0)$ such that $PM(u_i) \subset L(i)$. And the KKT conditions hold at x_0 with u_i for RNLP$_{L(i)}(\varepsilon_0)$. (2) Instead of proving that LI holds at x_0 for BNLP$_{L(i)}(\varepsilon_0)$, we can show that the index set $L(i)$ satisfies the assumption in Lemma 3.13, i.e., $\{\nabla_x g_k(x_0, \varepsilon_0) : k \in L(i) \cap J'(x_0, \varepsilon_0)\}$ are linearly independent. To show this, consider the sequence $\{\varepsilon_j\}$ with $\varepsilon_j \to \varepsilon_0$ such that for all j, $L(i) = L(\varepsilon_j) = PM(u(\varepsilon_j))$. Since $u(\varepsilon_j)$ is an extreme point of $M(x(\varepsilon_j), \varepsilon_j)$, the vectors $\{\nabla_x g_k(x(\varepsilon_j), \varepsilon_j) : k \in L(i) = PM(u(\varepsilon_j))\}$ are linearly independent. Therefore, the vectors in $\{\nabla_x g_k(x_0, \varepsilon_0) : k \in L(i) \cap J'(x_0, \varepsilon_0)\}$ are linearly independent since by WCR the set $\{\nabla_x g_k(x(\varepsilon), \varepsilon) : k \in L(i) \cap J'(x_0, \varepsilon_0)\}$ remains of constant rank. (3) It is easy to see by the definition of a relaxed subprogram that if (x_0, u) is a KKT point of RNLP$_{L(i)}(\varepsilon_0)$ then it is also a KKT point of NLP(ε_0). Hence the fact that GSSOSC holds at x_0 for NLP(ε_0) implies that GSSOSC holds at x_0 for RNLP$_{L(i)}(\varepsilon_0)$. Now applying Lemma 3.13 to each RNLP$_{L(i)}(\varepsilon)$ we obtain that x_i is directionally differentiable at ε_0 in any direction d. Then from Theorem 2.4 we find that x is directionally differentiable at ε_0 in the direction d. The calculation of $Dx(\varepsilon_0; d)$ is straightforward by Lemma 3.13 and the fact that for each $i \in I(x_0, \varepsilon_0)$, since $g_i(x(\varepsilon), \varepsilon) \le 0$, the inequality

$$\nabla_x g_i(x_0, \varepsilon_0)^{\mathsf{T}} z + \nabla_\varepsilon g_i(x_0, \varepsilon_0)^{\mathsf{T}} d \le 0$$

is always true. This completes the proof. □

The following example demonstrates a situation where the KKT conditions, MFCQ, WCR, and GSSOSC hold at a local solution, so Theorem 3.14 can apply to the case. However, SMFCQ or CR does not hold at the local solution. Hence Lemma 3.13 and the sensitivity results in Kyparisis [19] are not applicable here. It also provides a counterexample showing that WCR $\not\Rightarrow$ SMFCQ and WCR $\not\Rightarrow$ CR.

Example 3.15. Consider the parametric nonlinear program

$$\begin{aligned}\underset{x}{\text{minimize}}\ f = x_1^2 + (x_2 - 1 - \varepsilon)^2, \quad \text{s.t.}\ g_1 &= x_2 \le 0,\\ g_2 &= x_2 - \varepsilon \le 0,\\ g_3 &= -x_1 - x_2 \le 0,\\ g_4 &= (x_1 - 1)^2 + (x_2 - 1)^2 - 2 - \varepsilon \le 0,\end{aligned}$$

where $\varepsilon \in R$ and $\varepsilon_0 = 0$. Let $x_0 = (0,0)^{\mathsf{T}}$. It is easy to check that the KKT conditions, MFCQ, and GSSOSC hold at x_0 and that $I(x_0,\varepsilon_0) = \{1,2,3,4\}$, $M(x_0,\varepsilon_0) = \{(u_1,u_2,u_3,u_4)^{\mathsf{T}} : u_1 \ge 0, u_2 \ge 0, u_1 + u_2 = 2, u_3 = 0, u_4 = 0\}$, $J(x_0,\varepsilon_0) = \{1,2\}$. In addition, $\nabla_x g_1(x,\varepsilon) = (0,1)^{\mathsf{T}}$, $\nabla_x g_2(x,\varepsilon) = (0,1)^{\mathsf{T}}$, $\nabla_x g_3(x,\varepsilon) = (-1,-1)^{\mathsf{T}}$, $\nabla_x g_4(x,\varepsilon) = (2(x_1-1),2(x_2-1))^{\mathsf{T}}$. Thus WCR holds at (x_0,ε_0) and therefore by Theorem 3.14 we find that the above problem has a locally unique local minimizer $x(\varepsilon)$ near ε_0 that is directionally differentiable at ε_0 in any direction. The calculation of direction derivatives can be performed using Theorem 3.14; the details are omitted. Evidently, CR does not hold at (x_0,ε_0) since the set $\{\nabla_x g_3(x,\varepsilon), \nabla_x g_4(x,\varepsilon)\}$ does not remain a constant rank near (x_0,ε_0). Also, SMFCQ does not hold since $M(x_0,\varepsilon_0)$ is not a singleton set. □

It is worth pointing out that a major difference between CR and SMFCQ is that the former requires some rank condition on the *active* constraints while the latter imposes linear independence on the *Lagrange active* constraints. This fact somehow suggests that Lipschitz continuity of the perturbed local solution depends heavily on the stability behavior of the active constraints while directional differentiability of the perturbed local solution is mainly affected by the Lagrange active constraints. This is the main motivation for introducing the WCR condition.

As far as the directional differentiability of the perturbed local solution of a $C^{2,1}$ program is concerned, the conditions KKT + MFCQ + GSSOSC are the weakest ones for the existence and local uniqueness of the perturbed local solution (see [13]). Whether they are also sufficient for directional differentiability of the perturbed local solution should be a very interesting question. Our next counterexample gives the answer *no* to this question. This example may be regarded as a modified version of Example 3.11.

Example 3.16. Consider the following parametric convex quadratic program

$$\text{QP}(t) \qquad \begin{aligned}&\underset{x}{\text{minimize}}\ f = \frac{1}{2}\|x - (1,0)^{\mathsf{T}}\|^2 + \frac{1}{2}\|x\|^2,\\ &\text{s.t.}\ g = A(t)x + a(t) \le 0,\end{aligned}$$

where

$$A(t) = \begin{bmatrix} -1 & 0 \\ -1 & -t \end{bmatrix}, \qquad a(t) = \begin{bmatrix} 1 \\ 1 + \frac{1}{2}t^2\sin(1/t)^{1/3} \end{bmatrix} \quad [\sin(1/0) := 0],$$

$t \in R$, and $t_0 = 0$. Let $x_0 = (1,0)^{\mathsf{T}}$. The vector $a(t)$ is once continuously differentiable at t_0. Note that QP(t) is a special realization of Example 3.11 with $\varepsilon_1 = 1$, $\varepsilon_2 = t$, and $\varepsilon_3 = \frac{1}{2}t^2\sin(1/t)^{1/3}$. It is easy to check (see Example 3.11) that the KKT conditions, MFCQ and GSSOSC hold at x_0 for QP(0). And for $t > 0$ one has

$$x(t) = (1,0)^{\mathsf{T}} + \frac{1}{2}t\sin(1/t)^{1/3}(0,1)^{\mathsf{T}}.$$

Thus

$$(x(t) - x(t_0))/t = \frac{1}{2}\sin(1/t)^{1/3}(0,1)^{\mathsf{T}}.$$

This implies that $x(t)$ is not directionally differentiable at t_0 in the direction $d = 1$. □

We close this section by summarizing informally the main results obtained in this section as follows.

(1) KKT + MFCQ + CR + GSSOSC $+C^{k+1,k}$ $(k \geq 1)$ $\Rightarrow$ Lipschitz continuity + piecewise differentiability + directional differentiability of order $h \leq k$ + Lipschitz continuous directional differentiability.

(2) KKT + SMFCQ + SSOSC $+C^{2,1}$ $\Rightarrow$ Lipschitz continuous directional differentiability.

(3) KKT + MFCQ + WCR + GSSOSC $+C^{2,1}$ $\Rightarrow$ directional differentiability.

(4) KKT + SMFCQ + SSOSC $+C^{k+1,k}$ $(k \geq 1)$ $\not\Rightarrow$ Lipschitz continuity.

(5) KKT + MFCQ + GSSOSC $+C^{2,1}$ $\not\Rightarrow$ directional differentiability.

4. Lipschitz continuity and directional differentiability of solutions of VI. Similar to those in §3, parallel results for VI problems can be established. It is shown under the hypotheses of the Mangasarian–Fromovitz constraint qualification and general strong second-order condition that the perturbed stationary point of the original problem can be continuously selected from the perturbed stationary points of the binding subproblems and relaxed subproblems. Then using the idea of continuous selections, we derive several sensitivity results concerning Lipschitz continuity, piecewise differentiability, and directional differentiability of the perturbed stationary point. We should mention that we shall only consider stationary points of VI problems in this paper. The corresponding results for local solutions of VI problems can be obtained if the VI problems in question satisfy the convexity condition since then a stationary point of a VI problem is actually a local solution of the problem.

The necessary conditions for VI problems resemble those for NLP. If x is a local solution to VI(ε) and some constraint qualifications hold at x, then the general Karush–Kuhn–Tucker (GKKT) conditions or stationary conditions hold at x (see [18]). There exist multipliers $u \in R^p_+ \times R^q$ such that (x, u) satisfies the following generalized equation:

$$0 \in \begin{bmatrix} L_D(x,u,\varepsilon) \\ -g(x,\varepsilon) \end{bmatrix} + N_{R^n \times R^p_+ \times R^q}(x,u), \tag{4.1}$$

where $L_D(x,u,\varepsilon) := F(x,\varepsilon) + \nabla_x g(x,\varepsilon)^{\mathsf{T}} u$. A point x that satisfies (4.1) with some u is said to be a stationary point of VI(ε) and the pair (x, u) is said to be a GKKT point of VI(ε). The set of multipliers associated with a stationary point x of VI(ε) are denoted by $M(x,\varepsilon)$ and its extreme point set by $E(x,\varepsilon)$. The sets of stationary points and local solutions of VI(ε) are denoted by $SP(\varepsilon)$ and $LS(\varepsilon)$, respectively. These sets will be localized, i.e., assuming x_0 is a stationary point or a local solution of VI(ε_0), for any $\delta > 0$ the stationary point set and local solution set are localized by letting $SP_\delta(\varepsilon) := SP(\varepsilon) \cap B(x_0;\delta)$ and $LS_\delta(\varepsilon) := LS(\varepsilon) \cap B(x_0;\delta)$, respectively.

Note that if x is a stationary point of NLP(ε), then various second-order sufficient conditions can ensure that x is a local solution of NLP(ε). However, solution conditions for VI problems are usually more restrictive. A commonly used one is the so-called convexity assumption (see [18]). We say that VI(ε) satisfies the convexity assumption if $g_i(\cdot,\varepsilon)$, $i = 1,\ldots,p$, are convex and $g_i(\cdot,\varepsilon)$, $i = p+1,\ldots,p+q$, are affine. In this case a stationary point of VI(ε) is also a local solution.

Let L be a subset of $\{1,\ldots,p,p+1,\ldots,p+q\}$. The corresponding binding subproblem is defined by

BVI$_L(\varepsilon)$

$$\text{find}\, x \in BK_L(x) \quad \text{such that}\, F(x,\varepsilon)^{\mathsf{T}}(x'-x) \geq 0\, \text{for any}\, x' \in BK_L(\varepsilon),$$

where $BK_L(\varepsilon) = \{x \in R^n : g_L(x, \varepsilon) \in \{0\}^\ell\}$, $g_L(x, \varepsilon)$ consists of functions $g_i(x, \varepsilon)$, $i \in L$, and ℓ is the appropriate corresponding dimensionality, and the relaxed subproblem is

$\mathrm{RVI}_L(\varepsilon)$

$$\text{find}\, x \in RK_L(x) \quad \text{such that}\, F(x, \varepsilon)^{\mathsf{T}}(x' - x) \geq 0\, \text{for any}\, x' \in RK_L(\varepsilon),$$

where $RK_L(\varepsilon) = \{x \in R^n : g_L(x, \varepsilon) \in R_-^{p'} \times \{0\}^{q'}\}$, and p', q' are the appropriate corresponding dimensionalities.

Suppose x_0 is a stationary point of VI(ε_0). The index sets $I(x_0, \varepsilon_0)$, $J(x_0, \varepsilon_0)$, and the regularity conditions LI, MFCQ, SMFCQ, CR, and WCR used in §3 apply to VI(ε_0) without any modification. The difference in the second-order conditions between NLP (ε_0) and VI(ε_0) is the replacement of $\nabla_x L(x, u, \varepsilon)$ with $L_D(x, u, \varepsilon)$.

(a) The SSOC [18] holds at x_0 with $u \in M(x_0, \varepsilon_0)$ if

$$z^{\mathsf{T}} \nabla_x L_D(x_0, u, \varepsilon_0) z > 0 \quad \text{for}\, 0 \neq z \in Z(x_0, u),$$

where $Z(x_0, u)$ is defined as in §3.

(b) The GSSOC holds at x_0 if SSOC holds at x_0 with u for all $u \in M(x_0, \varepsilon_0)$.
The differentiability assumptions for the VI problem functions are introduced in the following text. We say that the parametric VI problem VI(ε) is $C^{k,\ell}$ ($k \geq 1$, $\ell \geq 0$) near a point (x_0, ε_0) if the problem functions F and g are $(k - 1)$ and k times continuously differentiable with respect to x (near x_0) for ε near ε_0, respectively, F and $\nabla_x g$ are ℓ times continuously differentiable in (x, ε) near (x_0, ε_0).

The VI versions of Fiacco's basic sensitivity theorem and Kojima's strong stability theorem in nonlinear programs were obtained in Tobin [37] and Liu [23], which we shall state first.

PROPOSITION 4.1 (Tobin [37]). *Let L be a subset of* $\{1, \ldots, p, p+1, \ldots, p+q\}$. *Suppose KKT, SSOC, and LI hold at* x_0 *with multipliers* u_L *for* $BVI_L(\varepsilon_0)$ *and that* $BVI_L(\varepsilon)$ *is* $C^{k+1,k}$ ($k \geq 1$) *near* (x_0, ε_0). *Then*

(a) *for* ε *in a neighborhood of* ε_0, *there exists a unique function* $y_L(\varepsilon) = (x_L(\varepsilon), u_L(\varepsilon))^{\mathsf{T}} \in C^k$, *where* $(x_L(\varepsilon), u_L(\varepsilon))$ *satisfies the GKKT conditions for* $BVI_L(\varepsilon)$ *and* $y(\varepsilon_0) = (x_0, u_L)$.

(b) *The Jacobian* $Q_L(\varepsilon)$ *of the system*

$$\begin{aligned} L_D(x_L, u_L, \varepsilon) &= 0, \\ g_i(x_L, \varepsilon) &= 0, \quad i \in L, \end{aligned} \tag{4.2}$$

with respect to (x_L, u_L) *is locally nonsingular and*

$$\nabla_\varepsilon y(\varepsilon) = Q(\varepsilon)^{-1} N(\varepsilon),$$

where $-N(\varepsilon)$ *is the Jacobian of the system* (4.2) *with respect to* ε.

PROPOSITION 4.2 (Liu [23]). *Let* x_0 *be a stationary point of VI*(ε_0). *Suppose that MFCQ and GSSOC hold at* x_0 *for VI*(ε_0) *and that VI*(ε) *is* $C^{2,0}$ *near* (x_0, ε_0). *Then there exist* δ, $r > 0$, *and a continuous function* $x : B(\varepsilon_0; r) \to R^n$ *such that* $x(\varepsilon) = SP_\delta(\varepsilon)$ *and MFCQ holds at* $x(\varepsilon)$ *for all* $\varepsilon \in B(\varepsilon_0; r)$.

The limiting index sets for VI(ε_0) at x_0 can be defined in a way similar to that used in §3. The union of limiting index sets are denoted again by LIS(x_0, ε_0) and its elements by $L(1), \ldots, L(s)$. For each following binding VI subproblem

$\mathrm{BVI}_{L(i)}(\varepsilon)$

$$\text{find } x \in BK_{L(i)}(x) \quad \text{such that } F(x,\varepsilon)^{\mathsf{T}}(x'-x) \geq 0 \text{ for any } x' \in BK_{L(i)}(\varepsilon),$$

where $i \in \{1,\ldots,s\}$, under the hypotheses of Proposition 4.2 and CR condition, it can be shown (similar to the case of nonlinear programs) that the GKKT conditions, MFCQ, and GSSOC hold at x_0 for $\mathrm{BVI}_{L(i)}(\varepsilon_0)$. Then by Proposition 4.2 there exist δ_i, $r_i > 0$, and a continuous function $x_i : B(\varepsilon_0; r_i) \to R^n$ such that $x_i(\varepsilon)$ is the unique stationary point of $\mathrm{BVI}_{L(i)}(\varepsilon)$ in $B(x_0;\delta_i)$. Furthermore, it can be proved that x can be continuously selected from $\{x_i\}$ locally. Thus we have the following continuous selection theorem.

THEOREM 4.3. *Under the hypotheses of Proposition* 4.2 *and the CR condition, there exist* δ, $r > 0$, *and continuous functions* x, $x_i : B(\varepsilon_0; r) \to R^n$, $i = 1,\ldots,s$, *such that for any* $\varepsilon \in B(\varepsilon_0; r)$, $x(\varepsilon)$ *and* $x_i(\varepsilon)$, $i = 1,\ldots,s$, *are the unique stationary points of VI*(ε) *and* $BVI_{L(i)}(\varepsilon)$, $i = 1,\ldots,s$, *on* $B(x_0;\delta)$, *respectively. Moreover,* $x(\cdot)$ *is a continuous selection from* $\{x_i(\cdot) : i = 1,\ldots,s\}$ *on* $B(\varepsilon_0; r)$, *i.e., for each* $\varepsilon \in B(\varepsilon_0; r)$, $x(\varepsilon) = x_i(\varepsilon)$ *for some* $i \in \{1,\ldots,s\}$.

The next result provides sufficient conditions for Lipschitz continuity, piecewise differentiability, and directional differentiability of the perturbed stationary point.

THEOREM 4.4. *Let* x_0 *be a stationary point of VI*(ε_0). *Suppose that MFCQ, CR, and GSSOSC hold at* x_0 *for VI*(ε_0) *and that VI*(ε) *is* $C^{k+1,k}$ $(k \geq 1)$ *near* (x_0,ε_0). *Then*

(a) *there exist* δ, $r > 0$, *and a unique Lipschitz continuous function* $x : B(\varepsilon_0; r) \to R^n$ *such that* $x(\varepsilon) = SP_\delta(\varepsilon) = LS_\delta(\varepsilon)$ *for all* $\varepsilon \in B(\varepsilon_0; r)$;

(b) x *is piecewise differentiable at* ε_0 *and directionally differentiable at* ε_0 *of order* $h \leq k$ *along any direction d*;

(c) *the directional derivatives* $D^h x(\varepsilon_0; d)$ *of order* $h \leq k$ *as functions of* d *are Lipschitzian.*

Proof. The proof of the theorem is similar to that of Theorem 3.6 except that it uses Theorem 4.3 and Proposition 4.1 instead of using Theorem 3.5 and Proposition 3.3. □

The following theorem improves the classical results of Kyparisis [17] concerning the directional differentiability of the perturbed GKKT point. Its proof is similar to that of Theorem 3.7 and is omitted. Conclusion (a) in this theorem is proved by Robinson [29].

THEOREM 4.5. *Let* x_0 *be a stationary point of VI*(ε_0) *with associated multiplier* u. *Suppose that LI and SSOC hold at* x_0 *for VI*(ε_0) *and that VI*(ε) *is* $C^{k+1,k}$ $(k \geq 1)$ *near* (x_0,ε_0). *Then*

(a) *for* ε *in a neighborhood of* ε_0, *there exists a unique Lipschitz continuous function* $y(\varepsilon) = [x(\varepsilon), u(\varepsilon)]^{\mathsf{T}}$ *such that* $[x(\varepsilon), u(\varepsilon)]$ *is the GKKT point of VI*(ε) *and* $y(\varepsilon_0) = [x_0, u]$;

(b) y *is piecewise differentiable at* ε_0 *and directionally differentiable at* ε_0 *of order* $h \leq k$ *along any direction d*;

(c) *the directional derivatives* $D^h y(\varepsilon_0; d)$ *of order* $h \leq k$ *as functions of d are Lipschitzian.*

In order to establish the directional differentiability of the perturbed stationary point, we need the following result concerning a relaxed VI subproblem.

LEMMA 4.6. *Let* x_0 *be a stationary point of VI*(ε_0). *Suppose that MFCQ and GSSOC hold at* x_0 *for VI*(ε_0) *and that VI*(ε) *is* $C^{2,1}$ *near* (x_0,ε_0). *Then, for every index set* $L \subset \{1,\ldots,p,p+1,\ldots,p+q\}$ *with* $L \supset PM(u)$ *for some* $u \in E(x_0,\varepsilon_0)$ *such that* $\{\nabla_x g_i(x_0,\varepsilon_0) : i \in L \cap J'(x_0,\varepsilon_0)\}$ *is linearly independent, where* $J'(x_0,\varepsilon_0) = J(x_0,\varepsilon_0) \cup \{p+1,\ldots,p+q\}$, *there is a locally unique continuous stationary point* $x_L(\varepsilon)$ *of* $RVI_L(\varepsilon)$. *Moreover,* $x_L(\varepsilon)$ *is directionally differentiable at* ε_0 *in any direction* d *and its directional derivative* $Dx(\varepsilon_0; d)$ *uniquely solves the following linear variational inequality*

LRVI$_L(u)$

find $z \in LRK_L(u)$ *such that* $(\nabla_x L_D z + \nabla_\varepsilon L_D d)^{\mathsf{T}}(z' - z) \geq 0$ *for all* $z' \in LRK_L(u)$

where

$$LRK_L(u) = \{z : \nabla_x g_i^{\mathsf{T}} z + \nabla_\varepsilon g_i^{\mathsf{T}} d = 0, \quad i \in M(u),$$
$$\nabla_x g_i^{\mathsf{T}} z + \nabla_\varepsilon g_i^{\mathsf{T}} d \leq 0, \quad i \in L \backslash M(u)\},$$

where the functions g_i *are evaluated at* (x_0, ε_0) *and the function* L_D *at* (x_0, u, ε_0).

Proof. The proof is analogous to that of Lemma 3.13. The differences are that we use Proposition 4.2 instead of Proposition 3.4 and Corollary 3.1 in Qiu and Magnanti [28] instead of Proposition 3.9. □

Finally, we give below a sensitivity result for VI problems concerning the directional differentiability, assuming GKKT, MFCQ, WCR, and GSSOC. It covers the results of both Kyparisis [19] and Qiu and Magnanti [28] on directional differentiability. The former [19] assumes GKKT, MFCQ, CR, and GSSOC, while the latter [28] uses GKKT, SMFCQ, and SSOC.

THEOREM 4.7. *Let* x_0 *be a stationary point of VI*(ε_0). *Suppose that MFCQ, WCR, and GSSOC hold at* x_0 *for VI*(ε_0) *and that VI*(ε) *is* $C^{2,1}$ *near* (x_0, ε_0). *Then for* ε *in some neighborhood of* ε_0, *there exists a locally unique continuous stationary point* $x(\varepsilon)$ *to VI*(ε) *that is directionally differentiable at* ε_0 *in any direction* d *and its directional derivative* $Dx(\varepsilon_0; d)$ *uniquely solves the following linear variational inequality for some* $u \in E(x_0, \varepsilon_0)$:

LVI(u)

find $z \in LK(u)$ *such that* $(\nabla_x L_D z + \nabla_\varepsilon L_D d)^{\mathsf{T}}(z' - z) \geq 0$ *for all* $z' \in LK_M(u)$,

where

$$LK(u) = \{z : \nabla_x g_i^{\mathsf{T}} z + \nabla_\varepsilon g_i^{\mathsf{T}} d = 0, \quad i \in PM(u),$$
$$\nabla_x g_i^{\mathsf{T}} z + \nabla_\varepsilon g_i^{\mathsf{T}} d \leq 0, \quad i \in I(x_0, \varepsilon_0) \backslash PM(u)\},$$

where the functions g_i *are evaluated at* (x_0, ε_0) *and the function* L_D *at* (x_0, u, ε_0). *Furthermore, if we strengthen MFCQ and WCR to SMFCQ, then the directional derivative* $Dx(\varepsilon_0; d)$ *is Lipschitzian as a function of* d.

Proof. The first part of the theorem can be proved in the same way as in the proof of Theorem 3.14 except that we use Lemma 4.6 instead of Lemma 3.13. The second part of the conclusions can be shown by applying Theorem 4.4 to LVI(u). □

To see an example where Theorem 4.7 is applicable but the sensitivity results of Kyparisis [19] and Qiu and Magnanti [28] are not, we let $F(x, \varepsilon) = \nabla_x f(x, \varepsilon)$, where $f(x, \varepsilon)$ is the objective function in Example 3.15 and there is no change in the constraints. We omit the details.

Note that the KKT conditions of the parametric convex quadratic programs in both Examples 3.11 and 3.16 are variational inequalities defined over perturbed polyhedral sets. Therefore, letting $F(x, \varepsilon) = \nabla_x f(x, \varepsilon)$, we obtain two nice examples in variational inequalities that demonstrate that GKKT + SMFCQ + SSOC $+C^{2,1} \not\Rightarrow$ Lipschitz continuity of the perturbed stationary point and that GKKT + MFCQ + GSSOC $+C^{2,1} \not\Rightarrow$ directional differentiability of the perturbed stationary point, respectively.

As in §3, we end this section by informally summarizing the main results derived as follows. Note that these relations hold only for the perturbed stationary point. But if in addition each VI(ε) meets the convexity assumption, then these conclusions are also valid for the perturbed local solution.

(1) GKKT + MFCQ + CR + GSSOC $+C^{k+1,k}$ $(k \geq 1)$ $\Rightarrow$ Lipschitz continuity + piecewise differentiability + directional differentiability of order $h \leq k$+ Lipschitz continuous directional differentiability.

(2) GKKT + SMFCQ + SSOC $+C^{2,1}$ $\Rightarrow$ Lipschitz continuous directional differentiability.

(3) GKKT + MFCQ + WCR + GSSOC $+C^{2,1}$ $\Rightarrow$ directional differentiability.

(4) GKKT + SMFCQ + SSOC $+C^{2,1}$ $\not\Rightarrow$ Lipschitz continuity.

(5) GKKT + MFCQ + GSSOC $+C^{2,1}$ $\not\Rightarrow$ directional differentiability.

Acknowledgments. The author is indebted to Professor S. M. Robinson for [34], which led to Example 3.11, to Professor A.V. Fiacco for his interest, and to two anonymous referees and the associate editor who handled this paper for their valuable comments and helpful references.

REFERENCES

[1] A. L. Dontchev and H. Th. Jongen, *On the regularity of the Kuhn-Tucker curve*, SIAM J. Control Optim., 24 (1986), pp. 169–176.

[2] A. V. Fiacco, *Introduction to Sensitivity and Stability Analysis in Nonlinear Programming*, Academic Press, New York, 1983.

[3] A. V. Fiacco and J. Liu, *Degeneracy in NLP and the development of results motivated by its presence*, Ann. Oper. Res., 46 (1993), pp. 61 – 80.

[4] M. S. Gowda and J. S. Pang, *Stability analysis of variational inequalities and nonlinear complementarity problems, via the mixed linear complementarity problem and degree theory*, Math. Oper. Res., 19 (1994), pp. 831 – 879.

[5] W. W. Hager, *Lipschitz continuity for constrained processes*, SIAM J. Control Optim., 17 (1979), pp. 321–338.

[6] R. Janin, *Directional derivative of the marginal function in nonlinear programming*, Math. Progr. Study, 21 (1984), pp. 110–126.

[7] K. Jittorntrum, *Solution point differentiability without strict complementarity in nonlinear programming*, Math. Progr. Study, 21 (1984), pp. 127–138.

[8] H. Th. Jongen, T. Moebert, and K. Tammer, *On iterated minimization in nonconvex optimization*, Math. Oper. Res., 11 (1986), pp. 679–691.

[9] H. Th. Jongen, W. Wetterling, and G. Zwier, *On sufficient conditions for local optimality in semi-infinite optimization*, Optimization, 18 (1987), pp. 165–178.

[10] H. Th. Jongen, F. Twilt, and G. W. Weber, *Semi-infinite optimization: Structure and stability of the feasible set*, J. Optim. Theory Appl., 72 (1992), pp. 529–552.

[11] A. J. King and R. T. Rockafellar, *Sensitivity analysis for nonsmooth generalized equations*, Math. Programming, 55 (1992), pp. 193–212.

[12] D. Klatte, *Stability of stationary solution in semi-infinite optimization via the reduction approach*, in Advances in Optimization, W. Oettli, D. Pallaschke, eds., Springer, Berlin, 1992, pp. 155–170.

[13] M. Kojima, *Strongly stable stationary solutions in nonlinear programs*, in Analysis and Computation of Fixed Points, S. M. Robinson, ed., Academic Press, New York, 1980, pp. 93–138.

[14] B. Kummer, *Newton's method for non-differentiable functions*, in Advances in Mathematical Optimization, J. Guddat, et al., eds., Akademic-Verlag, Berlin, 1988, pp. 114–125.

[15] ———, *An implicit-function theorem for $C^{0,1}$ — equations and parametric $C^{1,1}$ — optimization*, J. Math. Anal. Appl., 158 (1991), pp. 35–46.

[16] J. Kyparisis, *On uniqueness of Kuhn–Tucker multipliers in nonlinear programming*, Math. Programming, 32 (1985), pp. 242–246.

[17] ———, *Sensitivity analysis framework for variational inequalities*, Math. Programming, 38 (1987), pp. 203–213.

[18] ———, *Sensitivity analysis for variational inequalities and nonlinear complementarity problems*, Ann. Oper. Res., 27 (1990), pp. 143–174.

[19] ———, *Sensitivity analysis for nonlinear programs and variational inequalities with nonunique multipliers*, Math. Methods Oper. Res., 15 (1990), pp. 286–298.

[20] ———, *Parametric variational inequalities with multivalued solution sets*, Math. Methods Oper. Res., 12 (1992), pp. 341–364.

[21] J. Liu, *Linear stability of generalized equations. Part* I. *Basic theory*, Math. Oper. Res., 19 (1994), pp. 706 - 720.

[22] ———, *Linear stability of generalized equations. Part* II. *Applications to nonlinear programming*, Math. Oper. Res. 19 (1994), pp. 721 - 742.

[23] ———, *Strong stability in variational inequalities*, SIAM J. Control Optim., 33 (1995), pp. 725 - 749.

[24] J. Liu and J. E. Falk, *On bi-level programming. Part* I. *General nonlinear case*, Tech. Paper T-575, Department of Operations Research, George Washington University, 1993.

[25] W. H. Marlow, *Mathematics For Operations Research*, John Wiley & Sons, New York, 1978.

[26] B. Mordukhovich, *Stability theory for parametric generalized equations and variational inequalities via nonsmooth analysis*, Institute for Mathematics and Its Applications, University of Minnesota, 1992, preprint.

[27] J. S. Pang, *A degree-theoretic approach to parametric nonsmooth equations with multivalued solution sets*, Math. Programming, to appear.

[28] Y. Qiu and T. L. Magnanti, *Sensitivity analysis for variational inequalities*, Math. Methods Oper. Res., 17 (1992), pp. 61-76.

[29] S. M. Robinson, *Strongly regular generalized equations*, Math. Methods Oper. Res., 5 (1980), pp. 43–62.

[30] ———, *Generalized equations and their solutions. Part* II. *Applications to nonlinear programming*, Math. Progr. Study, 19 (1982), pp. 200–221.

[31] ———, *Local structure of feasible sets in nonlinear programming. Part* III. *Stability and sensitivity*, Math. Progr. Study, 30 (1987), pp. 45–66.

[32] ———, *Homeomorphism conditions for normal maps of polyhedra*, in Optimization and Nonlinear Analysis, A. Ioffe, M. Marcus, and S. Reich, eds., Longman, Harlow, U.K., to appear.

[33] ———, *Normal maps induced by linear transformations*, Math. Oper. Res., 17 (1992), pp. 691–714.

[34] ———, Personal communication, August 1992.

[35] A. Shapiro, *Second order sensitivity analysis and asymptotic theory of parameterized nonlinear programs*, Math. Programming, 33 (1985), pp. 280–299.

[36] ———, *Sensitivity analysis of nonlinear programs and differentiability properties of metric projections*, SIAM J. Control Optim., 26 (1988), pp. 628–645.

[37] R. L. Tobin, *Sensitivity analysis for variational inequalities*, J. Optim. Theory Appl., 48 (1986), pp. 191–204.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 33, No. 4, pp. 1061–1070, July 1995

© 1995 Society for Industrial and Applied Mathematics

006

A RESULT CONCERNING CONTROLLABILITY FOR THE NAVIER–STOKES EQUATIONS*

E. FERNÁNDEZ-CARA[†]AND M. GONZÁLEZ-BURGOS[†]

Abstract. The main goal of this paper is to present a new result concerning controllability of the time-dependent Navier–Stokes equations. Here, the control variable is the trace of the velocity field on a "small" part of the boundary. The main result states that the linear space spanned by final states is dense in the L^2 space of admissible fields. For the proof, one uses a duality argument that is suggested by the linear theory. This reduces the task to an existence/regularity result for a nonlinear problem.

Key words. approximate controllability, Navier–Stokes equations, nonlinear parabolic partial differential equations

AMS subject classifications. 93C20, 93B05, 76D05, 35D05

1. Statement of the problem: The main result.

In what follows it will be assumed that $\Omega \subset \mathbb{R}^N$ is a bounded open set ($N = 2$ or 3) whose boundary $\partial\Omega$ is of class $C^{1,1}$. We denote by γ a component of $\partial\Omega$ and we assume that $\partial\Omega\backslash\gamma$ has positive measure. We consider the following spaces:

$$\tilde{\mathcal{V}}(\Omega) = \{\mathbf{v}; \mathbf{v} \in \mathcal{D}(\bar{\Omega})^N,\ \nabla \cdot \mathbf{v} = 0,\ \operatorname{Supp} \mathbf{v} \subset \Omega \cup \gamma\},$$

$$\tilde{H}(\Omega) = \text{the closure of } \tilde{\mathcal{V}}(\Omega) \text{ in the space } L^2(\Omega)^N,$$

$$\tilde{V}(\Omega) = \text{the closure of } \tilde{\mathcal{V}}(\Omega) \text{ in the space } H^1(\Omega)^N.$$

Obviously, $\tilde{V}(\Omega)$ and $\tilde{H}(\Omega)$ are Hilbert spaces for the usual scalar products in $H^1(\Omega)^N$ and $L^2(\Omega)^N$, respectively. Furthermore, in $\tilde{V}(\Omega)$, the seminorm

$$\mathbf{u} \longrightarrow \|\nabla \mathbf{u}\|_{L^2}$$

is in fact a norm, equivalent to the norm in $H^1(\Omega)^N$. For simplicity, we put $\tilde{V}$ and $\tilde{H}$ instead of $\tilde{V}(\Omega)$ and $\tilde{H}(\Omega)$, resp.

Let $T > 0$ be given. Consider the following Navier–Stokes problem in $Q_T = \Omega \times (0,T)$, where we impose nonzero Dirichlet data:

$$(1)\qquad \left\{\begin{array}{rl} \dfrac{\partial \mathbf{y}}{\partial t} + (\mathbf{y}\cdot\nabla)\mathbf{y} - \nu\Delta\mathbf{y} + \nabla\pi = 0, & \nabla\cdot\mathbf{y} = 0 \text{ in } Q_T, \\ \mathbf{y} = \mathbf{v} & \text{on } \Lambda_T = \gamma \times (0,T), \\ \mathbf{y} = 0 & \text{on } S_T = (\partial\Omega\backslash\gamma) \times (0,T), \\ \mathbf{y}(0) = 0 & \text{in } \Omega. \end{array}\right.$$

Here, ν is the kinematic viscosity ($\nu > 0$) and $\mathbf{v} \in L^2(0,T;H^{-1/2}(\gamma)^N)$.

THEOREM 1.1. (a) *Assume* $\mathbf{v} = \mathbf{curl}\,\zeta\,|_\gamma$, *with*

$$(2)\qquad \begin{array}{ll} \boldsymbol{\zeta} \in L^2(0,T;H^2(\Omega)^M), & \dfrac{\partial\boldsymbol{\zeta}}{\partial t} \in L^2(0,T;H^1(\Omega)^M), \\ \boldsymbol{\zeta} \in L^\infty(0,T;W^{1,p}(\Omega)^3) & \textit{for some } p > 3 \textit{ if } N = 3, \\ \boldsymbol{\zeta} = \zeta \in L^\infty(0,T;W^{1,p}(\Omega)) & \textit{for some } p > 2 \textit{ if } N = 2, \\ \mathbf{v}(0)\cdot\mathbf{n} = 0 & \textit{in } H^{-1/2}(\gamma)^N \end{array}$$

* Received by the editors August 13, 1993; accepted for publication (in revised form) March 7, 1994.

[†] Departamento de Ecuaciones Diferenciales y Análisis Numérico, University of Sevilla, Tarfia s/n, E-41012 Sevilla, Spain.

(here, $\mathbf{n}$ *is the unit outward normal vector on* $\partial\Omega$*;* $M = 1$ *if* $N = 2$ *and* $M = 3$ *if* $N = 3$*). Then,* (1) *possesses at least one weak solution* $(\mathbf{y}_{\mathbf{v}}, \pi_{\mathbf{v}})$*. One has*

$$\mathbf{y}_{\mathbf{v}} \in L^2(0,T;\tilde{V}) \cap L^\infty(0,T;\tilde{H}),$$

$$\frac{\partial \mathbf{y}_{\mathbf{v}}}{\partial t} \in L^\sigma(0,T;H^{-1}(\Omega)^N) \qquad (\sigma = 2 \textit{ if } N = 2 \textit{ and } \sigma = 4/3 \textit{ if } N = 3),$$

$$\mathbf{y}_{\mathbf{v}} \in C^0([0,T];L^2(\Omega)^N) \quad \textit{if } N = 2,$$

$$\pi_{\mathbf{v}} \in L^2(Q_T).$$

(b) *If* $N = 2$*, there exists at most one weak solution to* (1) *(of course,* $\pi_{\mathbf{v}}$ *is unique up to a constant).*

The proof of this result can be easily obtained arguing as in [8], [9], [12]. Now, for each $\mathbf{v} \in L^2(0,T;H^{1/2}(\gamma)^N)$, let us set

$$\tilde{Y}_{\mathbf{v}}(T) = \{\mathbf{y}_{\mathbf{v}}(T); \mathbf{y}_{\mathbf{v}} \text{ solves, together with } \pi_{\mathbf{v}}, \text{ problem } (1)\}.$$

In this paper, we are concerned with the following problems.

PROBLEM (P). *Prove that the set*

$$\left(\bigcup_{\mathbf{v}} \tilde{Y}_{\mathbf{v}}(T)\right) \cap \tilde{H}$$

is dense in $\tilde{H}$.

PROBLEM (Q). *Let* $\tilde{Z}$ *be the subspace of* $\tilde{H}$ *spanned by*

$$\left(\bigcup_{\mathbf{v}} \tilde{Y}_{\mathbf{v}}(T)\right) \cap \tilde{H}.$$

Prove that $\tilde{Z}$ *is dense in* $\tilde{H}$.

Problem (P) is an approximate controllability problem in the sense of [10]. It admits the following physical interpretation: assume (for instance) that $\Omega = \mathcal{O}\backslash\bar{\Delta}$, where $\mathcal{O}$ and Δ are bounded and simply connected open sets. Also, assume that $\gamma = \partial\Delta$. If Problem (P) is solved, then a viscous incompressible fluid in $\mathcal{O}\backslash\bar{\Delta}$ that is initially at rest can be conduced to a mechanical state arbitrarily close to a given desired field acting exclusively on $\partial\Delta$.

Unfortunately, we are not able to solve Problem (P); instead, we solve Problem (Q) in this paper (see Theorem 1.2 below). Of course, the former is a much more interesting question. However, it must be noticed that in a similar linear situation Problems (P) and (Q) are equivalent. This happens, for instance, with (1) being replaced by the Stokes problem; thus, arguing as in the proof of Theorem 1.2, we obtain approximate controllability in this case (and this no matter how small γ is!).

On the other hand, recall that in the Navier–Stokes case not much is known on the nature of the set formed by all final states $\mathbf{y}_{\mathbf{v}}(T)$. In particular, it is not clear at all whether this set is very different from its linear span $\tilde{Z}$. In our opinion, this suffices to justify an analysis of Problem (Q).

Let us denote by U_{ad} the family of all admissible control functions:

$$U_{\text{ad}} = \{\mathbf{v};\ \mathbf{v} \in L^2(0,T;H^{1/2}(\gamma)^N),\ \exists \text{ solution } (\mathbf{y}_{\mathbf{v}}, \pi_{\mathbf{v}}) \text{ to } (1)\}.$$

The main result in this paper is as follows.

THEOREM 1.2. (a) *Assume $N = 2$ and let $\tilde{Y}$ be the subspace of $\tilde{H}$ spanned by the set*

$$\{\mathbf{y}_{\mathbf{v}}(T); \mathbf{v} \in U_{\text{ad}}\}.$$

Then $\tilde{Y}$ is dense in $\tilde{H}$.

(b) *Assume $N = 3$ and let $\tilde{Z}$ be the subspace of $\tilde{H}$ spanned by*

$$\left(\bigcup_{\mathbf{v}} \tilde{Y}_{\mathbf{v}}(T)\right) \cap \tilde{H}.$$

Then $\tilde{Z}$ is dense in $\tilde{H}$.

Theorem 1.2 is related to a conjecture formulated by Lions in [11]. In this reference, one is also concerned with approximate controllability, but there one imposes vanishing Dirichlet conditions on the whole $\partial\Omega \times (0, T)$ and one introduces L^2 control functions in the right side of the Navier–Stokes equations. In what follows this will be referred to as the distributed control variant of Problem (P). Bardos and Tartar [1] have considered in their paper a similar question; this time, the control is exerted on the initial condition and boundary data and second members vanish. Our result is similar to that in [1] (for $N = 2$ and initial data control) and also to those in [4] and [5] (for distributed control). See also [6] and the references therein for some related questions.

2. Some technical lemmas. Before we give the proof of Theorem 1.2, we present some technical results. First, we establish existence and regularity for the stationary Stokes problem with boundary conditions of different kinds on γ and on $\partial\Omega\backslash\gamma$ (recall that $\partial\Omega$ is a $C^{1,1}$ boundary and γ is a component of $\partial\Omega$). Let $\mathbf{f} \in L^2(\Omega)^N$, $g \in L^2(\Omega)$, and $\mathbf{b} \in H^{-1/2}(\gamma)^N$ be given and consider the following problem:

$$-\nu\Delta\mathbf{y} + \nabla\pi = \mathbf{f}, \qquad \nabla\cdot\mathbf{y} = g \quad \text{in}\,\Omega, \tag{3}$$

$$(-\pi\mathbf{Id} + \nu\nabla\mathbf{y})\cdot\mathbf{n} = \mathbf{b} \quad \text{on}\,\gamma, \tag{4}$$

$$\mathbf{y} = 0 \quad \text{on}\,\partial\Omega\backslash\gamma. \tag{5}$$

LEMMA 2.1. *There exists one and only one solution to* (3)–(5), $(\mathbf{y}, \pi) \in \tilde{V} \times L^2(\Omega)$. *For this couple,* (3) *is satisfied almost everywhere* (*a.e.*) *in* Ω, (4) *is satisfied as an equality in* $H^{-1/2}(\gamma)^N$, *and* (5) *is satisfied in the sense of the trace on* $\partial\Omega\backslash\gamma$. *Finally, there exists a constant* $C > 0$, *only depending on* Ω *and* γ, *such that*

$$\|\mathbf{y}\|_{H^1} + \|\pi\|_{L^2} \le C(\|\mathbf{f}\|_{L^2} + \|g\|_{L^2} + \|\mathbf{b}\|_{H^{-1/2}}).$$

The proof of this lemma can be achieved by means of well-known arguments. One also has the following.

LEMMA 2.2. *Let $m \ge 0$ be an integer. If $\partial\Omega$ is $C^{m+1,1}$*, $\mathbf{f} \in H^m(\Omega)^N$, $g \in H^{m+1}(\Omega)$, *and* $\mathbf{b} \in H^{m+1/2}(\gamma)^N$, *then* $(\mathbf{y}, \pi) \in H^{m+2}(\Omega)^N \times H^{m+1}(\Omega)$. *Furthermore, there exists a constant* $C > 0$, *only depending on* Ω, γ, *and* m, *such that*

$$\|\mathbf{y}\|_{H^{m+2}} + \|\pi\|_{H^{m+1}} \le C(\|\mathbf{f}\|_{H^m} + \|g\|_{H^{m+1}} + \|\mathbf{b}\|_{H^{m+1/2}}).$$

The proof of this result is rather technical. For instance, when $m = 0$, it relies on adequate uniform bounds for the finite difference quotients

$$\frac{1}{h}(\mathbf{y}(x + he_i) - \mathbf{y}(x)) \quad \text{and} \quad \frac{1}{h}(\pi(x + he_i) - \pi(x))$$

in $H^1(\Omega)^N$ and $L^2(\Omega)$, resp. The details are given in [7] (see also [2] and the references therein for other related results).

LEMMA 2.3. *There exists a sequence* $\{\lambda_j\}$, *with*

$$0 < \lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_j \leq \lambda_{j+1} \leq \dots, \qquad \lambda_j \nearrow \infty,$$

and an orthonormal basis of $\tilde{H}$, *denoted* $\{\mathbf{w}_j\}$, *such that, for all* j, *one has*

$$\mathbf{w}_j \in C^\infty(\Omega)^N \cap H^2(\Omega)^N \cap \tilde{V}$$

and

$$\int_\Omega \nabla \mathbf{w}_j : \nabla \mathbf{v}\, dx = \lambda_j \int_\Omega \mathbf{w}_j \cdot \mathbf{v}\, dx \quad \forall \mathbf{v} \in \tilde{V}.$$

The function $\mathbf{w}_j$ *is, together with some* $q_j \in C^\infty(\Omega) \cap H^1(\Omega)$, *the unique solution to*

$$\begin{cases} -\Delta \mathbf{w}_j + \nabla q_j = \lambda_j \mathbf{w}_j, \qquad \nabla \cdot \mathbf{w}_j = 0 \quad in\, \Omega, \\ (-q_j\, \mathbf{Id} + \nabla \mathbf{w}_j) \cdot \mathbf{n} = 0 \quad on\, \gamma, \\ \mathbf{w}_j = 0 \quad on\, \partial\Omega \backslash \gamma, \\ \|\mathbf{w}_j\|_{L^2} = 1. \end{cases}$$

Of course, the proof of Lemma 2.3 relies on the fact that the embedding $\tilde{V} \hookrightarrow \tilde{H}$ is dense and compact (see [7] for the details).

DEFINITION 2.4. *We introduce the trilineal form* $\tilde{b}$ *on* $H^1(\Omega)^N$ *by putting*

$$\tilde{b}(\mathbf{u},\ \mathbf{v}, \mathbf{w}) \equiv \frac{1}{2}[((\mathbf{u} \cdot \nabla)\mathbf{v}, \mathbf{w}) - ((\mathbf{u} \cdot \nabla)\mathbf{w}, \mathbf{v})].$$

Here, $(\cdot, \cdot)$ *stands for the usual scalar product in* $L^2(\Omega)^N$. *We also introduce the bilinear operator* $\tilde{B} : \tilde{V} \times \tilde{V} \to \tilde{V}'$ *by putting*

$$\langle \tilde{B}(\mathbf{u},\ \mathbf{v}), \mathbf{w} \rangle = \tilde{b}(\mathbf{u}, \mathbf{v}, \mathbf{w}) \quad \forall \mathbf{u}, \mathbf{v}, \mathbf{w}\, \in \tilde{V}.$$

Now, $\langle \cdot, \cdot \rangle$ *stands for the duality pairing between* $\tilde{V}'$ *and* $\tilde{V}$.

Assume that $\mathbf{u}, \mathbf{v}, \mathbf{w} \in H^1(\Omega)^N$ and $\nabla \cdot \mathbf{u} = 0$ in Ω. Then

$$\tilde{b}(\mathbf{u}, \mathbf{v}, \mathbf{w}) = ((\mathbf{u} \cdot \nabla)\mathbf{v}, \mathbf{w}) - \frac{1}{2} \int_{\partial\Omega} (\mathbf{u} \cdot \mathbf{n})\mathbf{v} \cdot \mathbf{w}\, dS.$$

On the other hand,

$$\tilde{b}(\mathbf{u}, \mathbf{v}, \mathbf{v}) = 0 \quad \forall \mathbf{u}, \mathbf{v} \in H^1(\Omega)^N.$$

Finally, notice that if **u** and **v** belong to $L^2(0, T; \tilde{V}) \cap L^\infty(0, T; \tilde{H})$, then

$$\tilde{B}(\mathbf{u}, \mathbf{v}) \in L^\sigma(0, T; \tilde{V}'),$$

where σ is arbitrary in $[1, 2)$ if $N = 2$ and $\sigma = 4/3$ if $N = 3$.

3. The existence of a solution to a coupled nonlinear problem. In order to prove Theorem 1.1, it will be convenient to demonstrate an existence result for a certain nonlinear problem. More precisely, for each $\mathbf{w} \in \tilde{H}$, let us introduce the system

$$
(6) \qquad \begin{cases} \dfrac{\partial \mathbf{y}}{\partial t} + (\mathbf{y} \cdot \nabla)\mathbf{y} - \nu \Delta \mathbf{y} + \nabla \pi = 0, & \nabla \cdot \mathbf{y} = 0 \quad \text{in } Q_T, \\ -\dfrac{\partial \mathbf{q}}{\partial t} - (\mathbf{y} \cdot \nabla)\mathbf{q} - \nu \Delta \mathbf{q} + \nabla Q = 0, & \nabla \cdot \mathbf{q} = 0 \quad \text{in } Q_T, \\ (-\pi \, \mathbf{Id} + \nu \nabla \mathbf{y}) \cdot \mathbf{n} - \dfrac{1}{2}(\mathbf{y} \cdot \mathbf{n})\mathbf{y} = \mathbf{q} \quad \text{on } \Lambda_T, \\ (-Q \, \mathbf{Id} + \nu \nabla \mathbf{q}) \cdot \mathbf{n} + \dfrac{1}{2}(\mathbf{y} \cdot \mathbf{n})\mathbf{q} = 0 \quad \text{on } \Lambda_T, \\ \mathbf{y} = \mathbf{q} = 0 \quad \text{on } S_T, \\ \mathbf{y}(0) = 0, \quad \mathbf{q}(T) = \mathbf{w} \quad \text{in } \Omega. \end{cases}
$$

Then one has the following theorem.

THEOREM 3.1. *If* $\mathbf{w} \in \tilde{H}$, *then the corresponding problem* (6) *possesses at least one weak solution* $(\mathbf{y}, \pi, \mathbf{q}, Q)$ *also satisfying*:

$$
(7) \qquad \begin{cases} \mathbf{y}, \mathbf{q} \in L^2(0,T;\tilde{V}) \cap L^\infty(0,T;\tilde{H}), & \dfrac{\partial \mathbf{y}}{\partial t}, \dfrac{\partial \mathbf{q}}{\partial t} \in L^\sigma(0,T;\tilde{V}'), \\ \mathbf{y}, \mathbf{q} \in C^0([0,T];\tilde{V}') \cap C^0_w([0,T];\tilde{H}), & \pi, Q \in L^2(Q_T), \end{cases}
$$

(*again*, σ *is arbitrary in* $[1,2)$ *if* $N = 2$ *and* $\sigma = 4/3$ *if* $N = 3$). *Moreover*, $\mathbf{y}$ *satisfies the energy inequalities*

$$
(8) \qquad \|\mathbf{y}(t)\|^2_{L^2} + 2\nu \int_0^t \|\nabla \mathbf{y}(s)\|^2_{L^2} \, ds \le 2 \int_0^t \int_\gamma \mathbf{q}(s) \cdot \mathbf{y}(s) dS \, ds
$$

and one has

$$
(9) \qquad (\mathbf{y}(T), \mathbf{w}) = \int \int_{\Lambda_T} |\mathbf{q}|^2 \, dS \, dt.
$$

Proof. Let us see that there exist functions

$$
\mathbf{y}, \mathbf{q} \in L^2(0,T;\tilde{V}) \cap L^\infty(0,T;\tilde{H}),
$$

which solve the weak formulation of (6), i.e., such that

$$
(10) \qquad \begin{cases} \left\langle \dfrac{\partial \mathbf{y}}{\partial t}, \mathbf{v} \right\rangle + \tilde{b}(\mathbf{y}, \mathbf{y}, \mathbf{v}) + \nu(\nabla \mathbf{y}, \nabla \mathbf{v}) = \int_\gamma \mathbf{q}(t) \cdot \mathbf{v} \, dS \quad \forall \mathbf{v} \in \tilde{V}, \\ -\left\langle \dfrac{\partial \mathbf{q}}{\partial t}, \mathbf{v} \right\rangle - \tilde{b}(\mathbf{y}, \mathbf{q}, \mathbf{v}) + \nu(\nabla \mathbf{q}, \nabla \mathbf{v}) = 0 \quad \forall \mathbf{v} \in \tilde{V}, \\ \mathbf{y}(0) = 0, \qquad \mathbf{q}(T) = \mathbf{w}. \end{cases}
$$

The proof consists of three steps.

First step: The existence of approximate solutions. We use the orthonormal basis furnished by Lemma 2.3. We denote by $\tilde{V}_m$ the linear space spanned by $\mathbf{w}_1, \mathbf{w}_2, \dots, \mathbf{w}_m$ and we put

$$\mathbf{w}_{0m} = \sum_{j=1}^{m} (\mathbf{w}, \mathbf{w}_j)\mathbf{w}_j,$$

i.e., $\mathbf{w}_{0m}$ is the orthogonal projection of $\mathbf{w}$ on $\tilde{V}_m$. For each $m \geq 1$, we search for functions

$$\mathbf{y}_m, \mathbf{q}_m \in C^0([0,T]; \tilde{V}_m)$$

such that

$$\text{(11)} \qquad \begin{cases} (\mathbf{y}'_m, \mathbf{w}_j) + \tilde{b}(\mathbf{y}_m, \mathbf{y}_m, \mathbf{w}_j) + \nu(\nabla \mathbf{y}_m, \nabla \mathbf{w}_j) = \int_\gamma \mathbf{q}_m \cdot \mathbf{w}_j \, dS \\ (1 \leq j \leq m), \qquad \mathbf{y}_m(0) = 0, \end{cases}$$

$$\text{(12)} \qquad \begin{cases} -(\mathbf{q}'_m, \mathbf{w}_j) - \tilde{b}(\mathbf{y}_m, \mathbf{q}_m, \mathbf{w}_j) + \nu(\nabla \mathbf{q}_m, \nabla \mathbf{w}_j) = 0 \\ (1 \leq j \leq m), \qquad \mathbf{q}_m(T) = \mathbf{w}_{0m}. \end{cases}$$

We argue as follows. If the function $\mathbf{p}_m$ is given in $C^0([0,T]; \tilde{V}_m)$, there exists exactly one maximal (in time) solution $\mathbf{y}_m = \mathbf{y}_m(\mathbf{p}_m)$ to the ordinary differential problem (11) with $\mathbf{q}_m = \mathbf{p}_m$. It is not difficult to check that

$$\frac{1}{2}\frac{d}{dt}\|\mathbf{y}_m(t)\|_{L^2}^2 + \nu\|\nabla \mathbf{y}_m(t)\|_{L^2}^2 = \int_\gamma \mathbf{p}_m(t) \cdot \mathbf{y}_m(t) dS.$$

Hence,

$$\|\mathbf{y}_m(t)\|_{L^2}^2 + \nu \int_0^t \|\nabla \mathbf{y}_m(s)\|_{L^2}^2 \, ds \leq C \int\int_{\Lambda_T} |\mathbf{p}_m(t)|^2 \, dS \, dt$$

for some C only depending on Ω, γ, and ν. From this inequality, we deduce that $\mathbf{y}_m$ is defined for all $t \in [0,T]$. Now, let us denote by $\mathbf{q}_m = \mathbf{q}_m(\mathbf{y}_m)$ the unique maximal solution to (12). It is clear that $\mathbf{q}_m$ is also defined for all $t \in [0,T]$. Moreover,

$$-\frac{1}{2}\frac{d}{dt}\|\mathbf{q}_m(t)\|_{L^2}^2 + \nu\|\nabla \mathbf{q}_m(t)\|_{L^2}^2 \equiv 0,$$

whence

$$\|\mathbf{q}_m(t)\|_{L^2}^2 + 2\nu \int_t^T \|\nabla \mathbf{q}_m(s)\|_{L^2}^2 \, ds = \|\mathbf{w}_{0m}\|_{L^2}^2 \leq \|\mathbf{w}\|_{L^2}^2.$$

This proves that $\mathbf{q}_m$ is bounded in $C^0([0,T]; \tilde{V}_m)$ independently from $\mathbf{y}_m$. Let W be the ball $\bar{B}(0; \|\mathbf{w}\|_{L^2}^2)$ in $C^0([0,T]; \tilde{V}_m)$ and let Φ be given as follows:

$$\Phi(\mathbf{p}_m) = \mathbf{q}_m(\mathbf{y}_m(\mathbf{p}_m)) \quad \forall \mathbf{p}_m \in W.$$

Then $\Phi : W \to W$ is a continuous compact mapping (due to the fact that $\Phi(\mathbf{p}_m) \in C^1([0,T]; \tilde{V}_m)$ for each $\mathbf{p}_m$). Consequently, Schauders' theorem applies and Φ possesses a fixed point $\mathbf{q}_m \in W$. Obviously, $\mathbf{q}_m$ and $\mathbf{y}_m = \mathbf{y}_m(\mathbf{q}_m)$ satisfy (11) and (12).

Second step: "A priori" estimates. From (11) and (12), one easily obtains

$$\mathbf{y}_m, \mathbf{q}_m \in \text{bounded set in } L^2(0,T;\tilde{V}) \cap L^\infty(0,T;\tilde{H}). \tag{13}$$

Consequently,

$$\tilde{B}(\mathbf{y}_m, \mathbf{y}_m), \tilde{B}(\mathbf{y}_m, \mathbf{q}_m) \in \text{bounded set in } L^\sigma(0,T;\tilde{V}'),$$

with σ being as before. Now, the choice of the basis $\{\mathbf{w}_j\}$ yields

$$\mathbf{y}'_m, \mathbf{q}'_m \in \text{bounded set in } L^\sigma(0,T;\tilde{V}'). \tag{14}$$

On the other hand, from (11) and (12), one easily deduces that

$$(\mathbf{y}_m(T), \mathbf{w}_{0m}) = \int\int_{\Lambda_T} |\mathbf{q}_m|^2 \, dS \, dt. \tag{15}$$

Third step: The choice of a convergent sequence — conclusion. From (13) and (14), one deduces that functions $\mathbf{y}$ and $\mathbf{q}$ and subsequences $\{\mathbf{y}_\rho\}$ and $\{\mathbf{q}_\rho\}$ must exist with

$$\mathbf{y}, \mathbf{q} \in L^2(0,T;\tilde{V}) \cap L^\infty(0,T;\tilde{H}) \cap C^0([0,T];\tilde{V}'),$$

$$\frac{\partial \mathbf{y}}{\partial t}, \frac{\partial \mathbf{q}}{\partial t} \in L^\sigma(0,T;\tilde{V}'),$$

and

$$\begin{cases} \mathbf{y}_\rho \,(\text{resp., } \mathbf{q}_\rho) \rightharpoonup \mathbf{y} \,(\text{resp., } \mathbf{q}) \text{ weakly in } L^2(0,T;\tilde{V}), \\ \mathbf{y}_\rho \,(\text{resp., } \mathbf{q}_\rho) \rightharpoonup \mathbf{y} \,(\text{resp., } \mathbf{q}) \text{ weakly} * \text{ in } L^\infty(0,T;\tilde{H}), \\ \mathbf{y}_\rho \,(\text{resp., } \mathbf{q}_\rho) \to \mathbf{y} \,(\text{resp., } \mathbf{q}) \text{ strongly in } L^2(0,T;\tilde{V}_s), \\ \dfrac{\partial \mathbf{y}_\rho}{\partial t} \left(\text{resp., } \dfrac{\partial \mathbf{q}_\rho}{\partial t}\right) \rightharpoonup \dfrac{\partial \mathbf{y}}{\partial t} \left(\text{resp., } \dfrac{\partial \mathbf{q}}{\partial t}\right) \text{ weakly in } L^\sigma(0,T;\tilde{V}'). \end{cases}$$

Here, $1/2 < s < 1$ and $\tilde{V}_s$ stands for the closure of $\tilde{V}$ with respect to the norm in $H^s(\Omega)^N$ (a new Hilbert space for the same norm). These convergence properties allow us to take limits in (11) and (12), which proves that $\mathbf{y}$ and $\mathbf{q}$ solve (10). Obviously, (8) is satisfied; on the other hand, from (15) and the previous properties, it is easy to deduce (9). This ends the proof of Theorem 3.1.

4. The proof of the main result.

From a well-known consequence of the Hahn–Banach theorem (for instance, see [3, Cor. I.8]), we know that the following is a statement equivalent to Theorem 1.2.

THEOREM 4.1. *Assume* $\mathbf{w} \in \tilde{H}$ *satisfies*

$$\begin{gathered} (\mathbf{y}_\mathbf{v}(T), \mathbf{w}) = 0 \quad \forall \mathbf{v} \in U_{\text{ad}} \quad \text{if } N = 2, \\ (\mathbf{y}_\mathbf{v}(T), \mathbf{w}) = 0 \quad \forall \mathbf{v} \in \left(\bigcup_{\mathbf{v} \in U_{\text{ad}}} \tilde{Y}_\mathbf{v}(T) \right) \cap \tilde{H} \quad \text{if } N = 3. \end{gathered} \tag{16}$$

Then $\mathbf{w} = 0$.

Proof. Let $\mathbf{w} \in \tilde{H}$ be given and assume that (16) is satisfied. Let $(\mathbf{y}^*, \pi^*, \mathbf{q}^*, Q^*)$ be the weak solution to (6) furnished by Theorem 3.1. Recall that $(\mathbf{y}^*, \pi^*, \mathbf{q}^*, Q^*)$ satisfies (7)–(9).

Let $\mathbf{v}$ be the trace of $\mathbf{y}^*$ on $\Lambda_T = \gamma \times (0,T)$. Then $\mathbf{v} \in U_{\text{ad}}$ and, moreover, the couple $(\mathbf{y}^*, \pi^*)$ is a state associated to $\mathbf{v}$. Accordingly, taking into account (9) and (16), one has

$$\mathbf{q}^* = 0 \quad \text{on } \Lambda_T. \tag{17}$$

From (8), we also deduce that $\mathbf{y}^* \equiv 0$. Thus, we have found a function $\mathbf{q}^*$ that vanishes on Λ_T and solves, together with Q^*, the following final value-boundary value problem:

$$-\frac{\partial \mathbf{q}}{\partial t} - \nu \Delta \mathbf{q} + \nabla Q = 0, \quad \nabla \cdot \mathbf{q} = 0 \quad \text{in } Q_T, \tag{18}$$

$$(-Q\,\mathbf{Id} + \nu \nabla \mathbf{q}) \cdot \mathbf{n} = 0 \quad \text{on } \Lambda_T, \tag{19}$$

$$\mathbf{q} = 0 \quad \text{on } S_T, \tag{20}$$

$$\mathbf{q}(T) = \mathbf{w} \quad \text{in } \Omega. \tag{21}$$

It is not difficult to prove that (18)–(21) possesses exactly one solution pair $(\mathbf{q}, Q)$, with (at least)

$$\mathbf{q} \in L^2(0,T;\tilde{V}) \cap C^0([0,T];\tilde{H}),$$

$$\frac{\partial \mathbf{q}}{\partial t} \in L^2(0,T;\tilde{V}'), \qquad Q \in L^2(Q_T).$$

Necessarily, $(\mathbf{q}, Q) = (\mathbf{q}^*, Q^*)$. Consequently, Theorem 4.1 is implied by Proposition 4.2 (see below).

PROPOSITION 4.2. *Assume the couple* $(\mathbf{q}^*, Q^*)$ *satisfies*

$$\mathbf{q}^* \in L^2_{\text{loc}}(0,T;\tilde{V}) \cap L^\infty_{\text{loc}}(0,T;\tilde{H}),$$

$$\frac{\partial \mathbf{q}^*}{\partial t} \in L^2_{\text{loc}}(0,T;\tilde{V}'), \qquad Q^* \in L^2_{\text{loc}}(0,T;L^2(\Omega))$$

and (17)–(20). *Then* $\mathbf{q}^* \equiv 0$.

Proof. Let $x_0 \in \gamma$ be given. Choose $r > 0$ such that

$$B(x_0;r) \cap \partial\Omega \subset \gamma$$

and consider the open sets $\omega = B(x_0;r)$ and $\tilde{\Omega} = \Omega \cup \omega$. Let $(\tilde{\mathbf{q}}, \tilde{Q})$ be the extension by zero of $(\mathbf{q}^*, Q^*)$ to the whole cylinder $\tilde{\Omega} \times (0,T)$. From (17), we see that

$$\tilde{\mathbf{q}} \in L^2_{\text{loc}}(0,T;V(\tilde{\Omega})) \cap L^\infty_{\text{loc}}(0,T;H(\tilde{\Omega})),$$

$$\tilde{Q} \in L^2_{\text{loc}}(\tilde{\Omega} \times (0,T)).$$

Here,

$$V(\tilde{\Omega}) = \{\mathbf{v}; \mathbf{v} \in H^1_0(\tilde{\Omega})^N,\ \nabla \cdot \mathbf{v} = 0 \text{ in } \tilde{\Omega}\},$$

$$H(\tilde{\Omega}) = \{\mathbf{v}; \mathbf{v} \in L^2(\tilde{\Omega})^N, \ \nabla \cdot \mathbf{v} = 0 \,\text{in}\, \tilde{\Omega}, \ \mathbf{v} \cdot \mathbf{n} = 0 \,\text{on}\, \partial\tilde{\Omega}\}.$$

$V(\tilde{\Omega})$ and $H(\tilde{\Omega})$ are endowed with the norms of $H^1(\tilde{\Omega})^N$ and $L^2(\tilde{\Omega})^N$, resp. It is easy to check that

$$\frac{\partial \tilde{\mathbf{q}}}{\partial t} \in L^2_{\text{loc}}(0,T; V(\tilde{\Omega})')$$

and

$$\begin{cases} -\dfrac{\partial \tilde{\mathbf{q}}}{\partial t} - \nu\Delta\tilde{\mathbf{q}} + \nabla\tilde{Q} = 0, \quad \nabla \cdot \tilde{\mathbf{q}} = 0 \quad \text{in}\, \tilde{\Omega} \times (0,T), \\ \tilde{\mathbf{q}} = 0 \quad \text{on}\, \partial\tilde{\Omega} \times (0,T). \end{cases}$$

In particular, we deduce that both $\tilde{\mathbf{q}}$ and $\tilde{Q}$ are analytical functions in the space in $\tilde{\Omega} \times (0,T)$ (cf. e.g. [8]). But $\tilde{\mathbf{q}} = 0$ in $(\tilde{\Omega}\backslash\bar{\Omega}) \times (0,T)$. Hence, necessarily $\tilde{\mathbf{q}} \equiv 0$.

For the sake of completeness, let us state (and prove) a regularity result for (18)–(21).

LEMMA 4.3. *Let* $\mathbf{w} \in \tilde{H}$ *and* $\delta > 0$ *be given. Then the unique solution* $(\mathbf{q}, Q)$ *to* (18)–(21) *satisfies*

$$\mathbf{q} \in L^2(0,T-\delta; H^2(\Omega)^N) \cap L^\infty(0,T-\delta; \tilde{V}) \cap L^2(0,T;\tilde{V}) \cap C^0([0,T];\tilde{H}),$$

$$\frac{\partial \mathbf{q}}{\partial t} \in L^2(0,T-\delta; H^1(\Omega)^N) \cap L^\infty(0,T-\delta; L^2(\Omega)^N) \cap L^2(0,T;\tilde{V}'),$$

$$Q \in L^2(0,T-\delta; H^1(\Omega)) \cap L^\infty(0,T-\delta; L^2(\Omega)) \cap L^2(Q_T).$$

Sketch of the proof. Let $\theta = \theta_\delta$ be a real-valued C^∞ function on $[0,+\infty)$ such that

$$\theta \equiv 1 \,\text{in}\, [0,T-\delta), \qquad \theta \equiv 0 \,\text{in}\, \left[T - \frac{\delta}{2}, +\infty\right).$$

Using θ, we introduce

$$\hat{\mathbf{q}} = \theta\mathbf{q} \quad \text{and} \quad \hat{Q} = \theta Q.$$

Then $\hat{\mathbf{q}} \in L^2(0,T;\tilde{V}) \cap L^\infty(0,T;\tilde{H})$ and, also,

$$\begin{cases} -\left\langle \dfrac{\partial \hat{\mathbf{q}}}{\partial t}(t), \mathbf{v} \right\rangle + \nu(\nabla\hat{\mathbf{q}}(t), \nabla\mathbf{v})_{0;\Omega} = (\mathbf{f}(t), \mathbf{v})_{0;\Omega} \quad \forall \mathbf{v} \in \tilde{V}, \ t \in (0,T) \,\text{a.e.}, \\ \hat{\mathbf{q}}(T) = 0, \end{cases}$$

where $\mathbf{f} = -\theta'\mathbf{q}$. Notice that

$$\mathbf{f} \in L^2(0,T;\tilde{V}) \cap L^\infty(0,T;\tilde{H}) \quad \text{and} \quad \frac{\partial \mathbf{f}}{\partial t} \in L^2(0,T;\tilde{V}').$$

It is clear that $\hat{\mathbf{q}}$ is the limit of approximate solutions $\hat{\mathbf{q}}_m$, with

$$\hat{\mathbf{q}}_m(t) = \sum_{j=1}^{m} \hat{q}^j_m(t)\mathbf{w}_j,$$

$$(22)\qquad \begin{cases} -(\hat{\mathbf{q}}'_m(t),\mathbf{w}_j)_{0;\Omega} + \nu(\nabla\hat{\mathbf{q}}_m(t),\,\nabla\mathbf{w}_j)_{0;\Omega} = (\mathbf{f}(t),\mathbf{w}_j)_{0;\Omega}, \\ (1\le j\le m), \qquad \hat{\mathbf{q}}_m(T)=0. \end{cases}$$

Differentiation with respect to t leads to the equalities

$$(23)\qquad -(\hat{\mathbf{q}}''_m(t),\mathbf{w}_j)_{0;\Omega} + \nu(\nabla\hat{\mathbf{q}}'_m(t),\nabla\mathbf{w}_j)_{0;\Omega} = \langle\mathbf{f}',\mathbf{w}_j\rangle.$$

Now, multiplying the jth equation in (22) by $\lambda_j\hat{q}^j_m(t)$, adding for $1\le j\le m$, and integrating with respect to t, we are led to the inequalities

$$\|\nabla\hat{\mathbf{q}}_m(t)\|^2_{0;\Omega} + \nu\int_t^T \|\Delta\hat{\mathbf{q}}'_m(s)\|^2_{0;\Omega}\,ds \le C\int_0^T \|\mathbf{f}(t)\|^2_{0;\Omega}\,dt,$$

where C is a constant. This proves that

$$\hat{\mathbf{q}}\in L^2(0,T;H^2(\Omega)^N)\cap L^\infty(0,T;\tilde{V}).$$

On the other hand, multiplying (23) by $(\hat{q}^j_m)'(t)$ and adding for $1\le j\le m$, we obtain

$$-\frac{1}{2}\frac{d}{dt}\|\hat{\mathbf{q}}'_m(t)\|^2_{0;\Omega} + \nu\|\nabla\hat{\mathbf{q}}'_m(t)\|^2_{0;\Omega} = -\left\langle\frac{\partial\mathbf{f}}{\partial t}(t),\hat{\mathbf{q}}'_m(t)\right\rangle \quad \forall t\in[0,T].$$

After integration with respect to t, one has

$$\|\hat{\mathbf{q}}'_m(t)\|^2_{0;\Omega} + \nu\int_t^T \|\delta\hat{\mathbf{q}}'_m(s)\|^2_{0;\Omega}\,ds \le C\left\|\frac{\partial\mathbf{f}}{\partial t}\right\|^2_*,$$

where $\|\cdot\|_*$ stands for the norm in $L^2(0,T;\tilde{V}')$. Hence,

$$\frac{\partial\hat{\mathbf{q}}}{\partial t}\in L^2(0,T;H^1(\Omega)^N)\cap L^\infty(0,T;L^2(\Omega)^N).$$

This proves the lemma.

REFERENCES

[1] C. Bardos and L. Tartar, *Sur l'unicité rétrograde des équations paraboliques et quelques questions voisines*, Arch. Rat. Mech. Anal., 50 (1973), pp. 10–25.

[2] J. A. Bello, Diferenciación respecto de dominios, thesis, University of Sevilla, Spain, 1992.

[3] H. Brézis, *Analyse Fonctionnelle. Théorie et Applications*, Masson, Paris, 1983.

[4] E. Fernández-Cara and J. Real, *On a conjecture due to J.L. Lions concerning weak controllability for Navier–Stokes Flows*, in Proceedings of the EQUADIFF'91 Conference, C. Perelló, C. Simó, and J. Solà-Morales, eds., World Scientific, Barcelona, 1993.

[5] ———, *On a conjecture due to J.L. Lions*, Nonlinear Anal. T.M.A., 21 (1993), pp. 835–847.

[6] A. V. Fursikov, *Properties of the solutions of some control problems connected with the Navier–Stokes system*, Soviet Math. Dokl., 25 (1982), pp. 40–45.

[7] M. González-Burgos, Dos problemas relacionados con E. D. P. de evolución no lineale, thesis, University of Sevilla, Spain, 1993.

[8] O. A. Ladyzhenskaya, *The Mathematical Theory of Viscous Incompressible Flow*, 2nd ed., Gordon and Breach, New York, 1969.

[9] J. L. Lions, *Quelques Méthodes de Résolution des Problèmes aux Limites Non Linéaires*, Dunod, Gauthiers-Villars, Paris, 1969.

[10] ———, *Exact controllability, stabilization and perturbations for distributed systems*, SIAM Rev., 30 (1988), pp. 1–68.

[11] ———, *Remarques sur la contrôlabilité approchée*, in Actas de las Jornadas Hispano-Francesas sobre Control de Sistemas Distribuidos, University of Málaga, Spain, 1990.

[12] R. Témam, *Navier–Stokes Equations*, North-Holland, Amsterdam, 1977.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 33, No. 4, pp. 1071–1085, July 1995

© 1995 Society for Industrial and Applied Mathematics
007

SMOOTHLY GLOBAL STABILIZABILITY BY DYNAMIC FEEDBACK AND GENERALIZATIONS OF ARTSTEIN'S THEOREM*

JOHN TSINIAS†

Abstract. The purpose of this paper is to explore the dynamic feedback stabilization problem for general nonlinear systems extending previous works of Artstein, Sontag, and the author. First, sufficient conditions for global stabilization by smooth dynamic feedback are provided concerning a wide class of triangular systems, where the linearization at the equilibrium is controllable. For the general nonlinear case, we derive a necessary and sufficient condition for global stabilization by means of an almost smooth dynamic feedback.

Key words. global, dynamic stabilizability, smooth, almost smooth feedback

AMS subject classification. 93D15

1. Introduction.

We consider general nonlinear single-input systems of the form

$$\dot{x} = f(x,u), \qquad (x,u) \in \mathbb{R}^n \times \mathbb{R}, \tag{1.1}$$

where the map $f : \mathbb{R}^{n+1} \to \mathbb{R}$ is continuous vanishing at zero.

We say that (1.1) is static asymptotically stabilizable (S.A.S.) at the origin if there exists an ordinary state feedback $u = u(x)$, vanishing at zero, which is smooth (C^∞) on $\mathbb{R}^n \backslash \{0\}$, such that $0 \in \mathbb{R}^n$ is asymptotically stable with respect to the closed-loop system $\dot{x} = f(x, u(x))$. We say that (1.1) is dynamic asymptotically stabilizable (D.A.S.) at the origin if the extended $(n+1)$-dimensional affine in the control system

$$\dot{x} = f(x,y); \qquad \dot{y} = v \tag{1.2}$$

is S.A.S. Finally, (1.1) is said to be smoothly S.A.S. (D.A.S.) when the corresponding feedback stabilizers are smooth on the whole space $\mathbb{R}^n$ ($\mathbb{R}^{n+1}$, respectively).

Our purpose is to analyze the feedback stabilzability problem in terms of control Lyapunov functions (clf) extending previous works on the same problem (see for instance [1], [2], [5], [8], [11], [13] and references therein).

First, we derive sufficient conditions for smoothly dynamic global stabilization for general triangular systems of the form

$$\begin{gathered} \dot{x}_1 = f_1(x_1, x_2), \\ \dot{x}_2 = f_2(x_1, x_2, x_3), \\ \cdots\cdots \\ \dot{x}_n = f_n(x_1, x_2, x_3, \ldots, x_n, x_{n+1}), x_{n+1} \div u, \end{gathered} \tag{1.3}$$

where the mappings $f_i : \mathbb{R}^{i+1} \to \mathbb{R}$, $i = 1, \ldots, n$ are C^∞ vanishing at zero. A typical example of nonlinearizable systems having triangular structure are systems of the form $x^{(n)} = \phi(x, x^{(1)}, \ldots, x^{(n-1)}, u)$, $x \in \mathbb{R}$, where the map ϕ is purely nonlinear with respect to u. Necessary and sufficient conditions for a general system to be triangularizable by applying change of coordinates plus dynamic feedback are given in [10].

The following theorem is one of the main results of our paper extending previous works on the feedback stabilization for triangular systems (see for instance [6], [9], [12], [14]).

THEOREM 1.1. *Suppose that for each $1 \le i \le n$ the following conditions hold:*

* Received by the editors December 28, 1992; accepted for publication (in revised form) March 24, 1994.

† National Technical University, Department of Mathematics, Zografou Campus, 15780 Athens, Greece.

(i) *for every* $x_1, \dots, x_i$ *the map* $f_i(x_1, \dots, x_i, \cdot) : \mathbb{R} \to \mathbb{R}$ *is a surjection, namely,* $f_i(x_1, \dots, x_i, \mathbb{R}) = \mathbb{R}$. *Furthermore, assume that*

$$f_i(0, \dots, 0, \mathbb{R}^+) = \mathbb{R}^+(\mathbb{R}^-); \qquad f_i(0, \dots, 0, \mathbb{R}^-) = \mathbb{R}^-(\mathbb{R}^+), \tag{1.4a}$$

whereas

$$f_i(0, \dots, 0, x_{i+1}) = 0 \quad \textit{iff}\ x_{i+1} = 0; \tag{1.4b}$$

(ii) *for each compact set* $A \subset \mathbb{R}^i$ *and for every unbounded subset* B *of the real line, the set*

$$\{t \in \mathbb{R} : t = f_i(x_1, \dots, x_i; y), (x_1, \dots, x_i; y) \in A \times B\}$$

is unbounded;

(iii) $$\frac{\partial f_i}{\partial x_{i+1}}(0, \dots, 0, 0) \neq 0,$$

which means that the linearization of (1.3) *at the origin is controllable.*
Then the system (1.3) *is smoothly globally D.A.S. If in addition we assume that*

(iv) $$\frac{\partial f_n}{\partial u}(x, u) \neq 0, \quad \forall (x, u) \in \mathbb{R}^n \times \mathbb{R},$$

then (1.3) *is smoothly globally S.A.S.*

Remark 1.2. A particular case of systems (1.3) satisfying all conditions of Theorem 1.1 arises when we assume that each f_i has the form

$$f_i(x_1, \dots, x_i, x_{i+1}) = x_{i+1} + g_i(x_1, \dots, x_i), \tag{1.5}$$

where each g_i is smooth and independent of x_{i+1}. It is well known that every single-input system (1.1), which is affine in the control, can locally be transformed into (1.3) with dynamics (1.5), provided that Brockett's linearization conditions are satisfied (see [3], [10]). Furthermore, as it was pointed out in [10], [14], the system (1.3) with dynamics (1.5) is globally feedback equivalent to a controllable linear system. Since the latter is globally S.A.S. by means of a linear feedback, it follows that the original system is smoothly globally S.A.S. and therefore according to [13, Thm. 4] it is smoothly globally D.A.S. Theorem 1.1 of the present paper consists of an extension of the previous case for general triangular systems (1.3).

The proof of Theorem 1.1 is based on the following result providing sufficient conditions for dynamic stabilization for the general case (1.1).

THEOREM 1.3 [15]. *Suppose that there exist a closed subset* $M \subset \mathbb{R}^{n+1}$, *a pair of disjoint open subsets* $U^+, U^- \subset \mathbb{R}^{n+1}$, *and a* C^1 *positive definite function* $V : \mathbb{R}^n \to \mathbb{R}$ *such that*

(i) $$0 \in M, \quad \mathbb{R}^{n+1} = U^+ \cup U^- \cup M, \quad \pi(M) = \mathbb{R}^n,$$

where $\pi(M)$ *denotes the projection of* M *on* $\mathbb{R}^n$ *along the y axis*;

(ii) *the following conditions hold*:

$$\{0\} \times (0 + \infty) \subset U^+, \qquad \{0\} \times (-\infty, 0) \subset U^-, \tag{1.6}$$

and further for each compact set $Q \subset \mathbb{R}^n$ *the set* $S_Q \doteqdot \{(x, y) \in M,\ x \in Q\}$ *is compact*;

(iii) *for each nonzero* $(x, y) \in M$ *we have*

$$DV(x) f(x, y) < 0, \tag{1.7}$$

where DV denotes the derivative of V;

(iv) *V is uniformly unbounded on $\mathbb{R}^n$, i.e., $V(x) \to +\infty$ as $\|x\| \to +\infty$, where $\| \ \|$ denotes the usual Euclidean norm.*
Then there exists a C^1 map $\Phi(x, y)$ such that the function $L(x, y) \div V(x) + \Phi(x, y)$ is a global clf with respect to (1.2), *namely, L is C^1, positive definite, uniformly unbounded on $\mathbb{R}^{n+1}$, and satisfies*

$$((\partial L/\partial x)f)(x,y) < 0, \quad \forall\,(x,y) \neq 0 : (\partial L/\partial y)(x,y) = 0.$$

Hence, according to Artstein's theorem (1.1) *is globally D.A.S. The system* (1.1) *is locally D.A.S., if $U^+ \cup U^- \cup M$ and $\pi(M)$ are neighborhoods of* $0 \in \mathbb{R}^{n+1}$ *and* $0 \in \mathbb{R}^n$, *respectively, and* (1.6) *and condition* (iii) *are fulfilled.*

Remark 1.4. The well-known Artstein's theorem [2] asserts that the Lyapunov condition (1.7) is satisfied for each $x \neq 0$ and for some $y = y(x)$ depending on x, if and only if (1.1) is stabilizable by means of a relaxed static feedback. Stabilization by means of an ordinary static feedback, which is smooth on $\mathbb{R}^n \backslash \{0\}$, is in general feasible if we further assume that (1.1) is affine in the control. It should be noted that conditions (i)–(iv) are satisfied in the particular case where (1.1) is S.A.S. by means of a feedback law that is continuous at zero (see [15]). In that case the assumption $0 \in M$ is equivalent to the fact that the control function V satisfies the "small control property" (see [2], [11], [13]).

A second aim of the paper is to provide a generalization of Theorem 1.3. In particular, in Theorem 3.1 we provide a necessary and sufficient condition for global stabilization of (1.1) by means of a dynamic ordinary feedback that is smooth on $\mathbb{R}^{n+1} \backslash \{0\}$. The proof of Theorem 3.1, similar to that of Theorem 1.3, is based on the construction of an appropriate clf guaranteeing feedback stabilization. However, the weaker assumptions imposed in Theorem 3.1 do not permit following the same procedure as in [15] to construct the desired clf. A more careful analysis is required.

Finally, relationships between global controllability and conditions of Theorem 1.1, as well as remarks on global stabilization by means of a continuous static feedback for planar systems, are included in §§2.2 and 3, respectively.

2. The triangular case.

2.1. Smoothly global stabilization.

To prove Theorem 1.1, we first need the following elementary lemmas.

LEMMA 2.1. *Suppose that conditions* (i) *and* (ii) *of Theorem* 1.1 *are satisfied and without any loss of generality assume that $f_i(0, \mathbb{R}^+) = \mathbb{R}^+$ and $f_i(0, \mathbb{R}^-) = \mathbb{R}^-$ for every $i = 1, \ldots, n$. Let $s : \mathbb{R}^i \to \mathbb{R}$ be a continuous function vanishing at zero. Then conditions* (i) *and* (ii) *of Theorem* 1.3 *are satisfied with*

$$\begin{gathered}
x = w_i \div (x_1, \ldots, x_i),\ y \div x_{i+1}, \\
M \div \{(w_i, y) \in \mathbb{R}^{i+1} : f_i(w_i, y) = s(w_i)\}, \\
U^+ \div \{(w_i, y) \in \mathbb{R}^{i+1} : f_i(w_i, y) > s(w_i)\}; \\
U^- \div \{(w_i, y) \in \mathbb{R}^{i+1} : f_i(w_i, y) < s(w_i)\}.
\end{gathered}$$

Proof. Conditions (i) and (1.6) of Theorem 1.3 follow immediately by taking into account the definitions of M, U^+, and U^- and the fact that $f_i(0, \mathbb{R}^\pm) = \mathbb{R}^\pm$ and $f_i(w_i, \mathbb{R}) = \mathbb{R}$ for all w_i. The latter also asserts that M, U^+, and U^- are nonempty. Next we show that for each compact set $Q \subset \mathbb{R}^i$ the set S_Q, as defined in the statement of Theorem 1.3, is compact. Indeed, suppose on the contrary that there exists a sequence $(w_{i\nu}, y_\nu) \in M$, $w_{i\nu} \in Q$, $\nu = 1, 2, \ldots$, with $w_{i\nu} \to w \in Q$ and $|y_\nu| \to +\infty$. This, in conjunction with assumption (ii) of

Theorem 1.1, implies that $\overline{\lim}|f_i(w_{i\nu}, y_\nu)| = +\infty$. On the other hand, by the definition of M, $f_i(w_{i\nu}, y_\nu) = s(w_{i\nu})$ for all ν and so $\lim f_i(w_{i\nu}, y_\nu) = s(w)$, a contradiction. □

LEMMA 2.2. *In addition to the hypothesis of Lemma* 2.1, *assume that condition* (iii) *of Theorem* 1.1 *is satisfied and the functions f_i and s are C^∞ near zero. Then there is a constant $\rho > 0$ and a unique map $\phi_i : \mathbb{R}^i \to \mathbb{R}$ vanishing at zero that is C^∞ on the region $\{(w_i, y) \in \mathbb{R}^{i+1}, \|w_i\| \le \rho\}$, where $w_i \doteq (x_1, \dots, x_i)$ and $y \doteq x_{i+1}$, such that the intersection of this region with the set M as defined in Lemma* 2.1 *coincides with the graph of the mapping $y = \phi_i(w_i)$.*

Proof. Since f_i satisfies condition (iii) of Theorem 1.1 and both f_i and s are C^∞ vanishing at zero, by the implicit function theorem there exists a constant $\rho > 0$ and a unique smooth map $\phi_i : \mathbb{R}^i \to \mathbb{R}$ with $\phi_i(0) = 0$ such that $f_i(w_i, \phi_i(w_i)) = s(w_i)$ for $\|w_i\| \le \rho$. This implies that the graph of ϕ_i restricted to the sphere S_ρ of radius ρ centered at $0 \in \mathbb{R}^{i+1}$ is contained in M. To complete the proof, suppose on the contrary that there exists a sequence $(w_{i\nu}, y_\nu) \in M \backslash S_\rho$ with $w_{i\nu} \to 0$. Then $f_i(w_{i\nu}, y_\nu) = s(w_{i\nu})$, whereas by using assumption (ii) and the continuity of f_i and s we may assume that $\lim y_\nu = \alpha < +\infty$. Consequently, $f(0, \alpha) = 0$ with $|\alpha| \ge \rho$, which contradicts (1.4b). The latter implies the desired conclusion. □

Proof of Theorem 1.1. For reasons of simplicity we consider the case $n = 2$, namely, we show that the system

$$\begin{pmatrix} \dot{x}_1 \\ \dot{x}_2 \end{pmatrix} = \begin{pmatrix} f_1(x_1, x_2) \\ f_2(x_1, x_2, x_3) \end{pmatrix}, \qquad x_3 \doteq u \tag{2.1}$$

is globally D.A.S. by means of a dynamic feedback, which is smooth on $\mathbb{R}^2$. The proof of the general case follows by induction and applying the same arguments. We divide our proof into two steps. In Step I we use assumptions (i)–(iii) of Theorem 1.1 for the map f_1 in order to establish that the system

$$\dot{x}_1 = f_1(x_1, x_2), \qquad x_2 \doteq u \tag{2.2}$$

is smoothly globally D.A.S. or, equivalently, the system

$$\dot{x}_1 = f_1(x_1, x_2), \quad \dot{x}_2 = v \tag{2.3}$$

is globally S.A.S. by means of a smooth static feedback $v = v_1(x_1, x_2)$. In Step II, taking into account the smoothness of the map ν_1, assumptions (i)–(iii) imposed for the map f_2, and following the same procedure as in Step I, we prove that (2.1) is smoothly globally D.A.S. or, equivalently, the system

$$\dot{x}_1 = f_1(x_1, x_2), \quad \dot{x}_2 = f_2(x_1, x_2, x_3), \quad \dot{x}_3 = v \tag{2.4}$$

is globally S.A.S. by means of a smooth static feedback $v = v_2(x_1, x_2, x_3)$. As in the statement of Lemma 2.1 we assume in what follows that $f_1(0, \mathbb{R}^\pm) = f_2(0, 0, \mathbb{R}^\pm) = \mathbb{R}^\pm$.

Step I: Smoothly global stabilization of (2.3). We define

$$M_1 \doteq \{(x_1, x_2) \in \mathbb{R}^2 : f_1(x_1, x_2) = -x_1\}, \qquad U_1^+ \doteq \{(x_1, x_2) \in \mathbb{R}^2 : f_1(x_1, x_2) > -x_1\},$$
$$U_1^- \doteq \{(x_1, x_2) \in \mathbb{R}^2 : f_1(x_1, x_2) < -x_1\}.$$

Taking into account assumptions (i) and (ii) of Theorem 1.1 and Lemma 2.1, it follows that these sets are nonempty, disjoint, and further that all conditions of Theorem 1.3 are fulfilled with $x_1, x_2, \mathbb{R}, V_1(x_1) \doteq \frac{1}{2}x_1^2, M_1, U_1^+$ and U_1^- instead of $x, y, \mathbb{R}^n, V(x), M, U^+$, and U^-, respectively. Hence according to Theorem 1.3 the system (2.2) is globally D.A.S. In particular, there exists a C^1 nonnegative map $\Phi(x_1, x_2)$ such that the function

$$L(x_1, x_2) \doteqdot V_1(x_1) + \Phi(x_1, x_2)$$

is a global clf with respect to (2.3). Equivalently, the following Lyapunov condition holds:

$$\text{(2.5a)}\qquad \begin{gathered} \frac{\partial L}{\partial x_2}(x_1, x_2) = \frac{\partial \Phi}{\partial x_2}(x_1, x_2) = 0, \qquad (x_1, x_2) \neq 0 \Rightarrow (x_1, x_2) \in M, \\ \left(\frac{\partial L}{\partial x_1} f_1\right)(x_1, x_2) = x_1 f_1(x_1, x_2) + \left(\frac{\partial \Phi}{\partial x_1} f_1\right)(x_1, x_2) < 0, \end{gathered}$$

$$\text{(2.5b)}\qquad \begin{gathered} \frac{\partial \Phi}{\partial x_2}(x_1, x_2) > 0 \quad \text{for } (x_1, x_2) \in U_1^+, \\ \frac{\partial \Phi}{\partial x_2}(x_1, x_2) < 0 \quad \text{for } (x_1, x_2) \in U_1^-, \end{gathered}$$

and since L is uniformly unbounded

$$\text{(2.5c)}\qquad \Phi(x_{1\nu}, x_{2\nu}) \to +\infty$$

for any sequence $\{x_{1\nu}, x_{2\nu}\}$ with $\lim |x_{1\nu}| < +\infty$ and $\lim |x_{2\nu}| = +\infty$.
Moreover, from condition (iii) we get $(\partial f_1/\partial x_2)(0,0) \neq 0$, and so by Lemma 2.2 there exist a positive constant ρ and a real map $x_2 = \phi_1(x_1)$ vanishing at zero, which is smooth on $\{(x_1, x_2) \in \mathbb{R}^2 : |x_1| < \rho\}$, and the intersection of this region with the set M_1 equals $\{(x_1, x_2) \in \mathbb{R}^2 : x_2 = \phi_1(x_1),\ |x_1| < \rho\}$; equivalently

$$\text{(2.6)}\qquad f_1(x_1, \phi_1(x_1)) = -x_1, \qquad |x_1| \leq \rho.$$

(Note that (2.6) implies that the feedback law $x_2 = \phi_2(x_1)$ locally asymptotically stabilizes (2.2) at $0 \in \mathbb{R}$.) It turns out that the map Φ can be constructed so that the following additional conditions are satisfied:

$$\text{(2.7a)}\qquad \Phi(x_1, x_2) = 0, \qquad |x_1| < \rho \Leftrightarrow x_2 = \phi_1(x_1),$$

$$\text{(2.7b)}\qquad (\partial \Phi/\partial x_2)(x_1, x_2) > 0, \quad \forall\, (x_1, x_2) \in \mathbb{R}^2 : |x_1| < \rho;\ x_2 > \phi_1(x_1),$$

$$\text{(2.7c)}\qquad (\partial \Phi/\partial x_2)(x_1, x_2) < 0, \quad \forall\, (x_1, x_2) \in \mathbb{R}^2 : |x_1| < \rho;\ x_2 < \phi_1(x_1),$$

and therefore, since Φ is positive definite,

$$\text{(2.7d)}\qquad D\Phi(x_1, x_2) = 0, \quad \forall\, (x_1, x_2) \text{ with } x_2 = \phi_1(x_1), |x_1| < \rho.$$

(See [15] or §3 of the present paper for a detailed construction of Φ.)

Until now we have shown that the function L is a global clf with respect to (2.3) and so (2.3) is globally asymptotically stabilizable by means of an ordinary static feedback that is in general discontinuous at $0 \in \mathbb{R}^2$. Next, using the additional properties (2.6), (2.7), and the fact that ϕ_1 is C^∞ near zero, we build a smooth stabilizer. First, we recall Theorem 4 in [13], which asserts that since ϕ_1 is smooth and satisfies (2.6), the function

$$\Phi_s(x_1, x_2) \doteqdot V_1(x_1) + \frac{1}{2}(x_2 - \phi_1(x_1))^2, \qquad \left(V_1(x_1) = \frac{1}{2}x_1^2\right)$$

is a local clf with respect to (2.3); in particular, there exists a smooth feedback $\nu = \nu_s(x_1, x_2)$ and a positive constant $\rho_1 \leq \rho$ such that

$$\left(\frac{\partial \Phi_s}{\partial x_1} f_1\right)(x_1, x_2) + \nu_s(x_1, x_2)\frac{\partial \Phi_s}{\partial x_2}(x_1, x_2) < 0, \quad \text{for } 0 < \|(x_1, x_2)\| \leq \rho_1. \tag{2.8}$$

Let $k : \mathbb{R}^2 \to \mathbb{R}^+$ be a smooth map taking values on the interval [0,1] such that $k(x_1, x_2) = 0$ for $\|(x_1, x_2)\| \leq \rho_1/2$ and $k(x_1, x_2) = 1$ for $\|(x_1, x_2)\| \geq \rho_1$. We define

$$V_2(x_1, x_2) \doteqdot V_1(x_1) + k(x_1, 0)\Phi(x_1, x_2) + \frac{1}{2}(1 - k(x_1, 0))(x_2 - \phi_1(x_1))^2. \tag{2.9}$$

Obviously, V_2 is C^1 positive definite. Indeed, V_2 is nonnegative definite, whereas if $V_2(x_1, x_2) = 0$ it follows that $x_1 = 0$, $k(x_1, 0) = 0$, and so $x_2 = \phi_1(0) = 0$. Next we establish that V_2 is a global clf with respect to (2.3) and construct the desired global smooth stabilizer. First, we show that V_2 is uniformly unbounded on $\mathbb{R}^2$. Consider any sequence $(x_{1\nu}, x_{2\nu}) \in \mathbb{R}^2$ with $\|(x_{1\nu}, x_{2\nu})\| \to +\infty$. Without any loss of generality we distinguish two cases. The first is $|x_{1\nu}| \to +\infty$ and $\lim |x_{2\nu}| < +\infty$. Then $k(x_{1\nu}, 0) = 1$ for sufficiently large ν and so

$$\lim V_2(x_{1\nu}, x_{2\nu}) = \lim V_1(x_{1\nu}) = +\infty.$$

The other case is $|x_{2\nu}| \to +\infty$ and $\lim |x_{1\nu}| < +\infty$. Then by (2.5c) $\Phi(x_{1\nu}, x_{2\nu}) \to +\infty$, $|x_{2\nu} - \phi_1(x_{1\nu})| \to +\infty$, $0 \leq k(x_{1\nu}, 0) \leq 1$ for all ν and so

$$\lim V_2(x_{1\nu}, x_{2\nu}) = \lim(k(x_{1\nu}, 0)\Phi(x_{1\nu}, x_{2\nu}) + \frac{1}{2}(1 - k(x_{1\nu}, 0))(x_{2\nu} - \phi_1(x_{1\nu}))^2) = +\infty;$$

therefore V_2 is uniformly unbounded on $\mathbb{R}^2$. We now show that $(\partial V_2/\partial x_1)f_1(x_1, x_2) < 0$ for every $(x_1, x_2) \neq 0$ with $(\partial V_2/\partial x_2)(x_1, x_2) = 0$, which asserts that V_2 is a global clf with respect to (2.3). Indeed, for each nonzero vector (x_1, x_2) with

$$(\partial V_2/\partial x_2)(x_1, x_2) = k(x_1, 0)(\partial \Phi/\partial x_2)(x_1, x_2) + (1 - k(x_1, 0))(x_2 - \phi_1(x_1)) = 0,$$

we get by (2.5) – (2.9) and the definition of k that $(x_1, x_2) \in M_1$ and

$$\left(\frac{\partial V_2}{\partial x_1} f_1\right)(x_1, x_2) = \begin{cases} x_1 f_1(x_1, x_2) < 0, & (x_2, x_2) \in M_1 : |x_1| \leq \rho_1, \\ x_1 f_1(x_1, x_2) + \left(\dfrac{\partial \Phi}{\partial x_1} f_1\right)(x_1, x_2) < 0; & (x_1, x_2) \in M_1, |x_1| > \rho_1. \end{cases}$$

Hence according to Artstein's theorem there exists an ordinary map $v = v_1(x_1, x_2)$ that is smooth on $\mathbb{R}^2 \backslash \{0\}$ and satisfies

$$\left(\frac{\partial V_2}{\partial x_1} f_1\right)(x_1, x_2) + v_1(x_1, x_2)\frac{\partial V_2}{\partial x_2}(x_1, x_2) < 0, \quad \forall (x_1, x_2) \neq 0. \tag{2.10}$$

We are now in a position to prove that the map

$$v_2(x_1, x_2) \doteqdot k(2x_1, 2x_2)v_1(x_1, x_2) + (1 - k(2x_1, 2x_2))v_s(x_1, x_2),$$

where v_s and v_1 are defined in (2.8) and (2.10), respectively, is the desired global smooth stabilizer for (2.3). Obviously, v_2 is smooth on the whole space $\mathbb{R}^2$ and vanishes at $0 \in \mathbb{R}^2$. We use the Lyapunov inequalities (2.8) and (2.10) to prove that the map v_2 globally asymptotically stabilizes (2.3). Indeed, we evaluate the derivative $\dot{V}_2$ of V_2 along the trajectories of the closed-loop system

$$\dot{x}_1 = f_1(x_1, x_2), \qquad \dot{x}_2 = v_2(x_1, x_2).$$

We obtain

$$\begin{aligned}\dot{V}_2(x_1, x_2) &= \left(\frac{\partial V_2}{\partial x_1} f_1\right)(x_1, x_2) + \left(v_2 \frac{\partial V_2}{\partial x_2}\right)(x_1, x_2)\\ &= k(2x_1, 2x_2)\left(\left(\frac{\partial V_2}{\partial x_1} f_1\right)(x_1, x_2) + \left(v_1 \frac{\partial V_2}{\partial x_2}\right)(x_1, x_2)\right)\\ &\quad + (1 - k(2x_1, 2x_2))\left(\left(\frac{\partial V_2}{\partial x_1} f_1\right)(x_1, x_2) + \left(v_s \frac{\partial V_2}{\partial x_2}\right)(x_1, x_2)\right).\end{aligned}$$

For $\|(x_1, x_2)\| \geq \rho_1/2$ we find $k(2x_1, 2x_2) = 1$ and so the previous expression is equal to $(\partial V_2/\partial x_1)f_1 + v_1(\partial V_2/\partial x_2)$, which by (2.10) is strictly negative. For $0 < \|(x_1, x_2)\| < \rho_1/2$ it follows that $|x_1| < \rho_1/2$, $k(x_1, 0) = 0$, $V_2(x_1, x_2) = \Phi_s(x_1, x_2)$, and so

$$\begin{aligned}\dot{V}_2(x_1, x_2) &= k(2x_1, 2x_2)\left(\frac{\partial V_2}{\partial x_1} f_1 + v_1 \frac{\partial V_2}{\partial x_2}\right)(x_1, x_2)\\ &\quad + (1 - k(2x_1, 2x_2))\left(\frac{\partial \Phi_s}{\partial x_1} f_1 + v_s \frac{\partial \Phi_s}{\partial x_2}\right)(x_1, x_2),\end{aligned}$$

which by (2.8) and (2.10) is also strictly negative. Therefore $\dot{V}_2(x_1, x_2) < 0$ for all $(x_1, x_2) \neq 0$; hence we conclude that the map v_2 globally asymptotically stabilizes (2.3) at $0 \in \mathbb{R}^2$.

Step II: Smoothly global stabilization of (2.4). We now show that (2.1) is smoothly globally D.A.S. The procedure is analogous with that of Step I so we present it briefly. Since the map f_2 satisfies conditions (i) – (iii) of Theorem 1.1 we can establish by again using Lemma 2.1 that all conditions of Theorem 1.3 are satisfied with respect to (2.1) with (x_1, x_2), $x_3, \mathbb{R}^2, V_2(x_1, x_2)$,

$$M_2 \doteq \{(x_1, x_2, x_3) \in \mathbb{R}^3 : f_2(x_1, x_2, x_3) = v_2(x_1, x_2)\},$$

$$U_2^+ \doteq \{(x_1, x_2, x_3) \in \mathbb{R}^3 : f_2(x_1, x_2, x_3) > v_2(x_1, x_2)\},$$

$$U_2^- \doteq \{(x_1, x_2, x_3) \in \mathbb{R}^3 : f_2(x_1, x_2, x_3) < v_2(x_1, x_2)\},$$

instead of $x, y, \mathbb{R}^n$, $V(x), M, U^+$, and U^-, respectively. For reasons of completeness we note that the definition of M_2 and the fact that $(\partial V_2/\partial x_1)f_1 + (\partial V_2/\partial x_2)v_2$ is strictly negative for all $(x_1, x_2) \neq 0$ imply

$$\text{(2.11)} \quad \left(\frac{\partial V_2}{\partial x_1} f_1\right)(x_1, x_2) + \frac{\partial V_2}{\partial x_2}(x_1, x_2) f_2(x_1, x_2, x_3) < 0, \quad \forall\, (x_1, x_2, x_3) \in M_2 \backslash \{0\}.$$

Furthermore, from condition (iii) of Theorem 1.1 we get $(\partial f_2/\partial x_3)(0, 0, 0) \neq 0$, which by Lemma 2.2 implies that the restriction at the set M_2 near $0 \in \mathbb{R}^3$ is the graph of a smooth mapping $x_3 = \phi_2(x_1, x_2)$ vanishing at $0 \in \mathbb{R}^2$. Hence by (2.11) and the definition of M_2 we get

$$\text{(2.12)} \qquad \left(\frac{\partial V_2}{\partial x_1} f_1\right)(x_1, x_2) + \frac{\partial V_2}{\partial x_2}(x_1, x_2) f_2(x_1, x_2, \phi_2(x_1, x_2)) < 0$$

for $(x_1, x_2) \neq 0$ near zero, which means that the map $x_3 = \phi_2(x_1, x_2)$ locally asymptotically stabilizes (2.1) at $0 \in \mathbb{R}^2$. Using (2.11), (2.12), and the same procedure as in Step I, with V_2, ϕ_2 and M_2 instead of V_1, ϕ_1 and M_1, respectively, we can construct a smooth map $v = v_3(x_1, x_2, x_3)$ vanishing at zero, which globally asymptotically stabilizes (2.4) at the origin.

Finally, assume that property (iv) of Theorem 1.1 holds for $n = 2$. In that case the equation $f_2(x_1, x_2, x_3) = v_2(x_1, x_2)$ can globally be solved with respect to x_3, namely, there exists a smooth function $x_3 = \phi_2(x_1, x_2)$ with $\phi_2(0, 0) = 0$ such that $f_2(x_1, x_2, \phi_2(x_1, x_2)) = v_2(x_1, x_2)$ for all x_1 and x_2. Therefore (2.12) is fulfilled for every nonzero $(x_1, x_2) \in \mathbb{R}^2$ and so the static feedback ϕ_2 globally asymptotically stabilizes (2.1) at the origin. □

The following proposition provides an algebraic description of the sufficient conditions of Theorem 1.1.

PROPOSITION 2.3. *Suppose that for every i there exist smooth mappings $h_i, g_i : \mathbb{R}^{i+1} \to \mathbb{R}$ vanishing at zero such that $f_i = h_i + g_i$, and furthermore assume*

(A1) *for every $x_0 = (x_{10}, \dots, x_{i0}) \in \mathbb{R}^i$ there exist an odd integer k_i and a positive constant σ_i such that*

$$\frac{\partial^{k_i} h_i}{\partial x_{i+1}^{k_i}}(w_i, y) > \sigma_i \tag{2.13}$$

for every $w_i \div (x_1, \dots, x_i)$ in a neighborhood of x_0 and for every $y \div x_{i+1}$; in particular, if $k_{\min}(x_0)$ denotes the smallest odd integer for which (2.13) *is satisfied, then we assume that*

$$k_{\min}(0) = 1; \tag{2.14}$$

(A2) *there exists a positive definite C^1 function $M_i : \mathbb{R}^i \to \mathbb{R}$ such that*

$$|g_i(w_i, y)| \leq M_i(w_i) \tag{2.15}$$

for all w_i and y.

Then conditions (i), (ii), *and* (iii) *of Theorem* 1.1 *are satisfied and so* (1.3) *is smoothly globally D.A.S.*

Proof. According to assumption (A1), for every i and x_0 there exist an odd integer k_i and a compact neighborhood S of x_0 such that f_i is written

$$\begin{aligned} f_i(w_i, y) &= \sum_{l=0}^{k_i - 1} \frac{y^l}{l} \frac{\partial^l h_i}{\partial x_{i+1}^l}(w_i, 0) + C_i(w_i, y) + g_i(w_i, y), \\ C_i(w_i, y) &\div \int_0^y \int_0^{s_1} \cdots \int_0^{s_{k_i - 1}} \frac{\partial^{k_i} h_i}{\partial x_{i+1}^{k_i}}(w_i, s_{k_i})\, ds_{k_i} \dots ds_2\, ds_1 \end{aligned} \tag{2.16}$$

and (2.13) is satisfied for each $w_i \in S$ and for all $y = x_{i+1}$. Since k_i is odd, it follows by (2.13) that

$$\underline{\lim} |y_\nu^{-k_i}| C_i(w_{i\nu}, y_\nu) > 0, \quad \text{as } |y_\nu| \to +\infty \tag{2.17}$$

for every sequence $w_{i\nu} \in S$. From (2.15), (2.16), and (2.17) we get $\lim f_i(w_{i\nu}, y_\nu) = \pm\infty$ as $y_\nu \to \pm\infty$. This in conjunction with the continuity of f_i implies that for every $a \in \mathbb{R}$ and $w_i \in \mathbb{R}^i$ there is a y with $f_i(w_i, y) = a$ from which it follows that the map $f_i(w_i, \cdot)$ is a surjection. Similarly, condition (ii) of Theorem 1.1 is a direct consequence of (2.17). Finally, notice that, since g_i is C^∞ vanishing at zero, it follows by (2.15) that $g_i(0, y) = (\partial g_i / \partial x_{i+1})(0, y) = 0$ for every $y = x_{i+1}$. The latter in conjunction with (2.13) and (2.14) implies (1.4) and condition (iii) of Theorem 1.1. □

Example 2.4. Consider the system $\dot{x}_1 = x_2(1-x_1)+x_2^3+x_1\sin x_1x_2$, $\dot{x}_2 = u(1-x_1-x_2)+u^3+x_1x_2\sin x_1x_2u$. We define $f_1 \doteqdot h_1+g_1$, $f_2 \doteqdot h_2+g_2$, $h_1 \doteqdot x_2(1-x_1)+x_2^3$, $h_2 \doteqdot u(1-x_1-x_2)+u^3$, $g_1 \doteqdot x_1\sin x_1x_2$; $g_2 \doteqdot x_1x_2\sin x_1x_2u$. Then we can easily justify that conditions (A1) and (A2) of Proposition 2.3 are satisfied and so the system is smoothly globally D.A.S.

The following corollary is a direct consequence of Proposition 2.3 and extends the main result in [14] dealing with global stabilization by linear feedback for triangular systems of the form:

$$\begin{aligned} \dot{x}_1 &= x_2+g_1(x_1,x_2)\\ \dot{x}_2 &= x_3+g_2(x_1,x_2,x_3)\\ &\vdots\\ \dot{x}_{n-1} &= x_n+g_{n-1}(x_1,x_2,\ldots,x_{n-1},x_n)\\ \dot{x}_n &= u+g_n(x_1,x_2,\ldots,x_{n-1},x_n,u), \end{aligned} \tag{2.18}$$

where the mappings g_i, $1\le i\le n$ are globally Lipschitzian.

COROLLARY 2.5. *Suppose that each g_i is smooth vanishing at zero and satisfies* (A2) *of Proposition* 2.3. *Then* (2.18) *is smoothly globally D.A.S.*

2.2. Controllability. We conclude this section by giving some links between the main assumptions of Theorem 1.1 and the null controllability of (1.3). It is well known (see [7]) that if a nonlinear system, whose linearization at the origin is controllable, is globally S.A.S., then there exists a $t>0$ such that the domain C_t consisting of all points $x_0\in\mathbb{R}^n$, each of which can be steered to zero at time t by some measurable input u, is the whole state space. It turns out that the domain of null controllability, namely, the set $C \doteqdot \cup_{t\ge 0} C_t$, covers the whole state space. It is not difficult to establish that the same result holds under the weaker assumption that the system is globally D.A.S. Moreover, for triangular systems (1.3) this result is strengthened as follows.

PROPOSITION 2.6. *If conditions* (i), (ii), *and* (iii) *of Theorem* 1.1 *are satisfied, then $C_t=\mathbb{R}^n$ for every $t>0$, namely, each initial state $x_0\in\mathbb{R}^n$, can be steered to zero at any time $t>0$ by an appropriate measurable input.*

Proof. For reasons of simplicity consider the planar case (2.1). Condition (iii) of Theorem 1.1 guarantees that for every $t_1>0$ the set C_{t_1} covers an open neighborhood S of $0\in\mathbb{R}^2$. To complete the proof it suffices to show that for any initial state $x_0=(x_{10},x_{20})\in\mathbb{R}^2$ there exists a measurable controller such that the endpoint of the corresponding trajectory of (2.1) lies in S at time $t_2\doteqdot t-t_1$. Let (a_1,a_2) be an arbitrary vector in S and let $x_1:[0,t_2]\to\mathbb{R}$ be a C^1 function with $x_1(0)=x_{10}$ and $x_1(t_2)=a_1$. Then by using assumptions (i) and (ii) of Theorem 1.1 we can find a measurable function $u_1:[0,t_2]\to\mathbb{R}$ such that

$$\dot{x}_1(s)=f_1(x_1(s),u_1(s)),\quad \forall s\in[0,t_2] \tag{2.19}$$

(see also [7, p. 162] where analogous arguments are used). From (2.19) it follows that x_1 is the solution of the subsystem $\dot{x}_1=f_1(x_1,x_2)$ with input $x_2=u_1$ starting from x_{10} with endpoint $x_1(t_2)=a_1$; equivalently

$$x_1(s)=x_{10}+\int_0^s f_1(x_1(\rho),u_1(\rho))d\rho,\quad \forall s\in[0,t_2]. \tag{2.20}$$

Let $x_{2\nu}(s)$, $s\in[0,t_2]$ be a sequence of C^1 functions such that

$$\int_0^{t_2}|x_{2\nu}(s)-u_1(s)|ds\to 0,\quad \text{as}\,\nu\to+\infty \tag{2.21}$$

and $x_{2\nu}(0) = x_{20}$; $x_{2\nu}(t_2) = a_2$ for all ν. Then by (2.20), (2.21), and using Gronwall's inequality it follows that for every $\varepsilon > 0$ there is an integer k such that $|a_1 - \hat{x}_1(t_2)| < \varepsilon$, where $\hat{x}_1(s) \doteq x_1(s, x_{10}, x_{2k})$ denotes the trajectory of $\dot{x}_1 = f_1(x_1, x_2)$ with input $\hat{x}_2(s) = x_{2k}(s)$ starting from x_{10}. Consider finally a measurable function $u_2 : [0, t_2] \longrightarrow \mathbb{R}$ such that

$$\dot{\hat{x}}_2(s) = f_2(\hat{x}_1(s), \hat{x}_2(s), u_2(s)), \quad \forall\, s \in [0, t_2], \tag{2.22}$$

whose existence is also guaranteed by assumptions (i) and (ii) of Theorem 1.1. From (2.19) and (2.22) it follows that $(\hat{x}_1, \hat{x}_2) : [0, t_2] \longrightarrow \mathbb{R}$ is the trajectory of (2.1) with input $x_3 = u_2$, starting from (x_{10}, x_{20}) and with endpoint $\|(\hat{x}_1(t_2), \hat{x}_2(t_2)) - (a_1, a_2)\| < \varepsilon$. Taking into account that $(a_1, a_2) \in S$ and by choosing ε appropriately small it follows that $\hat{x}(t_2) \doteq (\hat{x}_1(t_2), \hat{x}_2(t_2))$ also lies in S. Since $S \subset C_{t_1}$, $\hat{x}(t_2)$ can be steered to zero by some input $w : [0, t_1] \longrightarrow \mathbb{R}$. We conclude that x_0 can be steered to the origin at time $t_1 + t_2 = t$ by the concatenation of u_2 and w. The proof of the general case follows similarly by induction. □

3. The general case: A necessary and sufficient condition. The following theorem generalizes Theorem 1.3. It provides a necessary and sufficient condition for dynamic stabilization for the general nonlinear case (1.1).

THEOREM 3.1. *The system* (1.1) *is globally D.A.S. if and only if there exist a closed subset $M \subset \mathbb{R}^{n+1}$, a pair of nonempty disjoint subsets U^+ and U^- of $\mathbb{R}^{n+1}$, and a C^1 nonnegative function $W : \mathbb{R}^{n+1} \longrightarrow \mathbb{R}$ such that*

(i) $0 \in M$, $\mathbb{R}^{n+1} = U^+ \cup U^- \cup M$; $\pi(M) = \mathbb{R}^n$;

(ii) *for any $\varepsilon > 0$ the following hold*:

$$\{0\} \times (0, \varepsilon) \cap U^+ \neq \emptyset, \qquad \{0\} \times (-\varepsilon, 0) \cap U^- \neq \emptyset; \tag{3.1}$$

(iii) *the function W is strictly positive on the region $M \backslash \{0\}$, whereas $W(0, 0) = 0$. Moreover assume that*

$$(x; y) \in M : \frac{\partial W}{\partial y}(x, y) = 0 \Rightarrow \left(\frac{\partial W}{\partial x} f\right)(x, y) \leq -c(\|(x, y)\|), \tag{3.2}$$

where $c : \mathbb{R}^+ \longrightarrow \mathbb{R}^+$ is a positive definite function with $c(0) = 0$ and further

$$\frac{\partial W}{\partial y}(x, y) \geq 0, \quad \forall\, (x, y) \in U^+; \qquad \frac{\partial W}{\partial y}(x, y) \leq 0, \quad \forall\, (x, y) \in U^-; \tag{3.3}$$

(iv) *W is uniformly unbounded on $\mathbb{R}^{n+1}$.*

The system (1.1) *is locally D.A.S. if and only if $U^+ \cup U^- \cup M$ and $\pi(M)$ are neighborhoods of $0 \in \mathbb{R}^{n+1}$, $0 \in \mathbb{R}^n$, respectively, and further* (3.1), (3.2), *and* (3.3) *are satisfied.*

Proof. We establish only the global part of our theorem. Similar arguments can be repeated in a local setting. We proceed to the construction of a global clf guaranteeing global dynamic stabilization as follows. For each integer ν consider a locally finite partition of $\mathbb{R}^n$ that consists of subsets $A_{\nu i} \subseteq \mathbb{R}^n$, $i = 1, 2, \ldots$, with $\operatorname{int} A_{\nu i_1} \cap \operatorname{int} A_{\nu i_2} = \emptyset$ for $i_1 \neq i_2$. The diameter of each $A_{\nu i}$ is tending to zero as $\nu \longrightarrow +\infty$ uniformly on i, there exists an integer $i_0 = i_0(\nu)$ such that

$$0 \in \operatorname{int} A_{\nu i_0} \tag{3.4a}$$

and further for every $x \in \mathbb{R}^n$ and for almost all ν we can find an integer $i = i(\nu)$ with $x \in \operatorname{int} A_{\nu i}$. Similarly, for each ν and i consider a partition of the y axis of subintervals $B_{\nu ij} \doteq [y_{\nu ij}, y_{\nu i(j+1)}]$, $j \in \{\pm 1, \pm 2, \ldots\}$ such that $y_{\nu ij} > 0$ for $j = 1, 2, \ldots$; $y_{\nu ij} < 0$ for

$j = -1, -2, \ldots$; $|y_{\nu ij} - y_{\nu i(j+1)}| < \frac{1}{\nu}$ for all ν, i, j; and further each nonzero vector (x, y) belongs to the interior of $S_{\nu ij} \doteqdot A_{\nu i} \times B_{\nu ij}$ for almost all ν and some i and j depending on ν. Obviously, for each ν the family $\{S_{\nu ij}\}$ consists of a partition of $\mathbb{R}^{n+1}$ and due to our assumption (3.1) we may assume that there exists an integer $j_0 = j_0(\nu, i_0)$ depending on ν and i_0 (i_0 being the integer defined by (3.4a)) such that

$$0 \in \text{int} S_{\nu i_0 j_0}, \quad S_{\nu i_0(j_0-1)} \subset U^-, \quad S_{\nu i_0(j_0+1)} \subset U^+. \tag{3.4b}$$

We are in a position to construct for each ν and i a nonnegative smooth function $\phi_{\nu i}(y)$, $y \in \mathbb{R}$ such that $D\phi_{\nu i}(y) = 0$ for $y \in clB_{\nu ij}$ with $S_{\nu ij} \cap M \neq \emptyset$; $D\phi_{\nu i}(y) > 0$ for $y \in \text{int}B_{\nu ij}$ with $S_{\nu ij} \subset U^+$, $D\phi_{\nu i}(y) < 0$ for $y \in \text{int}B_{\nu ij}$ with $S_{\nu ij} \subset U^-$ and $|D\phi_{\nu i}(y)| \leq 1$; $0 \leq \phi_{\nu i}(y) < 1$, for every $y \in \mathbb{R}$. In particular, $\phi_{\nu i}(y)$ is strictly positive for $y \in \text{int}B_{\nu ij}$ with $S_{\nu ij} \subset U^+ \cup U^-$ and furthermore by (3.4) the map $\phi_{\nu i_0}$ can be constructed such that the previous conditions are satisfied and in addition

$$\phi_{\nu i_0}(y) = 0 \quad \text{for} y \in B_{\nu i_0 j_0}. \tag{3.5}$$

Consider now for every ν and i a nonnegative smooth map $a_{\nu i} : \mathbb{R}^n \to \mathbb{R}$ such that $\|Da_{\nu i}(x)\| \leq 1$ for $x \in A_{\nu i}$, $0 < a_{\nu i}(x) < 1$ for $x \in \text{int}A_{\nu i}$; $a_{\nu i}(x) = 0$ otherwise, and define

$$\psi_\nu(x, y) \doteqdot \sum_{i=1}^{\infty} \phi_{\nu i}(y) a_{\nu i}(x), \qquad (x, y) \in \mathbb{R}^n \times \mathbb{R}. \tag{3.6}$$

Obviously, ψ_ν is smooth nonnegative definite and vanishes at zero. Moreover ψ_ν as well as its derivative are bounded. Indeed, $0 \leq \psi_\nu(x, y) = \phi_{\nu i}(y) a_{\nu i}(x) \leq 1$; $|D\psi_\nu(x, y)| \leq |D\phi_{\nu i}(y)| a_{\nu i}(x) + \|Da_{\nu i}(x)\| \phi_{\nu i}(y) \leq 2$ for all $y \in \mathbb{R}$ and $x \in A_{\nu i}$. It turns out that the map

$$\Phi(x, y) \doteqdot \sum_{\nu=1}^{+\infty} 2^{-\nu} \psi_\nu(x, y) \tag{3.7}$$

is well defined and C^1. In particular, $D\Phi(x, y) = \sum 2^{-\nu} D\psi_\nu(x, y)$, where the series on the right-hand side uniformly converges to $D\Phi$. Moreover, Φ is nonnegative definite, vanishes at $0 \in \mathbb{R}^{n+1}$, and satisfies $\Phi(x, y) > 0$ for all $(x, y) \notin M$. Indeed, for $(x, y) = 0$ we get from (3.5)–(3.7) that $\Phi(0, 0) = 0$, whereas if $(x, y) \notin M$, it follows that $(x, y) \in \text{int}S_{\nu ij} \subset U^+ \cup U^-$ for almost all integers ν and for some i, j depending on ν; therefore $\psi_\nu(x, y) > 0$ and so $\Phi(x, y) > 0$. We can also easily establish that

$$\frac{\partial \Phi}{\partial y}(x, y) > 0 \Leftrightarrow (x, y) \in U^+, \tag{3.8a}$$

$$\frac{\partial \Phi}{\partial y}(x, y) < 0 \Leftrightarrow (x, y) \in U^-, \tag{3.8b}$$

$$\frac{\partial \Phi}{\partial y}(x, y) = 0 \Leftrightarrow (x, y) \in M. \tag{3.8c}$$

For reasons of completeness we prove that $(x, y) \in U^+$ implies $(\partial\Phi/\partial y)(x, y) > 0$. Indeed, let $(x, y) \in U^+$. Then for almost all integers ν there corresponds a pair of indices i_ν, j_ν

depending on ν with $(x,y) \in \mathrm{int} S_{\nu i_\nu j_\nu} \subset U^+$. Consequently, $D\phi_{\nu i_\nu}(y) > 0$, $a_{\nu i_\nu}(x) > 0$, and so $(\partial\psi_\nu/\partial y)(x,y) = a_{\nu i_\nu}(x) D\phi_{\nu i_\nu}(y) > 0$ for almost all ν, whereas $(\partial\psi_\nu/\partial y)\,(x,y) \geq 0$ otherwise. Therefore $(\partial\Phi/\partial y)(x,y) = \sum 2^{-\nu}(\partial\psi_\nu/\partial y)(x,y) > 0$.

Consider now a pair of continuous functions $a, b : \mathbb{R}^+ \longrightarrow \mathbb{R}^+$ such that b is positive definite, a is increasing and everywhere strictly positive, and furthermore

$$b(\Phi(x,y)) < c(\|(x,y)\|), \tag{3.9}$$

$$\frac{1}{2} + \left| \left(\frac{\partial\Phi}{\partial x} f \right)(x,y) \right| \leq a(W(x,y)), \quad \forall\,(x,y) \in M, \tag{3.10}$$

where c and W are defined in (3.2). Note that the existence of the map satisfying (3.10) follows from our assumption that W is strictly positive on $M\backslash\{0\}$. Finally, we define

$$L(x,y) \div \int_0^{W(x,y)} a(r)dr + \int_0^{\Phi(x,y)} b(r)dr.$$

Obviously L is C^1 and positive definite (the latter follows by our assumption that $W(x,y) > 0$ for $(x,y) \in M\backslash\{0\}$), and because of assumption (iv) it is uniformly unbounded on $\mathbb{R}^{n+1}$. We complete the proof by showing that L is a clf with respect to (1.2). Indeed, for each nonzero $(x,y) \in \mathbb{R}^{n+1}$ with

$$\frac{\partial L}{\partial y}(x,y) = 0 \tag{3.11}$$

it follows that $((\partial W/\partial y)a(W) + (\partial\Phi/\partial y)b(\Phi))(x,y) = 0$ and so by (3.3) and (3.8) we get $(\partial W/\partial y)(x,y) = (\partial\Phi/\partial y)(x,y) = 0$ and $(x,y) \in M$. Therefore by (3.2), (3.9), and (3.10) it follows that

$$\begin{aligned} \left(\frac{\partial L}{\partial x} f\right)(x,y) &= a(W(x,y)) \left(\frac{\partial W}{\partial x} f\right)(x,y) + b(\Phi(x,y)) \left(\frac{\partial \Phi}{\partial x} f\right)(x,y) \\ &\leq -\frac{1}{2} c(\|(x,y)\|) < 0. \end{aligned} \tag{3.12}$$

The implication (3.11) $\Rightarrow$ (3.12) asserts that L is a global clf with respect to (1.2); hence, the system (1.1) is globally D.A.S.

The converse part of the proof is straightforward. Assume that there exists a map $v : \mathbb{R}^{n+1} \longrightarrow \mathbb{R}$ that is smooth for $(x,y) \neq 0$ and such that $0 \in \mathbb{R}^{n+1}$ is globally asymptotically stable with respect to $\dot{x} = f(x,y), \dot{y} = v(x,y)$. Then according to Artstein's version of the well-known Kurzweil's converse stability theorem (see [2]) there exists a uniformly unbounded smooth Lyapunov function W of $0 \in \mathbb{R}^{n+1}$ with respect to (1.2), namely, $((\partial W/\partial x)f + (\partial W/\partial y)v)(x,y) < 0$ for all $(x,y) \neq 0$. Using the previous inequality we can easily justify that all conditions of Theorem 3.1 are fulfilled with W as above: $M = \{(x,y) \in \mathbb{R}^{n+1} : (\partial W/\partial y)(x,y) = 0\}$, $U^+ = \{(x,y) \in \mathbb{R}^{n+1} : (\partial W/\partial y)(x,y) > 0\}$, and $U^- = \{(x,y) \in \mathbb{R}^{n+1} : (\partial W/\partial y)(x,y) < 0\}$. For reasons of completeness we note that the previous definitions of M, U^+, and U^- and the fact that W is positive definite and uniformly unbounded on $\mathbb{R}^{n+1}$ imply conditions (i) and (ii) of the present theorem. □

Next we show that if conditions (i)–(iv) of Theorem 1.3 are fulfilled, there exists a function W satisfying conditions (iii) and (iv) of Theorem 3.1. This in conjunction with the fact that assumption (1.6) is a special case of (3.1) asserts that Theorem 3.1 is indeed a generalization of Theorem 1.3.

PROPOSITION 3.2. *If conditions* (i) – (iv) *of Theorem* 1.3 *are fulfilled, there exists a* C^1 *nonnegative map* $W : \mathbb{R}^{n+1} \to \mathbb{R}$ *satisfying conditions* (iii) *and* (iv) *of Theorem* 3.1.

Proof (outline). By using condition (ii) of Theorem 1.3 we can construct a C^1 nonnegative function $W_0(x, y)$ that vanishes in a closed neighborhood of M, is strictly positive otherwise, and satisfies

$$\frac{\partial W_0}{\partial y}(x, y) \geq 0, \quad \forall (x, y) \in U^+; \qquad \frac{\partial W_0}{\partial y}(x, y) \leq 0, \quad \forall (x, y) \in U^-; \tag{3.13}$$

(3.14)

$$W_0(x_\nu, y_\nu) \to +\infty, \quad \text{for any sequence } (x_\nu, y_\nu) \text{ with } \lim \|x_\nu\| < +\infty \text{ and } \lim |y_\nu| \to +\infty.$$

Then the map $W(x, y) = V(x) + W_0(x, y)$ is C^1, nonnegative definite, and satisfies conditions (iii) and (iv) of Theorem 3.1. Indeed, $W(0, 0) = 0$, and for each nonzero $(x, y) \in M$ with $(\partial W/\partial y)(x, y) = 0$ it follows that $DW_0(x, y) = 0$ and $((\partial W/\partial x) f)(x, y) = (DVf)(x, y)$, which by (1.7) is strictly negative. Furthermore, similar to [15], by condition (ii) of Theorem 1.3 we can establish the existence of a positive definite continuous function $c : \mathbb{R}^+ \to \mathbb{R}^+$ such that (3.2) holds. Condition (3.3) is an immediate consequence of (3.13). Finally, (3.14) and the uniform unboundedness of V imply that W is uniformly unbounded on $\mathbb{R}^{n+1}$. □

The following example has been carefully devised to show the applicability of Theorem 3.1 for systems (1.1) that are not necessarily stabilizable by ordinary static feedback.

Example 3.3. Consider a planar nonlinear system (1.1) with

$$\begin{aligned} f_1(x_1, x_2; u) &= -x_1 + x_1^2(\phi(u) - u^2)^2, \\ f_2(x_1, x_2; u) &= -x_1^3(\phi(u) - u^2) - (\phi(u)/2)^2 + (x_2 - \phi(u))^2, \end{aligned}$$

where ϕ is continuous, vanishing at zero, and satisfies $\phi(u) \geq 2u^2$ for all $u \geq 0$. We show that this system satisfies all conditions of Theorem 3.1 and so is globally D.A.S. We define

$$W(x_1, x_2, y) \doteq \begin{cases} \frac{1}{2}x_1^2 + \frac{1}{2}(x_2 - y^2)^2 + 2\int_0^y t(\phi(t) - t^2)dt, & y > 0; \\ \frac{1}{2}x_1^2 + \frac{1}{2}x_2^2, \quad -1 \leq y \leq 0; \\ \frac{1}{2}x_1^2 + \frac{1}{2}x_2^2 + (y + 1)^2, & y \leq -1; \end{cases}$$

$$M \doteq \{(x_1, x_2, y) \in \mathbb{R}^3 : x_2 = \phi(y),\ y > 0\} \cup \{(x_1, x_2, y) \in \mathbb{R}^3 : y = 0,\ x_2 < 0\};$$

$$U^+ \doteq \{(x_1, x_2, y) \in \mathbb{R}^3 : y > 0,\ x_2 < \phi(y)\};$$

$$\begin{aligned} U^- \doteq \{(x_1, x_2, y) \in \mathbb{R}^3 : x_2 > \phi(y), y > 0\} &\cup \{(x_1, x_2, y) \in \mathbb{R}^3 : y = 0, x_2 > 0\} \\ &\cup \{(x_1, x_2, y) \in \mathbb{R}^3 : y < 0\}. \end{aligned}$$

Then W is nonnegative definite ($W(x_1, x_2, y) = 0$ for $x_1 = x_2 = 0$; $y \in [-1, 0]$ and $W(x_1, x_2, y) > 0$ otherwise) and uniformly unbounded on $\mathbb{R}^3$. We now evaluate $(\partial W/\partial y)(x_1, x_2, y) = 2y(\phi(y) - x_2), y \geq 0$. It follows that for each nonzero (x_1, x_2, y) with $(\partial W/\partial y)(x_1, x_2, y) = 0$ and $y > 0$ we have $x_2 = \phi(y)$ and so $((\partial W/\partial x) f)(x_1, x_2, y) =$

$-x_1^2 - \left(\frac{\phi}{2}\right)^2 (\phi - y^2) \leq -x_1^2 - y^6$, whereas for $x_2 < 0$ and $y = 0$ we find $((\partial W/\partial x)f)\cdot (x_1, x_2, y) = -x_1^2 + x_2^3 < 0$. Therefore (3.3) is fulfilled with W and M as previously defined. We can also easily justify that the remaining conditions of Theorem 3.1 are satisfied and so the system is globally D.A.S. It must be noted that the system above is not necessarily S.A.S. Indeed, suppose for instance that ϕ has the form $\phi(u) = 2u^2 + u\left(1 + \frac{1}{2} \sin \frac{1}{u}\right)$, $u > 0$. Since the region $\{(x_1, x_2) \in \mathbb{R}^2 : x_1 = 0\}$ is invariant with respect to $\dot{x}_1 = -x_1 + x_1^2(\phi(u) - u^2)^2$, it suffices to show that the system $\dot{x}_2 = -\phi^2(u)/4 + (x_2 - \phi(u))^2$ cannot be S.A.S. Suppose on the contrary that there exists an ordinary feedback $u = u(x_2)$ that is continuous for $x_2 \neq 0$ near zero and locally asymptotically stabilizes the previous system at $0 \in \mathbb{R}$. This implies $-\frac{1}{4}\phi^2(u(x_2)) + (x_2 - \phi(u(x_2)))^2 < 0$, or $\phi(u(x_2)) < x_2 < \frac{3}{2}\phi(u(x_2))$ for $x_2 > 0$ near zero. The latter is impossible because of the particular choice of ϕ (see also [15], where an analogous example is investigated).

4. Conclusion and further remarks.

Sufficient conditions for smoothly global dynamic stabilizations for a wide class of triangular controllable systems have been presented. A necessary and sufficient condition for almost smoothly global dynamic stabilization for the general nonlinear case has also been provided.

It is worth remarking that the approach of the present work can be applied to derive further interesting results. For instance, we can combine Theorem 1.3 and a well-known result in [4], [5] concerning the local stabilization problem for planar systems

$$\dot{x} = f(x, y), \quad \dot{y} = \nu, \quad (x, y) \in \mathbb{R}^2 \tag{4.1}$$

in order to prove that (4.1) is globally S.A.S. by means of a feedback law that is smooth on $\mathbb{R}^2\backslash\{0\}$ and further it is continuous at $0 \in \mathbb{R}^2$, provided that

(B1) the map $(x, y) \to f(x, y)$ is analytic near zero;

(B2) there exists a pair of disjoint open sets $U^+, U^- \subset \mathbb{R}^2$ and a closed subset $M \subset \mathbb{R}^2$ containing zero such that conditions (i) and (ii) of Theorem 1.3 hold and further

$$xf(x, y) < 0, \quad \forall\, (x, y) \in M. \tag{4.2}$$

For the sake of completeness we note that assumption B1 in conjunction with (4.2) implies the existence of a continuous map $x \to \phi(x)$, $\phi(0) = 0$, and a real constant $\rho > 0$ such that $xf(x, \phi(x)) < 0$ for $0 < |x| \leq \rho$. Then according to [4], [5] there exists a real map $\nu = v_1(x, y)$, $v_1(0, 0) = 0$ that is smooth for $(x, y) \neq 0$ near zero and continuous at zero and that locally stabilizes (4.1) at $0 \in \mathbb{R}^2$. The assumption B1 also asserts that the system (4.1) satisfies condition (B2) with $M' = \{(x, y) \in \mathbb{R}^2;\ 0 < |x| \leq \rho,\ y = \phi(x)\} \cap \{(x, y) \in M,\ |x| \geq \rho\}$ instead of M. Then we can combine Theorem 1.3 and the approach in [4] as in the proof of Theorem 1.3 to build a feedback law $v = v_2(x, y)$ that globally asymptotically stabilizes (4.1), is smooth on $\mathbb{R}^2\backslash\{0\}$, and coincides with $v_1(x, y)$ near zero; hence it is continuous on the whole state space. Details are found in [16].

REFERENCES

[1] A. Andreini, A. Bacciotti, and G. Stefani, *Global stabilizability of homogeneous vector fields of odd degree*, Systems Control Lett., 10 (1985), pp. 251–256.

[2] Z. Artstein, *Stabilization with relaxed controls*, Nonlinear Anal., 7 (1983), pp. 1163–1173.

[3] R. W. Brockett, *Feedback invariants for nonlinear systems*, in Proc. IFAC Congress, Helsinki, 1978.

[4] J. M. Coron and L. Praly, *Adding an integrator for the stabilization problem*, Systems Control Lett., 17 (1991), pp. 89–104.

[5] W. P. Dayawansa, C. F. Martin, and G. Knowles, *Asymptotic stabilization of a class of smooth two-dimensional systems*, SIAM J. Control. Optim., 28 (1990), pp. 1321–1349.

[6] I. Kanellakopoulos, P. V. Kokotovic, and A. S. Morse, *A toolkit for nonlinear feedback design*, Systems Control Lett., 18 (1992), pp. 83–92.

[7] E. B. Lee and L. Markus, *Foundations of Optimal Control Theory*, John Wiley, New York, 1967.

[8] Y. Lin and E. D. Sontag, *A universal formula for stabilization with bounded controls*, Systems Control Lett., 16 (1991), pp. 393–397.

[9] R. Marino and P. Tomei, *Robust stabilization of feedback linearizable time-varying uncertain nonlinear systems*, Automatica, 29(1) (1993), pp. 181–189.

[10] H. Nijmeijer and A. J. van der Schaft, *Nonlinear Dynamical Control Systems*, Springer-Verlag, New York, 1990.

[11] E. D. Sontag, *A "universal" construction of Artstein's theorem on nonlinear stabilization*, Systems Control Lett., 13 (1989), pp. 117–123.

[12] A. R. Teel, *Semi-global stabilization of minimum phase nonlinear systems in special normal form*, Systems Control Lett., 19 (1992), pp. 187–192.

[13] J. Tsinias, *Sufficient Lyapunov-like conditions for stabilization*, Math. Control Signals Systems, 2 (1989), pp. 343–357.

[14] ———, *A theorem on global stabilization of nonlinear systems by linear feedback*, Systems Control Lett., 17 (1991), pp. 357–362.

[15] ———, *An extension of Artstein's theorem on stabilization by using ordinary feedback integrators*, Systems Control Lett., 5 (1993), pp. 141–148.

[16] ———, *Global stabilization for planar systems*, IEEE Med. Conf. on Decision and Control, 1993.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 33, No. 4, pp. 1086–1101, July 1995

©1995 Society for Industrial and Applied Mathematics

008

A MIXED $l_\infty/\mathcal{H}_\infty$ OPTIMIZATION APPROACH TO ROBUST CONTROLLER DESIGN*

MARIO SZNAIER[†]

Abstract. In spite of its practical importance, the problem of designing controllers capable of satisfying mixed time/frequency domain performance requirements under model uncertainty remains, to a large extent, open. In this paper we propose a design procedure for minimizing the maximum amplitude of a regulated error to a specified input while, at the same time, addressing model uncertainty through bounds on the $\mathcal{H}_\infty$ norm of a relevant transfer function. This problem is of interest in optimal tracking applications where the objective is to achieve minimum tracking error while, at the same time, maintaining an adequate robustness level. We show that for the SISO case, the problem can be solved by solving sequence of problems, each one consisting of a finite-dimensional convex optimization and an unconstrained Nehari approximation problem

Key words. robust control synthesis, $\mathcal{H}_\infty$ control, l_∞ control, discrete-time systems

AMS subject classifications. 93, 93B35, 93B36, 93B51, 93C55

1. Introduction.

A large number of control problems involve designing a controller capable of achieving acceptable performance under system uncertainty and design constraints. However, in spite of its practical importance, this problem remains, to a large extent, open. During the last decade a large research effort led to procedures for designing robust controllers, capable of achieving desirable properties under various classes of plant uncertainties while, at the same time, satisfying frequency-domain constraints. However, these design procedures cannot accommodate directly time-domain performance specifications.

Recently, some progress has been made in this direction [1]–[5]. By using a parametrization of all stabilizing linear controllers in terms of a stable transfer matrix Q, the problem of finding the "best" linear controller can be formulated as the constrained optimization problem of minimizing a weighted ∞-norm over the set of suitable Q. In this formulation, additional specifications can be imposed by further constraining the problem. The resulting optimization problem has been solved using convex programming [1] and constrained nondifferentiable optimization [2]. However, although these methods are effective when the specifications are easily expressed in terms of the frequency response, presently they can handle time-domain specifications in a conservative fashion, through the use of several approximations. Additionally, they may require solving very large nondifferentiable optimization problems. A different approach has been pursued in [3]–[5], where time-domain constraints over a finite horizon are incorporated into an $\mathcal{H}_\infty$ optimal control problem that is then transformed into a finite-dimensional optimization problem. However, at this stage constraints over an infinite horizon can be handled only indirectly.

Finally, in [6] and [7] the problems of finding an internally stabilizing compensator that minimizes the maximum error to l_∞ bounded disturbances and to a fixed, given

*Received by the editors May 15, 1992; accepted for publication (in revised form) January 10, 1994. This research was supported in part by National Science Foundation grant ECS-9211169 and Florida Space Grant Consortium.

[†]Department of Electrical Engineering, The Pennsylvania State University, University Park, Pennsylvania 16802.

signal was solved. However, these designs cannot accommodate frequency-domain specifications.

In this paper we address the problem of finding an internally stabilizing compensator that minimizes the maximum amplitude of the error to a fixed given input subject to constraints upon the $\mathcal{H}_\infty$ norm of a relevant transfer function. This problem, which can be thought of as the dual of the problem proposed in [3]–[5], is of particular interest for optimal tracking problems where the objective is to achieve minimum error magnitude, while at the same time maintaining an adequate robustness level against model uncertainty.

The paper is organized as follows: In §2 we give a formal definition to the mixed $l_\infty/\mathcal{H}_\infty$ optimization problem. In §3 we propose a solution method using a technique similar to the one that we presented in [8]–[10]. The main result of this section shows that the mixed optimization problem can be solved by solving a sequence of modified problems, each one consisting of a finite-dimensional convex, constrained optimization problem, and an *unconstrained* Nehari approximation. In §4 we present a simple design example and compare our controller to the unconstrained optimal $\mathcal{H}_\infty$ controller. Finally, in §5, we summarize our results and indicate directions for future research.

2. Problem formulation.

2.1. Notation. By $\mathcal{L}_\infty$ we denote the Lebesgue space of complex-valued transfer functions essentially bounded on the unit circle, equipped with the norm $\|G(z)\|_\infty \stackrel{\Delta}{=} \sup_{|z|=1} |G(z)|$. $\mathcal{H}_\infty$ ($\mathcal{H}_\infty^-$) denotes the space of stable (antistable) complex functions $G(z) \in \mathcal{L}_\infty$, i.e analytic in $|z| \geq 1$ ($|z| \leq 1$), equipped with the norm $\|.\|_\infty$. The prefix $\mathcal{R}$ denotes subspaces formed by real rational transfer matrices. $\mathcal{RH}_\delta$ denotes the subspace of transfer matrices in $\mathcal{RH}_\infty$ that are analytic outside the disk of radius δ, $0 < \delta < 1$, equipped with the norm $\|G(z)\|_{\mathcal{H}_\delta} \stackrel{\Delta}{=} \sup_{0 \leq \theta \leq \pi} |G(\delta e^{j\theta})|$. l_∞ denotes the space of bounded real sequences $\{e_k\}$ equipped with the norm $\|e\|_{l_\infty} \stackrel{\Delta}{=} \sup_k |e_k|$. l_1 denotes the space of real sequences, equipped with the norm $\|q\|_1 = \sum_{k=0}^{\infty} |q_k| < \infty$. Given a sequence $q \in l_1$ we will denote its Z-transform by $Q(z) \in \mathcal{RH}_\infty$. To avoid confusion and by a slight abuse of notation, we will denote the $\|.\|_\infty$ norm of a transfer function as $\|.\|_{\mathcal{H}_\infty}$[1] and the l_∞ norm of a sequence as $\|.\|_{l_\infty}$.

Throughout the paper we will use packed notation to represent state-space realizations, i.e.,

$$G(z) = C(zI - A)^{-1}B + D \stackrel{\Delta}{=} \left(\begin{array}{c|c} A & B \\ \hline C & D \end{array} \right).$$

$P_+ \colon \mathcal{RL}_\infty \to \mathcal{RH}_\infty$ denotes the projection operator; i.e., given $G \in \mathcal{RL}_\infty$, $\mathcal{G} = P_+(G)$ is the stable part of G. For a transfer function $G(z)$, $G^{\sim} \stackrel{\Delta}{=} G(\frac{1}{z})$. Given $R \in \mathcal{RH}_\infty$, $\Gamma_{\mathrm{H}}(R)$ denotes its maximum Hankel singular value (for $R \in \mathcal{RH}_\infty^-$, $\Gamma_{\mathrm{H}}(R)$ denotes the maximum Hankel singular value of $R^{\sim}$).

Finally, given two transfer matrices $T = \begin{pmatrix} T_{11} & T_{12} \\ T_{21} & T_{22} \end{pmatrix}$ and Q with appropriate dimensions, the lower *linear fractional transformation* is defined as

$$\mathcal{F}_l(T, Q) \stackrel{\Delta}{=} T_{11} + T_{12}Q(I - T_{22}Q)^{-1}T_{21}.$$

[1] Strictly speaking this norm should be denoted as $\|.\|_{\mathcal{L}_\infty}$ since it will be applied to both functions in $\mathcal{H}_\infty$ and $\mathcal{H}_\infty^-$. However, we decided to use the notation $\|.\|_{\mathcal{H}_\infty}$ to avoid confusion with the continuous-time $\mathcal{L}_\infty$ norm.

2.2. Statement of the problem. Consider the SISO system represented by the block diagram in Fig. 1, where S represents the system to be controlled; the scalar signals w, θ, and u represent an exogenous disturbance, a *known, fixed* signal, and the control action, respectively; and ζ, e, and y represent the outputs subject to frequency-domain performance constraints, the tracking error to the signal θ, and the measurements respectively. Note that w and ζ include fictitious signals used to assess stability in the presence of model uncertainty. Then the problem that we address in this paper is the following:

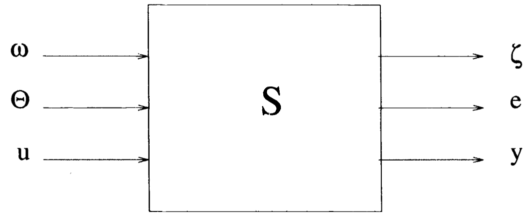


FIG. 1. *The generalized plant.*

Mixed $l_\infty/\mathcal{H}_\infty$ control problem. Given the nominal system (S), with frequency-domain performance specifications of the form

$$(P) \qquad \|W(z)T_{\zeta w}(z)\|_\infty \leq \gamma$$

where $W(z)$ is a suitable weighting function, find an internally stabilizing rational controller

$$(C) \qquad u(z) = K(z)y(z)$$

such that the maximum amplitude of the regulated output e due to θ is minimized subject to the performance specifications (P)

3. Problem solution. In this section we show that the mixed $l_\infty/\mathcal{H}_\infty$ problem can be solved by solving a sequence of problems, each one requiring the solution of a finite-dimensional convex optimization problem and an unconstrained Nehari extension problem.

3.1. Problem transformation. Assume that the system S has the following state-space realization (where without loss of generality we assume that all weighting factors have been absorbed into the plant):

$$(S) \qquad \left(\begin{array}{c|ccc} A & B_{1f} & B_{1t} & B_2 \\ \hline C_f & D_{ff} & D_{ft} & D_{f2} \\ C_t & D_{tf} & D_{tt} & D_{t2} \\ C_2 & D_{2f} & D_{2t} & D_{22} \end{array}\right)$$

where D_{f2} has full column rank, D_{2f} has full row rank, and the pairs (A, B_2) and (C_2, A) are stabilizable and detectable, respectively. It is well known (see for instance

[11]) that the set of all internally stabilizing controllers can be parametrized in terms of a free parameter $Q \in \mathcal{H}_\infty$ as

$$K = \mathcal{F}_l(J, Q) \tag{1}$$

where J has the following state-space realization:

$$(J) \qquad \left(\begin{array}{c|cc} A + B_2F + LC_2 + LD_{22}F & -L & B_2 + LD_{22} \\ \hline F & 0 & I \\ -(C_2 + D_{22}F) & I & -D_{22} \end{array}\right)$$

where F and L are selected such that $A + B_2F$ and $A + LC_2$ are stable. By using this parametrization, the closed-loop transfer functions $T_{\zeta w}$ and $T_{e\theta}$ can be written as

$$\begin{aligned} T_{\zeta w} &= \mathcal{F}_l(T, Q) = T_1 + T_2Q, \\ T_{e\theta} &= \mathcal{F}_l(T_\theta, Q) = T_1^\theta + T_2^\theta Q \end{aligned} \tag{2}$$

where $T_1 \stackrel{\Delta}{=} T_{11}$, $T_2 \stackrel{\Delta}{=} T_{12}T_{21}$, $T_1^\theta \stackrel{\Delta}{=} T_{11}^\theta$, $T_2^\theta \stackrel{\Delta}{=} T_{12}^\theta T_{21}^\theta \in \mathcal{RH}_\infty$, and T and T_θ have the following state-space realizations:

$$T = \left(\begin{array}{cc|cc} A + B_2F & -B_2F & B_{1f} & B_2 \\ 0 & A + LC_2 & B_{1f} + LD_{2f} & 0 \\ \hline C_f + D_{f2}F & -D_{f2}F & D_{ff} & D_{f2} \\ 0 & C_2 & D_{2f} & 0 \end{array}\right),$$

(3)

$$T_\theta = \left(\begin{array}{cc|cc} A + B_2F & -B_2F & B_{1t} & B_2 \\ 0 & A + LC_2 & B_{1t} + LD_{2t} & 0 \\ \hline C_t + D_{t2}F & -D_{t2}F & D_{tt} & D_{t2} \\ 0 & C_2 & D_{2t} & 0 \end{array}\right).$$

Moreover (see for instance [9]), it is possible to select F and L in such a way that $T_2(z)$ is inner (i.e., $T_2\tilde{\ }T_2 = I$). Since $\|.\|_{\mathcal{H}_\infty}$ is invariant under multiplication by an inner function, we have

$$\begin{aligned} \|T_{\zeta w}\|_{\mathcal{H}_\infty} &= \|T_1 + T_2Q\|_{\mathcal{H}_\infty} \\ &= \|T_1T_2\tilde{\ } + Q\|_{\mathcal{H}_\infty} = \|R + Q\|_{\mathcal{H}_\infty} \end{aligned} \tag{4}$$

where $R(z) \stackrel{\Delta}{=} T_1(z)T_2\tilde{\ }(z)$ has all its poles outside the unit disk [9].

By using this parametrization the mixed optimization problem can be now precisely stated as solving

$$(l_\infty/\mathcal{H}_\infty) \qquad \mu^o = \inf_{Q \in \mathcal{RH}_\infty} \|e_k\|_{l_\infty}$$

subject to

$$\|T_1(z) + T_2(z)Q(z)\|_{\mathcal{H}_\infty} \le \gamma \tag{5}$$

where

$$e_k = Z^{-1}\left\{E(z) = (T_1^\theta(z) + T_2^\theta(z)Q(z))\Theta(z)\right\}$$

and $\theta \in l_1{}^2$ is a known, fixed signal.

3.2. l_∞ Optimization analysis. In this section we analyze the l_∞ optimization and show that it can be handled by considering a finite number of constraints. We begin by recalling a result for the l_∞ problem without $\mathcal{H}_\infty$ constraints.

THEOREM 1(Dahleh and Pearson, [7]). *Let $T_2^\theta(z)\Theta(z)$ have n distinct zeros a_k outside the open unit disk. Then*

$$(6)\quad \mu^* = \inf_{K\text{stab}} \|e\|_{l_\infty} = \max_{\alpha_j}\left[\sum_{i=1}^{n} \alpha_i \text{Re}\{T_1^\theta(a_i)\Theta(a_i)\} + \sum_{i=1}^{n} \alpha_{i+n}\text{Im}\{T_1^\theta(a_i)\Theta(a_i)\}\right]$$

subject to

$$(7)\quad \sum_{j=0}^{\infty}\left|\sum_{i=1}^{n}\alpha_i\text{Re}\{a_i^{-j}\} + \sum_{i=1}^{n}\alpha_{i+n}\text{Im}\{a_i^{-j}\}\right| \leq 1.$$

Furthermore, let

$$(8)\quad r_j \stackrel{\Delta}{=} \sum_{i=1}^{n}\alpha_i\text{Re}\{a_i^{-j}\} + \sum_{i=1}^{n}\alpha_{i+n}\text{Im}\{a_i^{-j}\}.$$

Then the optimal error e_k satisfies the following condition:

$$|e_k| = \begin{cases} \mu^*, & \text{if } r_k \neq 0; \\ \leq \mu^*, & \text{if } r_k = 0. \end{cases}$$

Remark 1. Note that the optimal solution may have infinitely many terms such that $|e_i| = \mu^*$.

Since all the solutions to a suboptimal Nehari extension problem of the form $\|R + Q\|_{\mathcal{H}_\infty} \leq \gamma$ can be parametrized in terms of a free parameter $W(z) \in \mathcal{H}_\infty$, $\|W\|_{\mathcal{H}_\infty} \leq \gamma^{-1}$ problem $l_\infty/\mathcal{H}_\infty$ can be thought of as an optimization problem inside the origin centered γ^{-1}-ball. However, the γ^{-1}-ball is not compact in $\mathcal{H}_\infty$. Thus a minimizing solution may not exist. Motivated by this difficulty, we introduce the following *modified* mixed $l_\infty/\mathcal{H}_\infty$ problem: given $\delta < 1$ and $T_1(z), T_2(z)$, and $\Theta(z) \in \mathcal{RH}_\delta$, find[3]

$$(l_\infty/\mathcal{H}_\delta)\qquad \mu_o^\delta = \min_{Q \in \mathcal{RH}_\delta} \|e\|_{l_\infty}$$

subject to

$$\|T_1(z) + T_2(z)Q(z)\|_{\mathcal{H}_\delta} \leq \gamma.$$

Remark 2. Problem $l_\infty/\mathcal{H}_\delta$ can be thought of as solving the problem $l_\infty/\mathcal{H}_\infty$ with the additional constraint that all the poles of the closed-loop system must be inside the disk of radius δ. A parametrization of all achievable closed-loop transfer functions,

[2] This restriction on θ can be relaxed to include steps functions, by absorbing the pole at $z = 1$ into the plant, thus forcing a controller with integral action.

[3] Problem $l_\infty/\mathcal{H}_\delta$ was suggested by Dr. H. Rotstein and Prof. A. Sideris, Department of Electrical Engineering, Caltech.

such that T satisfies this additional constraint, can be obtained from (1) by simply changing the stability region from the unit-disk to the δ-disk using the transformation $z = \delta\hat{z}$ before performing the factorization. Furthermore, by combining this transformation with the inner factorization, the resulting $T_2(z)$ satisfies $T_2(z)T_2(\frac{1}{z})|_{|z|=\delta} = 1$ (i.e., T_2 is inner in $\mathcal{H}_\delta$).

In what follows we will show that if $l_\infty/\mathcal{H}_\delta$ is feasible, it always admits a minimizing solution. Moreover, this solution is rational (i.e., $Q \in \mathcal{RH}_\infty$) and requires considering only a finite number of elements $N(\delta)$ (independent of Q) of the sequence $\{e_k\}$. The proof is constructive and is based upon showing that $l_\infty/\mathcal{H}_\delta$ can be decoupled into a finite-dimensional convex optimization and an unconstrained Nehari approximation problem, both of them admitting a minimizing rational solution.

LEMMA 1. *Assume that $l_\infty/\mathcal{H}_\delta$ is feasible. Then there exists $N(\delta)$ such that for every $Q \in \mathcal{H}_\delta$ satisfying the constraint $\|R+Q\|_{\mathcal{H}_\delta} \le \gamma$, the corresponding sequence $\{e_k\}$ satisfies $|e_k| < \mu^*$ for all $k \ge N$, where μ^* indicates the unconstrained l_∞ optimum introduced in Theorem* 1.

Proof. Proof of Lemma 1 is given in Appendix A. □

COROLLARY 1. *Problem $l_\infty/\mathcal{H}_\delta$ is equivalent to the following semi-infinite convex optimization problem*:

$$\min_{\substack{Q\in\mathcal{H}_\delta \\ \|R+Q\|_{\mathcal{H}_\delta}\le\gamma}} \left\{ \max_{0\le k\le N-1} |(\underline{t}_1 + \tau\underline{q})_k| \right\} \tag{9}$$

where

$$\underline{t}_1 \stackrel{\Delta}{=} (\, t^\theta_{1o} \quad \cdots \quad t^\theta_{1N-1}\,)',$$

$$\tau = \begin{pmatrix} t^\theta_{2o} & 0 & \dots & 0 \\ t^\theta_{21} & t^\theta_{20} & \dots & 0 \\ \vdots & & \ddots & \\ t^\theta_{2N-1} & \dots & & t^\theta_{2o} \end{pmatrix}, \tag{10}$$

$$\underline{q}^o \stackrel{\Delta}{=} (\, q_o \quad \cdots \quad q_{N-1}\,)',$$

where t^θ_{ik}, q_k denote the kth element of the impulse response of $T^\theta_i(z)\Theta(z)$ and $Q(z)$, respectively.

Proof. Let μ^o denote the solution to $l_\infty/\mathcal{H}_\delta$. From Lemma 1 it follows that (since $\mu^o \ge \mu^*$)

$$\min_{\substack{Q\in\mathcal{H}_\delta \\ \|R+Q\|_{\mathcal{H}_\delta}\le\gamma}} \left\{ \sup_k |e_k| \right\} = \mu^o \ge \mu^* > |e_l|,\ l \ge N.$$

Therefore the peak value of $|e_k|$ is achieved for some $k < N$. The proof is completed by noting that (9) gives e_k in terms of the impulse responses of $T^\theta_1\Theta(z), T^\theta_2\Theta(z)$, and $Q(z)$. □

3.3. The $\mathcal{H}_\infty$ performance constraint. In this section we show that i) the modified $l_\infty/\mathcal{H}_\infty$ problem $l_\infty/\mathcal{H}_\delta$ has a global minimum that can be explicitly found by solving a constrained finite-dimensional convex optimization followed by the solution to an unconstrained Nehari approximation problem; and ii) the mixed $l_\infty/\mathcal{H}_\infty$ problem can be solved by solving a sequence of modified problems. Let q_i denote the terms of the impulse response of $Q(z)$. The key observation to show the first result is that (from Lemma 1) only the first N terms of this expansion appear in the l_∞ optimization. The second result follows then by constructing a nonincreasing sequence $\{\mu_i\}$.

LEMMA 2. *Consider the following Sylvester equation:*

$$A_R' Y A_q - Y = c_R' e_N' \tag{11}$$

where A_R is a nonsingular antistable matrix, c_R is a row vector, and

$$\begin{aligned} A_q &= \begin{pmatrix} 0 & \dots & 0 & 0 \\ 1 & \dots & 0 & 0 \\ \vdots & \ddots & \vdots & \vdots \\ 0 & \dots & 1 & 0 \end{pmatrix}, \\ e_N' &= \begin{pmatrix} 0 & \dots & 1 \end{pmatrix}. \end{aligned} \tag{12}$$

Then the solution Y to (11) *can be explicitly calculated as*

$$Y = -\begin{pmatrix} (A_R')^{N-1} c_R' & (A_R')^{(N-2)} c_R' & \dots & c_R' \end{pmatrix}. \tag{13}$$

Proof. The proof follows by successive right multiplications of (11) by the columns of the identity. □

In the following theorem we consider the case where the *first* N coefficients of the expansion of $Q(z)$ are *specified* and establish a necessary and sufficient condition for the existence of a tail to complete $Q(z)$ in such a way that the approximation error verifies $\|R + Q\|_{\mathcal{H}_\infty} \le \gamma$.

THEOREM 2. *Let*

$$R \stackrel{\Delta}{=} \left(\begin{array}{c|c} A_R & b_R \\ \hline c_R & d_R \end{array} \right) \in \mathcal{RH}_\infty^-,$$

with McMillan degree n, and $Q_F = \sum_{i=0}^{N-1} q_i z^{-i}$ be given. Then there exist $Q_R \in \mathcal{RH}_\infty$, such that $\|R + Q_F + z^{-N} Q_R\|_{\mathcal{H}_\infty} \le \gamma$, iff $\|\mathcal{Q}\|_2 \stackrel{\Delta}{=} \overline{\sigma}(\mathcal{Q}) \le \gamma$, where $\overline{\sigma}$ denotes the maximum singular value, $\mathcal{Q}$ is a matrix affine in the coefficients of Q_F of the following form:

$$\begin{aligned} \mathcal{Q} &= W^{\frac{1}{2}} \begin{pmatrix} I & 0 \\ 0 & \mathcal{H}' \end{pmatrix} L_c^{\frac{1}{2}}, \\ \mathcal{H} &= \begin{pmatrix} h_N & h_{N-1} & \dots & \dots & h_1 \\ & h_N & h_{N-1} & \dots & h_2 \\ & & \ddots & & \\ & & & h_N & h_{N-1} \\ & & & & h_N \end{pmatrix}, \\ h_i &= q_{N-i} + b_R' (A_R')^{N-1-i} c_R', \ 1 \le i \le N-1, \\ h_N &= q_o + d_R; \end{aligned} \tag{14}$$

and W and L_c are positive definite matrices depending only on R.

Proof. Let $G \stackrel{\Delta}{=} R + Q_F$. Given Q_F, there exist $Q_R \in \mathcal{H}_\infty$ such that $\|R + Q_F + z^{-N} Q_R\|_{\mathcal{H}_\infty} \le \gamma$ *iff* the corresponding *unconstrained* 1-block Nehari approximation problem [12] has a solution, i.e. if

$$\begin{aligned} \min_{Q_R \in \mathcal{H}_\infty} \|G + z^{-N} Q_R\|_{\mathcal{H}_\infty} &= \min_{Q_R \in \mathcal{H}_\infty} \|z^N G + Q_R\|_{\mathcal{H}_\infty} \\ &= \min_{Q_R \in \mathcal{H}_\infty} \|z^{-N} G^\sim + Q_R^\sim\|_{\mathcal{H}_\infty} \\ &= \Gamma_H(z^{-N} G^\sim) \le \gamma, \end{aligned} \tag{15}$$

where we used the facts that z^N is an inner function and $\|G\|_{\mathcal{H}_\infty} = \|G^\sim\|_{\mathcal{H}_\infty}$. Moreover, from Nehari's theorem it also follows that $Q_R \in \mathcal{RH}_\infty$. In order to compute Γ_H we need a state-space realization for the stable part of $z^{-N}G^\sim$. Let $G_1 \overset{\Delta}{=} R^\sim z^{-N}$. Standard space-state manipulations [13] yield

$$(16) \qquad R^\sim = \left(\begin{array}{c|c} (A_R')^{-1} & -(A_R')^{-1}c_R' \\ \hline b_R'(A_R')^{-1} & d_R' - b_R'(A_R')^{-1}c_R' \end{array}\right), \qquad z^{-N} = \left(\begin{array}{c|c} A_q & e_1 \\ \hline e_N' & 0 \end{array}\right)$$

where A_R^{-1} exists since $R \in \mathcal{H}_\infty{}^-$ and

$$A_q = \begin{pmatrix} 0 & \dots & 0 & 0 \\ 1 & \dots & 0 & 0 \\ \vdots & \ddots & \vdots & \vdots \\ 0 & \dots & 1 & 0 \end{pmatrix}, \qquad e_1 = \begin{pmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix},$$
$$e_N' = (\, 0 \quad \dots \quad 1 \,).$$

Hence

$$(17) \qquad G_1 = \left(\begin{array}{cc|c} (A_R')^{-1} & -(A_R')^{-1}c_R'e_N' & 0 \\ 0 & A_q & e_1 \\ \hline b_R'(A_R')^{-1} & d_R'e_N' - b_R'(A_R')^{-1}c_R'e_N' & 0 \end{array}\right).$$

Finally, the similarity transformation

$$T = \begin{pmatrix} I_n & Y \\ 0 & I_N \end{pmatrix}$$

where Y is the unique solution to the Sylvester equation

$$A_R'YA_q - Y = c_R'e_N'$$

yields

$$(18) \qquad G_1 = \left(\begin{array}{cc|c} (A_R')^{-1} & 0 & Ye_1 \\ 0 & A_q & e_1 \\ \hline b_R'(A_R')^{-1} & d_R'e_N' - b_R'(A_R')^{-1}(c_R'e_N' + Y) & 0 \end{array}\right).$$

Since A_R is antistable, A_R^{-1} is stable. Hence $P_+[G_1] = G_1$. Similarly,

$$(19) \qquad \begin{aligned} G_2 \overset{\Delta}{=} z^{-N}Q_F{}^\sim &= \sum_{i=1}^{N-1} q_{N-i}z^{-i} \\ &= \left(\begin{array}{c|c} A_q & e_1 \\ \hline c_q & 0 \end{array}\right) \end{aligned}$$

where

$$c_q = (\, q_{N-1} \quad \dots \quad q_0 \,).$$

Hence

$$\mathcal{G}\overset{\Delta}{=}P_+[G_1+G_2]=\left(\begin{array}{cc|c} (A'_R)^{-1} & 0 & Ye_1 \\ 0 & A_q & e_1 \\ \hline b'_R(A'_R)^{-1} & H & 0 \end{array}\right) \tag{20}$$

where

$$H\overset{\Delta}{=}c_q+d'_Re'_N-b'_R(A'_R)^{-1}(c'_Re'_N+Y)\overset{\Delta}{=}(\,h_1 \quad \dots \quad h_N\,)\,. \tag{21}$$

Finally, note that Y can be computed explicitly by using Lemma 2. Substituting (13) in (20) and (21) yields

$$\begin{aligned} \mathcal{G}=P_+[G_1+G_2]&=\left(\begin{array}{cc|c} (A'_R)^{-1} & 0 & -(A'_R)^{N-1}c'_R \\ 0 & A_q & e_1 \\ \hline b'_R(A'_R)^{-1} & H & 0 \end{array}\right); \\ h_i&=q_{N-i}+b'_R(A'_R)^{N-1-i}c'_R, \qquad 1\le i\le N-1; \\ h_N&=q_0+d_R. \end{aligned} \tag{22}$$

In order to compute the approximation error we need to compute the observability and controllability grammians of $\mathcal{G}$. Although, in principle, this requires the solution of two Lyapunov equations, with coefficients that are functions of Q_F, we will show that the particular structure of the problem allows for computing these solutions explicitly. For the controllability grammian L_c we have

$$\begin{aligned} &\begin{pmatrix} (A'_R)^{-1} & 0 \\ 0 & A_q \end{pmatrix}\begin{pmatrix} L^C_{11} & L^C_{12} \\ L'^C_{12} & L^C_{22} \end{pmatrix}\begin{pmatrix} (A'_R)^{-1} & 0 \\ 0 & A_q \end{pmatrix}'-\begin{pmatrix} L^C_{11} & L^C_{12} \\ L'^C_{12} & L^C_{22} \end{pmatrix} \\ &\qquad\qquad=-\begin{pmatrix} (A'_R)^{N-1}c'_Rc_R(A_R)^{N-1} & -(A'_R)^{N-1}c'_Re'_1 \\ -e_1c_R(A_R)^{N-1} & e_1e'_1 \end{pmatrix}. \end{aligned} \tag{23}$$

Solving for each of the blocks of the grammian yields

$$\begin{aligned} L^C_{11}&=L^C_o, \\ L^C_{12}&=-\left(\,(A'_R)^{N-1}c'_R \quad (A'_R)^{N-2}c'_R\dots \quad c'_R\,\right)=Y, \\ L^C_{22}&=I_N \end{aligned} \tag{24}$$

where L^C_o is the solution of the following Lyapunov equation:

$$A'_RL^C_oA_R-L^C_o=(A'_R)^Nc'_Rc_R(A_R)^N \tag{25}$$

and the expression for L^C_{12} was obtained from the corresponding equation by successive right multiplications by e_i. Note that the controllability grammian of $\mathcal{G}$ is independent of Q_F. Similarly, for the observability grammian L_o we have

$$\begin{aligned} &\begin{pmatrix} (A'_R)^{-1} & 0 \\ 0 & A_q \end{pmatrix}'\begin{pmatrix} L^0_{11} & L^0_{12} \\ L'^0_{12} & L^0_{22} \end{pmatrix}\begin{pmatrix} (A'_R)^{-1} & 0 \\ 0 & A_q \end{pmatrix}-\begin{pmatrix} L^0_{11} & L^0_{12} \\ L'^0_{12} & L^0_{22} \end{pmatrix} \\ &\qquad\qquad=-\begin{pmatrix} (A_R)^{-1}b_Rb'_R(A'_R)^{-1} & (A_R)^{-1}b_rH \\ H'b'_R(A'_R)^{-1} & H'H \end{pmatrix}. \end{aligned} \tag{26}$$

Solving for each of the blocks of the grammian yields

$$\begin{aligned} L^0_{11} &= L^0_o, \\ L^0_{12} &= \mathcal{A}\mathcal{H}', \\ L^0_{22} &= \mathcal{H}\mathcal{H}' \end{aligned} \tag{27}$$

where

$$\mathcal{H} \stackrel{\Delta}{=} \begin{pmatrix} h_N & h_{N-1} & \dots & \dots & h_1 \\ & h_N & h_{N-1} & \dots & h_2 \\ & & \ddots & & \\ & & & h_N & h_{N-1} \\ & & & & h_N \end{pmatrix}, \tag{28}$$
$$\mathcal{A} = \left(A_R^{-N} b_R \quad A_R^{-(N-1)} b_R \dots A_R^{-1} b_R \right),$$

and L^0_o is the solution to the following Lyapunov equation:

$$A_R L^0_o A'_R - L^0_o = b_R b'_R \tag{29}$$

(i.e., the controllability grammian for R), which is independent from Q_F. Finally, note that

$$\begin{aligned} L_o &= \begin{pmatrix} L^0_o & \mathcal{A}\mathcal{H}' \\ \mathcal{H}\mathcal{A}' & \mathcal{H}\mathcal{H}' \end{pmatrix} \\ &= \begin{pmatrix} I & 0 \\ 0 & \mathcal{H} \end{pmatrix} \begin{pmatrix} L^0_o & \mathcal{A} \\ \mathcal{A}' & I \end{pmatrix} \begin{pmatrix} I & 0 \\ 0 & \mathcal{H}' \end{pmatrix}. \end{aligned} \tag{30}$$

Let

$$W'^{\frac{1}{2}} W^{\frac{1}{2}} \stackrel{\Delta}{=} \begin{pmatrix} L^0_o & \mathcal{A} \\ \mathcal{A}' & I \end{pmatrix}. \tag{31}$$

Then

$$L_o = \begin{pmatrix} I & 0 \\ 0 & \mathcal{H} \end{pmatrix} W'^{\frac{1}{2}} W^{\frac{1}{2}} \begin{pmatrix} I & 0 \\ 0 & \mathcal{H}' \end{pmatrix}. \tag{32}$$

Hence, from (24) and (32) we have that

$$\begin{aligned} L_c^{\frac{1}{2}} L_o L_c^{\frac{1}{2}} &= \mathcal{Q}'\mathcal{Q} \\ \mathcal{Q} &\stackrel{\Delta}{=} W^{\frac{1}{2}} \begin{pmatrix} I & 0 \\ 0 & \mathcal{H}' \end{pmatrix} L_c^{\frac{1}{2}}. \end{aligned} \tag{33}$$

From Nehari's theorem [12] it follows that

$$\begin{aligned} \|R + Q\|_{\mathcal{H}_\infty} \le \gamma &\iff \rho^{\frac{1}{2}} \left(L_c^{\frac{1}{2}} L_o L_c^{\frac{1}{2}} \right) \le \gamma \\ &\iff \|\mathcal{Q}\|_2 \le \gamma \end{aligned} \tag{34}$$

where ρ indicates the spectral radius. □

Remark 3. Note that

$$\begin{pmatrix} L_o^0 & \mathcal{A} \\ \mathcal{A}' & I \end{pmatrix}$$

is positive definite, since from (29) it can be easily shown that

$$L_o^0 - \mathcal{A}\mathcal{A}' = A_R{}^{-N} L_o^0 A_R^{-N} > 0. \tag{35}$$

Remark 4. Although the similarity transformation used to diagonalize (17) is useful in establishing the desired result, attempting to compute $\mathcal{Q}$ from (33) may be numerically ill conditioned since the matrices L_c and $\mathcal{H}$ contain the powers A_R^i, $i = 1 \dots N$, and A_R all its eigenvalues outside the unit disk. An alternative expression for $\mathcal{Q}$ that avoids this ill-conditioning is given in Appendix B.

Remark 5. This result can be easily extended to the case $Q \in \mathcal{RH}_\delta$, $\|R+Q\|_{\mathcal{H}_\delta} \leq \gamma$ by using the change of variable $z = \delta\hat{z}$.

Combining Corollary 1 and Theorem 2 yields the main result of this section.

THEOREM 3. *$Q^o = Q_F^o + \hat{z}^{-N} Q_R^o$, where $Q_F^o = \sum_{i=0}^{N-1} q_i \hat{z}^{-i}$, solves the modified $l_\infty/\mathcal{H}_\infty$ control problem $l_\infty/\mathcal{H}_\delta$ iff $\underline{q}^o = (q_o \dots q_{N-1})'$ solves the following finite-dimensional convex optimization problem:*

$$\underline{q}^o = \underset{\substack{\underline{q} \in R^N \\ \|\mathcal{Q}\|_2 \leq \gamma}}{\operatorname{argmin}} \left\{ \max_{0 \leq k \leq N-1} |(\underline{t}_1 + \tau \underline{q})_k| \right\} \tag{36}$$

and Q_R^o solves the unconstrained Nehari approximation problem

$$Q_R^o(\hat{z}) = \underset{Q_R \in \mathcal{H}_\infty}{\operatorname{argmin}} \|R(\hat{z}) + Q_F^o + \hat{z}^{-N} Q_R(\hat{z})\|_{\mathcal{H}_\infty} \tag{37}$$

where $z = \delta\hat{z}$, R is defined in (4), N is selected such that

$$\begin{aligned} &K_e \delta^N \leq \mu^*, \\ &K_e = \left(\|T_1^\theta(\hat{z})\|_{\mathcal{H}_\infty} + \|T_2^\theta(\hat{z})\|_{\mathcal{H}_\infty} \left(\gamma + \|R(\hat{z})\|_{\mathcal{H}_\infty} \right) \right) \|\Theta(\hat{z})\|_{\mathcal{H}_\infty}, \end{aligned} \tag{38}$$

and μ^* is the unconstrained l_∞ optimum.

Remark 6. $Q^o \in \mathcal{RH}_\infty$ since Q_F^o is a finite impulse response filter and, from Nehari's theorem, $Q_R^o \in \mathcal{RH}_\infty$.

4. Synthesis algorithm. Based upon the results of Theorem 3 and §§3.2 and 3.3, the mixed $l_\infty/\mathcal{H}_\infty$ problem can be solved by using the following algorithm:

0) *Data:* An increasing sequence $\delta_i \to 1, \epsilon > 0$.
1) Solve the unconstrained l_∞ problem using Theorem 1. Compute $\|T_{\zeta w}\|_{\mathcal{H}_\infty}$. If $\|T_{\zeta w}\|_{\mathcal{H}_\infty} \leq \gamma$ stop, else set $i = 1$.
2) Solve the problem $l_\infty/\mathcal{H}_{\delta i}$ proceeding as follows:
   2.1) Let $z = \delta_i \hat{z}$ and consider the system $S(\hat{z})$.
   2.2) Perform the factorization (1) to obtain $T_i(\hat{z}), T_i^\theta(\hat{z})$.
   2.3) Compute N from (38).
   2.4) Find $\hat{Q}(\hat{z})$ and μ_i using Theorem 3.
3) Let $Q = \hat{Q}(\frac{z}{\delta_i}), K = F_l(J, Q)$. Compute $\|T_{\zeta w}(z)\|_{\mathcal{H}_\infty}$. If $\|T_{\zeta w}(z)\|_{\mathcal{H}_\infty} \geq \gamma - \epsilon$ stop, else set $i = i + 1$ and go to 2.

Step 2.4) entails solving first the nondifferentiable constrained optimization problem (36) and then the unconstrained $\mathcal{H}_\infty$ problem (37). The latter is a well-understood problem, and efficient computation techniques are available for finding the solution (see for instance [5]). The finite-dimensional problem (36) can be solved by using the deep-cut ellipsoid method [14, p. 329]. This algorithm requires computing subgradients of the objective function and the constraint $\overline{\sigma}(\mathcal{Q}) \leq \gamma$. Since the objective function is affine in Q, a subgradient is readily available. From the explicit expression for $\mathcal{Q}$ given in Appendix B. (eq. (B4)), it can be easily shown that a subgradient of the constraint is given by

$$\frac{\partial \overline{\sigma}}{\partial q_i} = (S_R^i v)u \tag{39}$$

where u and v are right and left singular vectors of $\mathcal{Q}$ corresponding to $\overline{\sigma}$ and S_R^i indicates the shift right operator applied i times, i.e.,

$$S_R^i v \overset{\Delta}{=} (0 \quad 0 \quad \dots \quad v_1 \quad \dots v_{n-i}).$$

THEOREM 4. *Assume that* $\inf_{Q \in \mathcal{RH}_\infty} \|T_1 + T_2 Q\|_{\mathcal{H}_\infty} = \Gamma_H(R) < \gamma$. *Then the sequence* $\mu_i \rightarrow \mu^o$, *the solution to the mixed* $l_\infty/\mathcal{H}_\infty$ *problem.*

Proof. From the maximum modulus theorem it follows that the solution Q_i to $l_\infty/\mathcal{H}_{\delta i}$ is a feasible solution for $l_\infty/\mathcal{H}_{\delta i+1}$. Thus, the sequence μ_i is nonincreasing, bounded below by the value of the unconstrained l_∞ controller. Therefore the sequence has a limit $\mu \geq \mu^o$. We will show next that $\mu = \mu^o$. Assume by contradiction that $\mu^o < \mu$ and select $\mu^o < \hat{\mu} < \mu$. Since $\Gamma_H(R) < \gamma$, there exists $Q_1 \in \mathcal{RH}_\infty$ such that $\|R + Q_1\|_{\mathcal{H}_\infty} < \gamma$. From the definition of μ^o it follows that, given $\eta > 0$, there exists $Q_o \in \mathcal{RH}_\infty$, $\|R + Q_o\|_{\mathcal{H}_\infty} \leq \gamma$, such that $\|e(Q_o)\|_{l_\infty} \leq \mu^o + \eta$, where $e(Q)$ denotes the output corresponding to the controller Q. Let $E(Q)$ denote the z-transform of $e(Q)$ and define $\hat{Q} \overset{\Delta}{=} Q_o + \epsilon(Q_1 - Q_o)$. Then

$$\|R + \hat{Q}\|_{\mathcal{H}_\infty} \leq (1-\epsilon)\|R + Q_o\|_{\mathcal{H}_\infty} + \epsilon\|R + Q_1\|_{\mathcal{H}_\infty} < \gamma$$

and

$$\begin{aligned} e_k(\hat{Q}) - e_k(Q_o) &= \frac{1}{2\pi j}\oint_{|z|=1} (E(\hat{Q}) - E(Q_o))z^{k-1}dz, \\ |e_k(\hat{Q}) - e_k(Q_o)| &\leq \frac{1}{2\pi}\oint_{|z|=1} |E(\hat{Q}) - E(Q_o)|dz \\ &\leq \|E(\hat{Q}) - E(Q_o)\|_{\mathcal{H}_\infty} = \epsilon\|T_2^\theta(Q_1 - Q_o)\Theta\|_{\mathcal{H}_\infty}. \end{aligned} \tag{40}$$

Hence

$$|e_k(\hat{Q})| \leq |e_k(Q_o)| + \epsilon\|T_2^\theta(Q_1 - Q_o)\Theta\|_{\mathcal{H}_\infty}.$$

It follows that

$$\begin{aligned} \|e(\hat{Q})\|_{l_\infty} &= \sup_k |e_k(\hat{Q})| \\ &\leq \sup_k |e_k(Q_o)| + \epsilon\|T_2^\theta(Q_1 - Q_o)\Theta\|_{\mathcal{H}_\infty} \\ &\leq \mu^o + \eta + \epsilon\|T_2^\theta(Q_1 - Q_o)\Theta\|_{\mathcal{H}_\infty}. \end{aligned}$$

Since $\hat{Q} \in \mathcal{RH}_\infty$, there exists $\delta_1 < 1$ such that $T_1 + T_2\hat{Q}$ is analytic in $|z| \geq \delta_1$. Moreover, since $\|T_1 + T_2\hat{Q}\|_{\mathcal{H}_\infty} < \gamma$, it follows (from continuity) that there exists $\delta_2 < 1$ such that $\|T_1 + T_2\hat{Q}\|_{\mathcal{H}_{\delta_2}} \leq \gamma$. Therefore, by taking ϵ and η small enough and $\delta \stackrel{\Delta}{=} \max\{\delta_1, \delta_2\} < 1$ we have that $\|T_1 + T_2\hat{Q}\|_{\mathcal{H}_\delta} \leq \gamma$ and $\|e(\hat{Q})\|_{l_\infty} < \hat{\mu}$. Hence for $\delta_i \geq \delta$, $\mu_i \leq \hat{\mu}$. However, this contradicts the fact that the sequence μ_i is nonincreasing and that $\hat{\mu} < \mu = \lim_{\delta_i \to 1} \mu_i$. □

Remark 7. Theorem 4 shows that as $\delta \to 1$ the objective μ_i of the modified problem converges to the solution μ of the original problem. Moreover, μ_i is a nonincreasing sequence of upper bounds of μ. Note, however, that as $\delta \to 1$, $N(\delta)$ will in general increase, thus increasing the size of the constrained optimization problem (36).

5. A simple example. Consider the problem of minimizing the step response error for the nonminimum phase plant used in [7], subject to robust stability against the unstructured multiplicative uncertainty shown in Fig. 2. Table 1 shows $\|e\|_{l_\infty}$ and $\|T_{\zeta w}\|_{\mathcal{H}_\infty}$ for different designs, with the corresponding step and frequency responses shown in Fig. 3. From Theorem 1 it can be easily shown that the infimum of the error is $\|e\|_{l_\infty} = \frac{8}{3}$, achieved with the controller $C(z) = \frac{z-1}{z}$. The same controller yields $\|T_{\zeta w}\|_{\mathcal{H}_\infty} = 5$, thus guaranteeing robust stability against unstructured perturbations $\|\Delta\|_{\mathcal{H}_\infty} \leq 0.2$. Note that this controller is *not internally stabilizing* due to the pole-zero cancellation at $z = 1$. The optimal $\mathcal{H}_\infty$ controller yields $\|T_{\zeta w}\|_{\mathcal{H}_\infty} = 3$ and $\|e\|_{l_\infty} = 4$. Mixed $l_\infty/\mathcal{H}_\infty$ optimization with $\|T_{\zeta w}\|_{\mathcal{H}_\infty} \leq 3.3$ yields $\|e\|_{l_\infty} = 3.31$. However, this procedure results in a controller with 104 states (since it can be easily shown from Appendix A that the l_∞ optimization needs to consider no more than 50 steps). Finally, the last entry in Table 1 corresponds to a reduced-order controller with five states. In spite of the substantial order reduction, this controller yields virtually the same performance as the mixed $l_\infty/\mathcal{H}_\infty$ controller.

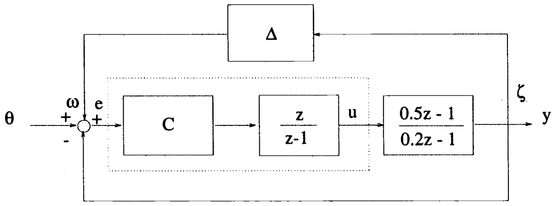


FIG. 2. *Block diagram with multiplicative uncertainty Δ "pulled-out."*

TABLE 1
$\|T_{\zeta w}\|_{\mathcal{H}_\infty}$ *vs* $\|e\|_{l_\infty}$ *for the example.*

	$\|T_{\zeta w}\|_{\mathcal{H}_\infty}$	$\|e\|_{l_\infty}$
l_∞	5	$\frac{8}{3}$
$\mathcal{H}_\infty$	3	4
$l_\infty/\mathcal{H}_\infty$	3.3	3.311
$l_\infty/\mathcal{H}_{\infty_{\mathrm{red}}}$	3.3	3.312

6. Conclusions. In this paper we address the problem of finding an internally stabilizing compensator that minimizes the maximum amplitude of the error to a fixed

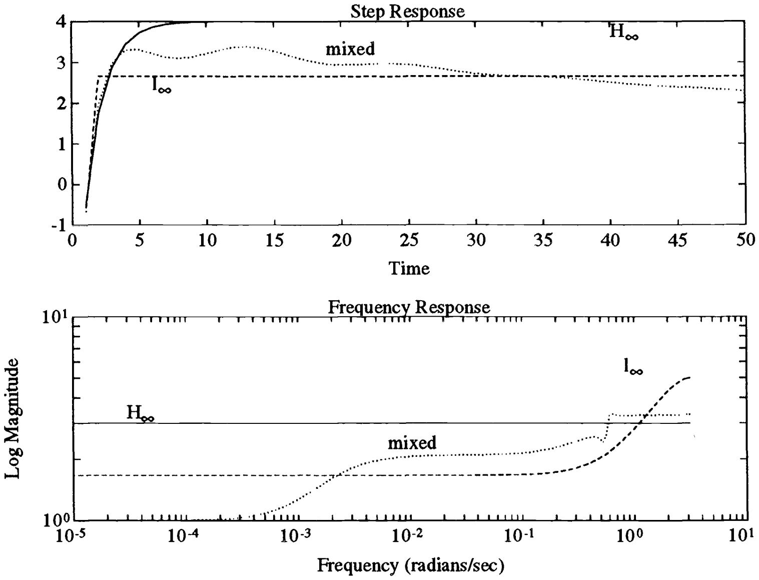


FIG. 3. *Step and frequency responses for different designs.*

given input subject to constraints upon the $\mathcal{H}_\infty$ norm of a relevant transfer function. This problem can be thought of as the problem of designing a controller capable of guaranteeing an adequate robustness level against dynamic uncertainty while using the extra available degrees of freedom to optimize a time-domain performance. We show that the problem can be solved by solving a sequence of modified problems, each one entailing solving a finite-dimensional constrained optimization problem and an unconstrained $\mathcal{H}_\infty$ problem.

Perhaps the most severe limitation of the proposed method is that may result in very-large-order controllers (roughly $2N$) necessitating some type of model reduction. The example of §5 suggests that substantial order reduction can be accomplished without performance degradation. Research is currently under way addressing this issue and pursuing the extension of the formalism to allow more control on the shape of the error response.

Acknowledgments. The author wishes to thank Dr. Héctor Rotstein and Prof. Athanasios Sideris, Department of Electrical Engineering, Caltech, for many comments and for suggesting problem $l_\infty/\mathcal{H}_\delta$.

Appendix A: Proof of Lemma 1. Since $E(z) \in \mathcal{RH}_\delta$, it is analytic in $|z| \geq \delta$ and

$$e_k = \frac{1}{2\pi j}\oint_C E(z)z^{k-1}dz \tag{A1}$$

where C is the origin centered circle with radius δ. From (A1) it follows that, for any $K_e \geq \sup_{z=\delta e^{j\theta}} |E(z)| = \|E\|_{\mathcal{H}_\delta}$, we have

$$|e_k| \leq K_e \delta^k. \tag{A2}$$

An upper bound of $\|E\|_{\mathcal{H}_\delta}$ can be found from (2) as follows. Since $\|.\|_{\mathcal{H}_\delta}$ is submultiplicative, we have

$$\begin{aligned}\|E(z)\|_{\mathcal{H}_\delta} &\leq \|T_{e\theta}\|_{\mathcal{H}_\delta}\|\Theta\|_{\mathcal{H}_\delta} \\ &\leq (\|T_1^\theta\|_{\mathcal{H}_\delta} + \|T_2^\theta\|_{\mathcal{H}_\delta}\|Q^o\|_{\mathcal{H}_\delta})\|\Theta\|_{\mathcal{H}_\delta}.\end{aligned} \tag{A3}$$

From the hypothesis we have that

$$\|Q\|_{\mathcal{H}_\delta} \leq \gamma + \|R\|_{\mathcal{H}_\delta} \stackrel{\Delta}{=} \gamma_q. \tag{A4}$$

Substitution of (A4) in (A3) yields

$$K_e = \left(\|T_1^\theta\|_{\mathcal{H}_\delta} + \|T_2^\theta\|_{\mathcal{H}_\delta}\gamma_q\right)\|\Theta\|_{\mathcal{H}_\delta} < \infty. \tag{A5}$$

It follows that if N is selected such that

$$K_e\delta^N < \mu^*, \tag{$A6$}$$

then, for $l > N$, $|e_l| < \mu^*$.

Appendix B: Some numerical considerations. In this appendix we give an alternative to (33) for computing $\mathcal{Q}$. Since this alternative expression does not involve increasing powers of A_R, it is preferable in cases where N is large or A_R has large eigenvalues. From (24) we have that

$$L_c = \begin{pmatrix} L_o^c & Y \\ Y' & I_N \end{pmatrix} = \begin{pmatrix} I & Y \\ 0 & I \end{pmatrix}\begin{pmatrix} W_{oR} & 0 \\ 0 & I \end{pmatrix}\begin{pmatrix} I & 0 \\ Y' & I \end{pmatrix} \tag{B1}$$

where $W_{oR} \stackrel{\Delta}{=} L_o^c - YY'$ satisfies

$$A'_R W_{oR} A_R - W_{oR} = c'_R c_R$$

(i.e., W_{oR} is the observability grammian of A_R). From (30) we have

$$\begin{aligned} L_o &= \begin{pmatrix} I & 0 \\ 0 & \mathcal{H} \end{pmatrix}\begin{pmatrix} L_o^0 & \mathcal{A} \\ \mathcal{A}' & I \end{pmatrix}\begin{pmatrix} I & 0 \\ 0 & \mathcal{H}' \end{pmatrix} \\ &= \begin{pmatrix} I & \mathcal{A} \\ 0 & \mathcal{H} \end{pmatrix}\begin{pmatrix} L_o^0 - \mathcal{A}\mathcal{A}' & 0 \\ 0 & I \end{pmatrix}\begin{pmatrix} I & 0 \\ \mathcal{A}' & \mathcal{H}' \end{pmatrix}. \end{aligned} \tag{B2}$$

Finally by using (35) we get

$$\begin{aligned} L_o &= \begin{pmatrix} I & \mathcal{A} \\ 0 & \mathcal{H} \end{pmatrix}\begin{pmatrix} A_R^{-N} L_o^0 A_R'^{-N} & 0 \\ 0 & I \end{pmatrix}\begin{pmatrix} I & 0 \\ \mathcal{A}' & \mathcal{H}' \end{pmatrix} \\ &= \begin{pmatrix} A_R^{-N} & \mathcal{A} \\ 0 & \mathcal{H} \end{pmatrix}\begin{pmatrix} L_o^0 & 0 \\ 0 & I \end{pmatrix}\begin{pmatrix} A_R'^{-N} & 0 \\ \mathcal{A}' & \mathcal{H}' \end{pmatrix}. \end{aligned} \tag{B3}$$

Since the spectral radius of L_oL_c is invariant under a similarity transformation, it follows that $\mathcal{Q}$ in (33) can be replaced by

$$\mathcal{Q} = \begin{pmatrix} L_o^{0\frac{1}{2}} & 0 \\ 0 & I \end{pmatrix}\begin{pmatrix} A_R'^{-N} & A_R'^{-N}Y \\ \mathcal{A}' & \mathcal{A}'Y + \mathcal{H}' \end{pmatrix}\begin{pmatrix} W_{oR}^{\frac{1}{2}} & 0 \\ 0 & I \end{pmatrix} \tag{B4}$$

where the only terms that contain powers $A_R^i, i = 1\ldots N$, are in $\mathcal{H}' + \mathcal{A}'Y$. Finally, note that by defining $\hat{H} \stackrel{\Delta}{=} H + b'_R(A'_R)^{-1}Y + b'_R(A'_R)^{-1}c'_R e'_N = (\hat{h}_1 \ldots \hat{h}_N)$, (22) yields

$$\begin{aligned} \hat{h}_i &= q_{N-i}, \qquad 1 \leq i \leq N-1; \\ \hat{h}_N &= q_0 + d_R. \end{aligned} \tag{B5}$$

Hence, we have that

$$
\tag{B6}
\begin{aligned}
\mathcal{H}' + \mathcal{A}'Y = &\begin{pmatrix}
\hat{h}_N & \hat{h}_{N-1} & \cdots & \cdots & \hat{h}_1 \\
 & \hat{h}_N & \hat{h}_{N-1} & \cdots & \hat{h}_2 \\
 & & \ddots & & \\
 & & & \hat{h}_N & \hat{h}_{N-1} \\
 & & & & \hat{h}_N
\end{pmatrix} \\
+ &\begin{pmatrix}
c_R A_R^{-1} b_R & c_R A_R^{-2} b_R & \cdots & \cdots & c_R A_R^{-N} b_R \\
 & c_R A_R^{-1} b_R & c_R A_R^{-2} b_R & \cdots & c_R A_R^{-(N-1)} b_R \\
 & & \ddots & & \\
 & & & c_R A_R^{-1} b_R & c_R A_R^{-2} b_R \\
 & & & & c_R A_R^{-1} b_R
\end{pmatrix}.
\end{aligned}
$$

Since (B6) does not contain increasing powers of A_R, it is preferable to (33), especially in cases where N or the spectral radius of A_R are large.

REFERENCES

[1] S. Boyd, V. Balakrishnan, C. H. Barratt, N. M. Khraishi, X. M. Li, D. G. Meyer, and S. A. Norman, *A new CAD method and associated architectures for linear controllers,* IEEE Trans. Automat. Control, 33 (1988), pp. 268–283.

[2] E. Polak and S. Salcudean, *On the design of linear multivariable feedback systems via constrained nondifferentiable optimization in $\mathcal{H}_\infty$ spaces,* IEEE Trans. Automat. Control, 34 (1989), pp. 268–276.

[3] J. W. Helton and A. Sideris, *Frequency response algorithms for $\mathcal{H}_\infty$ optimization with time domain constraints,* IEEE Trans. Automat. Control, 34 (1989), pp. 427–434.

[4] A. Sideris and H. Rotstein, *$\mathcal{H}_\infty$ optimization with time domain constraints over a finite horizon,* Proc. 29th IEEE CDC, Hawaii, December 5–7, 1990, pp. 1802–1807.

[5] H. Rotstein, *Constrained $\mathcal{H}_\infty$-optimization for discrete-time control,* Ph.D. Dissertation, California Institute of Technology, Pasadena, CA, 1992.

[6] M. A. Dahleh and J. B. Pearson, *l_1-optimal feedback controllers for MIMO discrete-time systems,* IEEE Trans. Automat. Control, AC-32 (1987), pp. 314–322.

[7] ———, *Minimization of a regulated response to a fixed input,* IEEE Trans. Automat. Control, AC-33 (1988), pp. 924–930.

[8] M. Sznaier and A. Sideris, *Suboptimal norm based robust control of constrained systems with an $\mathcal{H}_\infty$ cost,* Proc. 30th IEEE CDC, Brighton, England, December 11–13, 1991, pp. 2280–2286.

[9] M. Sznaier, *A mixed $l_\infty/\mathcal{H}_\infty$ approach to robust controller design,* Proc. 1992 American Control Conference, Chicago, IL, June 24–26, pp. 727–732.

[10] ———, *Robust controller design for the benchmark problem using a mixed $l_\infty/\mathcal{H}_\infty$ approach,* Proc. 1992 American Control Conference, Chicago, IL, June 24–26, pp. 2059–2060.

[11] J. Doyle, *Lecture Notes in Advances in Multivariable Control,* ONR/Honeywell Workshop, Minneapolis, MN, 1984.

[12] B. Francis, *A Course in H_∞ Optimization Theory,* Lecture Notes in Control and Inform. Sci. #88, Springer-Verlag, New York, 1987.

[13] K. Zhou and J. Doyle, *Notes on MIMO Control Theory,* Lecture Notes, California Institute of Technology, 1990.

[14] S. Boyd and C. Barratt, *Linear Controller Design: Limits of Performance,* Prentice-Hall Inform. Systems Sci. Ser., Englewood Cliffs, NJ, 1991.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 33, No. 4, pp. 1102–1126, July 1995

© 1995 Society for Industrial and Applied Mathematics

009

ORDERS OF INPUT/OUTPUT DIFFERENTIAL EQUATIONS AND STATE-SPACE DIMENSIONS*

YUAN WANG† AND EDUARDO D. SONTAG‡

Abstract. This paper deals with the orders of input/output equations satisfied by nonlinear systems. Such equations represent differential (or difference, in the discrete-time case) relations between high-order derivatives (or shifts, respectively) of input and output signals. It is shown that, under analyticity assumptions, there cannot exist equations of order less than the minimal dimension of any observable realization; this generalizes the known situation in the classical linear case. The results depend on new facts, themselves of considerable interest in control theory, regarding universal inputs for observability in the discrete case, and observation spaces in both the discrete and continuous cases. Included in the paper is also a new and simple self-contained proof of Sussmann's universal input theorem for continuous-time analytic systems.

Key words. control systems, input/output equations, observation spaces, universal inputs, observability

AMS subject classifications. 93B15, 93A25, 93B25, 93B27, 93B29

1. Introduction.

Previous papers by the authors (see [40], [41]) studied various relationships between realizability of continuous-time systems and the existence of algebraic or analytic input/output differential equations. These are equations of the form

$$E\left(u(t), u'(t), u''(t), \ldots, u^{(r-1)}(t), y(t), y'(t), y''(t), \ldots, y^{(r)}(t)\right) = 0 \tag{1}$$

that relate inputs $u(\cdot)$ and outputs $y(\cdot)$. Such equations, and their discrete-time analogues, are of interest in identification theory and arise also naturally in the "behavioral" approach to systems (see, e.g., [43]). They provide a natural generalization of the autoregressive moving-average representations that appear in linear systems theory, where E is linear (in that case, the Laplace transform of the equation leads to the usual transfer function).

The papers [37], [40], [41] (see also [28] for analogous work in the discrete-time case) dealt with the relationships between the existence of such equations and the possibility of realizing the corresponding input/output (i/o) operator $u(\cdot) \mapsto y(\cdot)$ by a state-space system of the type

$$x'(t) = f(x(t)) + G(x(t))u(t)\,, \quad y(t) = h(x(t)) \tag{2}$$

whose state $x(t)$ evolves in an n-dimensional manifold. (Precise definitions are given later; for the rest of the introduction we give an informal discussion. The main assumption will be that all functions appearing are analytic.) While i/o equation descriptions of type (1) are well suited to identification algorithms, state-space descriptions of type (2) are often the basis of feedback design tools and are needed for

* Received by the editors April 2, 1993; accepted for publication (in revised form) January 18, 1994.

† Department of Mathematics, Florida Atlantic University, Boca Raton, Florida 33431 (ywang@polya.math.fau.edu). This research was supported in part by National Science Foundation grant DMS-9108250.

‡ Department of Mathematics, Rutgers University, New Brunswick, New Jersey 08903 (sontag@hilbert.rutgers.edu). This research was supported in part by US Air Force grant AFOSR-91-0346.

the statement and solution of control problems. Thus, it is of great interest to study the possible relationships between the two kinds of descriptions.

A question that has not been sufficiently studied and was not addressed in [40], [41] is that of comparing the *order* r of an i/o equation (1) to the minimal possible dimension n of a realization (2). In discrete time, for the analogous equations

$$E\left(u(t-1),\, u(t-2),\, \ldots,\, u(t-r),\, y(t),\, y(t-1),\, \ldots,\, y(t-r)\right) = 0 \tag{3}$$

and systems

$$\Sigma: \quad x(t+1) = f(x(t), u(t)), \quad y(t) = h(x(t)), \quad t = 0,\, 1,\, 2, \ldots, \tag{4}$$

it was known for a long time (see [28]) that one may have $r < n$, even if the system in (4) is minimal. It turns out, perhaps surprisingly, that this cannot happen in the continuous-time case: we prove here that if there is a minimal realization of dimension n, then no i/o equation can have order less than n. Moreover, we show that the result holds true also for discrete-time systems that are *reversible*, that is, those for which the controls induce one-to-one maps on the state space (the examples in [28] were not reversible).

The results in [40], [41] depend on an important equality among observation spaces. The latter are sets of functions on the state space that are obtained by performing different kinds of "experiments" with the system and extracting infinitesimal information from the observed data. The basic fact needed was established in [39], and it related the space obtained by using piecewise constant controls (and derivatives of the output function with respect to switching times between constant pieces) to the space obtained when using differentiable inputs instead (and the corresponding jet of derivatives of the output at time zero). The new results given in this paper depend on new facts, themselves of considerable interest in control theory, regarding subspaces obtained by the application of "generic" smooth inputs.

The results in this paper were announced and their proofs sketched in the conference paper [32] (and for discrete time in [42]). To be more precise, in [32] we derived our results from an equality between observation spaces that is somewhat weaker than the corresponding one proved here; namely, instead of the current Lemma 2.1, we only had that $\mathrm{d}\mathcal{F}(x) = \mathrm{d}\mathcal{F}_\mu(x)$ for generic jets μ and for *generic* states x. This is all that is needed in order to establish the desired results on orders of i/o equations. However, while this journal version was being written, Coron [5] showed that the equality can be strengthened so that it holds for *all* (not merely generic) states (but still generic μ). Since it turns out that the stronger equality can in fact be obtained with essentially the same proof as in [32], we now present the result directly in that form. (Since we are only interested in analytic systems, we can use elementary facts from analytic geometry to present a simpler approach to the problem than in [5]; in that reference the techniques of proof are very different, as the focus is on applications to feedback control problems for smooth systems. See also [31] for remarks on applications of results of the type proved here to path planning and feedback.)

In the development of the new observation space results, we needed to extend to discrete time the well-known and fundamental theorem by Sussmann on universal inputs for distinguishability of continuous-time analytic systems. It turned out that our proof also applies in continuous time. The theorem is obtained in a fairly direct way from a stronger result, Lemma 2.1 in this work. The proof of Lemma 2.1 is very elementary and intuitive, as it does not use anything more complicated than the fact that every descending chain of sets defined by analytic equations stabilizes relative to any fixed compact. (The original proof of Sussmann's theorem relies heavily on the stratification theory of subanalytic sets, a considerably deeper set of tools. Thus one contribution of this paper is to provide an alternative and simpler proof of that important result.) In addition to its role in helping to derive the universal input theorem and our main results, Lemma 2.1 also has its own independent interest, as it provides relationships between observation spaces defined in different ways and, thus, provides connections between several different notions of observability. We also note the very recent work [33], where further results on universal inputs are presented; these results show in particular the existence of inputs that are universal uniformly over the class of all analytic systems.

Another set of results that arose naturally while studying the problems in this paper, and which are included here, deals with the relationships among various alternative notions of observability, especially those proposed in the context of the differential-algebraic approach to control theory. We are able to characterize, for instance, the notion of observability proposed in [10], [9] in terms of more standard local observability concepts.

1.1. Other related work.

In addition to the references already mentioned, work by many authors is related to the topic of i/o equations and realizability; see for instance [6], [13], [37]. In particular, [7], [8] showed that one must add inequality constraints to (1) in order to obtain a precise characterization of the behavior of a state-space system, unless stronger algebraic conditions hold. In [26], [38], [4], local i/o equations were derived under nondegeneracy rank conditions, for smooth systems, under observability assumptions. The notions of observation space and algebra that we employ were introduced for discrete-time systems in [28], and their analogous continuous-time versions were given in [2], [3].

1.2. Outline of paper.

In §2 we introduce continuous-time systems and a technical result on observation spaces for generic jets. Certain special cases for which stronger conclusions can be given, namely bilinear and rational systems, are also studied there. In §3, we define universal inputs and relate their properties to the results on equality of observation spaces and to the orders of i/o equations. The following section has a proof of the main technical results stated in §§2 and 3. After this, §5 provides the discrete-time results. There are also two appendices with some technical lemmas that are required by the proofs.

2. Observation spaces for continuous-time systems.

In this section we first discuss several natural ways of defining observation spaces for continuous-time systems and then explore the relationships between the different definitions.

2.1. Observation spaces.

Consider an *analytic system*

$$\Sigma : \begin{cases} x'(t) &= g_0(x(t)) + \sum_{i=1}^{m} g_i(x(t))u_i(t)\,, \\ y(t) &= h(x(t))\,, \end{cases} \tag{5}$$

where for each t, $x(t) \in \mathcal{M}$, which is an analytic (second countable) manifold of dimension n, $h : \mathcal{M} \longrightarrow \mathbb{R}$ is an analytic function and $g_0, g_1, \ldots, g_m$ are analytic vector fields defined on $\mathcal{M}$. Controls are measurable essentially bounded maps $u : [0, T] \longrightarrow \mathbb{R}^m$ defined on suitable intervals. In general, $\varphi(t, x, u)$ denotes the state trajectory of (5) corresponding to a control u and initial state x, defined at least for small t.

In the special case in which $\mathcal{M} = \mathbb{R}^n$ and the entries of the vector fields g_i's (on the natural global coordinates) and of the function h are rational (with no real poles), we call (5) a *rational system.* If, in addition, the entries of the g_i's and h are polynomials, we call (5) a *polynomial system.*

For a given continuous-time system, let $\mathcal{F}$ be the subspace of functions $\mathcal{M} \longrightarrow \mathbb{R}$ spanned by the Lie derivatives of h in the directions of $g_0, g_1, \dots, g_m$, i.e.,

$$\mathcal{F} := \operatorname{span}_{\mathbb{R}} \left\{ L_{g_{i_1}} L_{g_{i_2}} \dots L_{g_{i_r}} h : \ r \geq 0, \ 0 \leq i_j \leq m \right\} . \tag{6}$$

This is the *observation space* associated with (5) (see, e.g., [30, Rem. 5.4.2]) For each $x \in \mathcal{M}$, let $\mathcal{F}(x)$ denote the space obtained by evaluating the elements of $\mathcal{F}$ at x.

For each $\alpha \in \mathcal{F}$, we may consider its differential $\mathrm{d}\alpha$, seen as a 1-form. For each $x \in \mathcal{M}$, we let $\mathrm{d}\mathcal{F}(x)$ be the space of covectors defined by

$$\mathrm{d}\mathcal{F}(x) = \{\mathrm{d}\alpha(x) : \ \alpha \in \mathcal{F}\} \ .$$

We also let $\mathrm{d}\mathcal{F}$ be $\{d\alpha : \ \alpha \in \mathcal{F}\}$ as a space of 1-forms.

A related construction is as follows. First, let $\mathbb{R}^{m,\infty} = \prod_{i=1}^{\infty} \mathbb{R}^m$ endowed with the box topology, for which a base of open sets consists of all sets of the form $\prod_{i=1}^{\infty} \mathcal{U}_i$, where each $\mathcal{U}_i$ is an open set of $\mathbb{R}^m$. A *generic* subset $\mathcal{W}$ of $\mathbb{R}^{m,\infty}$ is one that contains a countable intersection of open dense sets. It can easily be shown that with the above topology, $\mathbb{R}^{m,\infty}$ is a Baire space; thus, a generic subset is always dense.

Now for any $\mu = (\mu_0, \mu_1, \dots)$ in $\mathbb{R}^{m,\infty}$, we define

$$\psi_i(x, \mu) = \left. \frac{d^i}{dt^i} \right|_{t=0} h(\varphi(t, x, u)) \tag{7}$$

for $i \geq 0$, where u is any C^∞ control with initial values $u^{(j)}(0) = \mu_j$. The functions $\psi_i(x, \mu)$ can be expressed—applying repeatedly the chain rule—as polynomials in the $\mu_j = (\mu_{1j}, \dots, \mu_{mj})$ whose coefficients are analytic functions (rational functions if the system is rational) of x. Take the single-input case

$$\dot{x} = f(x) + g(x), \quad y = h(x)$$

(for simplicity of notation) as an example. The functions are

$$\begin{aligned}
&\psi_0(x, \mu) = h(x), \\
&\psi_1(x, \mu) = L_f h(x) + \mu_0 L_g h(x), \\
&\psi_2(x, \mu) = L_f^2 h(x) + \mu_0 (L_g L_f h(x) + L_f L_g h(x)) + \mu_0^2 L_g^2 h(x) + \mu_1 L_g h(x),
\end{aligned}$$

and so forth. For instance, for single-input single-output linear systems

$$x' = Ax + bu, \quad y = cx,$$

we have,

$$\psi_l(x, \mu) = cA^l x + \sum_{i=1}^{l} \mu_{i-1} cA^{i-1} b, \quad l = 0, 1, \dots \ .$$

For each fixed $\mu \in \mathbb{R}^{m,\infty}$, let $\mathcal{F}_\mu$ be the subspace of functions from $\mathcal{M}$ to $\mathbb{R}$ defined by

$$\mathcal{F}_\mu = \operatorname{span}_{\mathbb{R}} \{\psi_0(\cdot, \mu), \ \psi_1(\cdot, \mu), \ \psi_2(\cdot, \mu), \dots\} \tag{8}$$

and let $\mathcal{F}_\mu(x)$ be the space obtained by evaluating the elements of $\mathcal{F}_\mu$ at x for each $x \in \mathcal{M}$. Let $\mathrm{d}\mathcal{F}_\mu(x)$ be the space of covectors given by

$$\mathrm{d}\mathcal{F}_\mu(x) = \{\mathrm{d}\psi(x,\mu) : \psi \in \mathcal{F}_\mu\}$$

for each $x \in \mathcal{M}$. For instance, for linear systems, $\mathrm{d}\psi_l(x,\mu) = cA^l$ and

$$\mathrm{d}\mathcal{F}_\mu(x) = \operatorname{span}\{c,\ cA,\ cA^2,\ \ldots\},$$

which is independent of μ (and x). We also let $\mathrm{d}\mathcal{F}_\mu$ be $\{d\psi(\cdot,\mu) : \psi \in \mathcal{F}_\mu\}$, seen as a space of covector fields.

Clearly, for each μ, $\mathcal{F}_\mu$ is a subspace of $\mathcal{F}$, and therefore, for each x also $\mathrm{d}\mathcal{F}_\mu(x)$ is a subspace of $\mathrm{d}\mathcal{F}(x)$. The main result in [39] says that

$$\boxed{\mathcal{F} = \sum_\mu \mathcal{F}_\mu\,.} \tag{9}$$

This equality is fundamental in establishing results linking realizability to the existence of i/o equations, in [40] and [41]. In intuitive but less rigorous terms, the equality in (9) can be interpreted as follows. We consider the successive derivatives $y(0)$, $y'(0)$, $y''(0)$, ... expressed as functions of $x(0)$ and $u(0)$, $u'(0)$, $u''(0)$, ... For particular controls $u(t)$, the $y(0)$, $y'(0)$, $y''(0)$, ... are just functions of x; taking the span of all such functions, over all possible *smooth* controls, one obtains the right-hand side of (9). On the other hand, taking all possible *piecewise-constant* instead of smooth controls and taking derivatives with respect to the times at which the controls switch values, one obtains the space $\mathcal{F}$ in the left-hand side of (9).

The following is a technical result for continuous-time systems, which will help in deriving the desired facts about i/o equations.

LEMMA 2.1. *Assume that* (5) *is an analytic system. Then there exists a generic subset $\mathcal{W}$ of $\mathbb{R}^{m,\infty}$ such that*

$$\mathcal{F}(x) = \mathcal{F}_\mu(x) \tag{10}$$

and

$$\mathrm{d}\mathcal{F}(x) = \mathrm{d}\mathcal{F}_\mu(x) \tag{11}$$

for every $x \in \mathcal{M}$ and all $\mu \in \mathcal{W}$.

Remark 2.2. The above conclusions are also true if instead of the box topology one uses the weak topology on $\mathbb{R}^{m,\infty}$. This is the topology for which a basis of open sets consists of all sets of the form $\prod_{i=1}^\infty \mathcal{U}_i$, where each $\mathcal{U}_i$ is an open subset of $\mathbb{R}^m$ and only finitely many of them are proper subsets of $\mathbb{R}^m$. Clearly, the weak topology is coarser than the topology used before. With this topology, $\mathbb{R}^{m,\infty}$ is again a Baire space. We will remark at the end of the proof of Lemma 2.1 in §4 that the conclusions of Lemma 2.1 also hold for the weak topology. Moreover, these conclusions can be established as consequences of a more general result about convergent generating series, that ensures there exists a generic subset $\mathcal{W}$ of $\mathbb{R}^{m,\infty}$ with the property that these jets suffice for distinguishing *all possible* convergent generating series; more details are given in [33].

Remark 2.3. The conclusions in Lemma 2.1 do not always hold for every $\mu \in \mathbb{R}^{m,\infty}$. Consider as an illustration the following bilinear system:

$$x_1' = x_2, \quad x_2' = x_2 + x_1 u, \quad y = x_2\,. \tag{12}$$

For this system, $\mathcal{F} = \operatorname{span}\{x_1, x_2\}$, thus, $\mathcal{F}(x) \neq 0$ for all $x \neq 0$. But on the other hand, we have

$$\psi_0(x, \mu) = x_2, \ . \ \psi_1(x, \mu) = x_2 + x_1\mu_0 \, ,$$

and in general, $\psi_k(x, \mu) = P_k(x, \mu_0, \mu_1, \ldots, \mu_{k-2}) + x_1\mu_{k-1}$, where P_k is some polynomial. Clearly, for every $x = (x_1, x_2)$ for which $x_1 \neq 0$, one can find a solution μ recursively for the equations $\psi_i(x, \mu) = 0$ for $i > 0$. Hence, as long as $x_1 \neq 0$ and $x_2 = 0$, there exists some jet μ such that $\mathcal{F}_\mu(x) = 0$, which is therefore different from $\mathcal{F}(x)$ when $x_1 \neq 0$ and $x_2 = 0$.

2.2. Algebraic formulation. In this section, we assume for simplicity that $\mathcal{M} = \mathbb{R}^n$; we could work with more general manifolds but this would complicate notation, and in any case we will only need to apply the results given here locally. We say a function β is a *meromorphic function* if $\beta = \frac{p}{q}$, where p and q are analytic functions defined on $\mathcal{M}$, and $q \not\equiv 0$. (Note that this global definition is different from the local definition usually given; see, e.g., [17]. It will be enough for our purposes.) For each function $\alpha \in \mathcal{F}$, $\mathrm{d}\alpha$ is a covector field defined on $\mathcal{M}$. If β is a meromorphic function defined on $\mathcal{M}$, then $\beta \mathrm{d}\alpha$ is a well-defined 1-form on some open dense subset of $\mathcal{M}$ and any finite sum of such partially defined covector fields is defined on a common open dense set. Thus, we may introduce the subspace $\widehat{\mathrm{d}\mathcal{F}}$ of the cotangent space defined by

$$\widehat{\mathrm{d}\mathcal{F}} := \operatorname{span}_{\mathbb{R}_x}\{\mathrm{d}\alpha : \ \alpha \in \mathcal{F}\} \, ,$$

where $\mathbb{R}_x$ is the field of meromorphic functions defined on $\mathcal{M}$. Similarly, one can define, for each $\mu \in \mathbb{R}^{m,\infty}$, the space $\widehat{\mathrm{d}\mathcal{F}}_\mu$ by

$$\widehat{\mathrm{d}\mathcal{F}}_\mu := \operatorname{span}_{\mathbb{R}_x}\{\mathrm{d}\alpha : \ \alpha \in \mathcal{F}_\mu\} \, .$$

Note that there are natural identifications $\widehat{\mathrm{d}\mathcal{F}} \simeq \mathrm{d}\mathcal{F} \otimes \mathbb{R}_x$ and $\widehat{\mathrm{d}\mathcal{F}}_\mu \simeq \mathrm{d}\mathcal{F}_\mu \otimes \mathbb{R}_x$.

Since $\mathcal{M} = \mathbb{R}^n$, we can identify elements of $\widehat{\mathrm{d}\mathcal{F}}$ with vectors

$$(\alpha_1(x), \alpha_2(x), \ldots, \alpha_n(x))$$

of meromorphic functions defined on $\mathcal{M}$. The dimension of $\widehat{\mathrm{d}\mathcal{F}}$ over $\mathbb{R}_x$ is the size of the largest matrix that can be formed out of such vectors and has full rank, i.e., has a minor that is not zero as a function. That is, $\dim_{\mathbb{R}_x} \widehat{\mathrm{d}\mathcal{F}}$ is the same as $\max_{x \in \mathcal{M}} \dim \mathrm{d}\mathcal{F}(x)$. A similar argument can be made for each $\mathrm{d}\mathcal{F}_\mu(x)$; together with Lemma 2.1, we can then conclude the following corollary.

COROLLARY 2.4. *For any analytic system, $\widehat{\mathrm{d}\mathcal{F}}_\mu = \widehat{\mathrm{d}\mathcal{F}}$ for all μ in a generic set of $\mathbb{R}^{m,\infty}$.*

Yet another object is obtained if one instead views the elements

$$\psi_i(x, U) \tag{13}$$

as rational functions (in particular polynomials), on the formal variables $U = \{U_{ij}\}$, whose coefficients are functions of x, as opposed to seeing them as functions of x for each numerical choice $U_{ij} = \mu_{ij}$. We proceed as follows. Let

$$K = \mathbb{R}\Big(\{U_{ij} : \ 1 \leq i \leq m, \, j \geq 0\} \Big)$$

be the field obtained by adjoining the indeterminates U_{ij} to $\mathbb{R}$, and let

$$K_x = \mathbb{R}_x\Big(\{U_{ij} :\ 1 \le i \le m,\, j \ge 0\}\Big)$$

be the field obtained by adjoining the indeterminates U_{ij} to $\mathbb{R}_x$. We then let $\mathfrak{F}$ be defined as the subspace of K_x spanned by the functions ψ_i over the field K, i.e.,

$$\mathfrak{F} := \operatorname{span}_K \{\psi_i :\ i \ge 0\}\ .$$

Thus, $\mathfrak{F}$ consists of finite linear combinations $\sum q_i(U)\psi_i(x,\, U)$, where the $q_i(\cdot)$ are rational functions on the variables $\{U_{ij}\}$. Such a linear combination can be seen as a rational function on the $\{U_{ij}\}$ whose coefficients are meromorphic functions of x (and hence also meromorphic functions) and, thus, elements of K_x. The differentials (with respect to x) of elements of K_x are viewed as rational functions in $\{U_{ij}\}$, whose coefficients are (in general, partially defined) covector fields. Finally we define

$$\widehat{\mathrm{d}\mathfrak{F}} := \operatorname{span}_{K_x}\{\mathrm{d}\psi :\ \psi \in \mathfrak{F}\}.$$

Then Lemma 2.1 implies the following corollary.

COROLLARY 2.5. *For any analytic system,* $\dim_{\mathbb{R}_x} \widehat{\mathrm{d}\mathcal{F}} = \dim_{K_x} \widehat{\mathrm{d}\mathfrak{F}}$.

Proof. Clearly $\dim_{K_x} \widehat{\mathrm{d}\mathfrak{F}} \le \dim_{\mathbb{R}_x} \widehat{\mathrm{d}\mathcal{F}}$. Conversely, $\dim_{K_x} \widehat{\mathrm{d}\mathfrak{F}} = \max_\mu \dim_{R_x} \widehat{\mathrm{d}\mathcal{F}}_\mu$. The desired conclusion then follows from Corollary 2.4. □

2.3. Bilinear and rational systems. Now consider the bilinear system

$$\begin{aligned} x' &= A_0 x + \sum_{i=1}^m u_i A_i x, \\ y &= cx, \end{aligned}$$

where $A_0, A_1, \ldots, A_m$ are $n \times n$ matrices and c is an $1 \times n$ matrix. For each multi-index $i_1 i_2 \ldots i_r$, where $0 \le i_j \le m$ for each $j \ge 0$,

$$L_{g_{i_1}} L_{g_{i_2}} \ldots L_{g_{i_r}} h(x) = c A_{i_r} A_{i_{r-1}} \ldots A_{i_1} x.$$

Note that ψ_i (as defined in (7)) is also linear in x for each i; for instance, in the single-input case (for simplicity of notation),

$$\psi_2(x,\, \mu_0,\, \mu_1) = c(A_0 + \mu_0 A_1)^2 x + \mu_1 c A_1 x\,.$$

Thus, for the bilinear case, we have the following corollary.

COROLLARY 2.6. *For a bilinear system,*

$$\mathcal{F} = \mathcal{F}_\mu \quad \text{and} \quad \mathrm{d}\mathcal{F} = \mathrm{d}\mathcal{F}_\mu \tag{14}$$

for every μ in a generic subset of $\mathbb{R}^{m,\infty}$.

Remark 2.7. We would like to point out that this corollary does *not* hold in general. The following simple example shows that for a general nonlinear system, $\mathcal{F}$ and $\mathcal{F}_\mu$ (respectively, $\mathrm{d}\mathcal{F}$ and $\mathrm{d}\mathcal{F}_\mu$) may not be the same for *any* μ, even though the two spaces $\widehat{\mathrm{d}\mathcal{F}}$ and $\widehat{\mathrm{d}\mathcal{F}}_\mu$ are the same.

Example 2.8. Consider the system

$$x' = x^3 + x^2 u, \quad y = x.$$

It is easy to see that

$$y = x, \quad y' = x^3 + x^2u,$$
$$y'' = 3x^5 + 5x^4u + 2x^3u^2 + x^2u',$$

and in general, $y^{(k)} = (2k-1)!!x^{2k+1} + p_k(u, u', \ldots, u^{(k-2)}, x) + x^2u^{(k-1)}$, where p_k is a polynomial in x of degree less than or equal to $2k$. It can be seen that

$$\mathcal{F} = \operatorname{span}_{\mathrm{IR}} \{x,\ x^2,\ x^3, \ldots\} \ .$$

However, $x^2 \notin \mathcal{F}_\mu$ for any μ for the following reason. Assume that

$$x^2 = \sum_{i=0}^{k} a_i\psi_i(x, \mu)$$

for some k and some $a_0, a_1, \ldots, a_k \in \mathrm{IR}$. Then $a_i = 0$ for $i \geq 2$, otherwise the degree of x in the left-hand side would be higher than 3. Thus the above equation becomes

$$x^2 = a_0x + a_1(x^3 + x^2\mu_0),$$

which is impossible. This shows that $\mathcal{F}_\mu \neq \mathcal{F}$ for any μ even though, in this case, $\widehat{\mathrm{d}}\mathcal{F} = \widehat{\mathrm{d}}\mathcal{F}_\mu = \operatorname{span}_{\mathrm{IR}_x}\{dx\}$ for all μ.

In this example, it is also true that $\mathrm{d}\mathcal{F} \neq \mathrm{d}\mathcal{F}_\mu$ for any μ. This can be shown as follows. If $\mathrm{d}\mathcal{F} = \mathrm{d}\mathcal{F}_\mu$, then $\mathrm{d}x^2 = 2x\mathrm{d}x \in \mathrm{d}\mathcal{F}_u$. From here it would follow that $x^2 = \alpha_1(x, \mu) + \alpha_2(x, \mu) + \cdots + \alpha_l(x, \mu) + c$ for some elements $\alpha_i \in \mathcal{F}_\mu$ and some constant $c \in \mathrm{IR}$. But it can be seen from the above argument that this is impossible.

Assume now that (5) is a rational system. Define $\mathcal{A}$ ($\mathcal{A}_\mu$, respectively) as the IR-algebra generated by the elements of $\mathcal{F}$ ($\mathcal{F}_\mu$, repectively). Then we define the observation field $\mathcal{Q}$ ($\mathcal{Q}_\mu$, respectively) as the quotient field of $\mathcal{A}$ ($\mathcal{A}_\mu$, respectively). For a field extension Q of IR, we use $\operatorname{trdeg}_{\mathrm{IR}}Q$ to denote the transcendence degree of Q over IR. Then we have the following conclusion for rational systems, in analogy to the above conclusion about bilinear systems.

COROLLARY 2.9. *For a rational system,*

$$\operatorname{trdeg}_{\mathrm{IR}}\mathcal{Q} = \operatorname{trdeg}_{\mathrm{IR}}\mathcal{Q}_\mu\ ,$$

for each μ in a generic subset of $\mathrm{IR}^{m,\infty}$.

3. Observability and universal inputs in continuous time. Consider an analytic system (5). Fix any two states p, $q \in \mathcal{M}$ and take an input u. We say p and q are *distinguished by u*, denoted by $p \not\sim_u q$, if $h(\varphi(\cdot, p, u)) \neq h(\varphi(\cdot, q, u))$ (considered as functions defined on the common domain of $\varphi(\cdot, p, u)$ and $\varphi(\cdot, q, u)$); otherwise we say p and q cannot be distinguished by u, denoted by $p \sim_u q$. If p and q cannot be distinguished by *any* input u, then we say p and q are *indistinguishable*, denoted by $p \sim q$. If for any two states, $p \sim q$ implies $p = q$, then we say that system (5) is *observable*. (See [30, Chap. 5].)

An input u is called a *universal* (*distinguishing*) input for system (5) if every distinguishable pair can be distinguished by u. The existence of universal inputs was first studied in [15] for bilinear systems, in [27] for analytic systems with compact state spaces, and for arbitrary analytic systems in [35] for the continuous case. In this work, we will provide a different and simpler proof of the general result in [35]. (Also, we later give a discrete-time version.) We now state the result to be proved.

For each $T > 0$, we consider $\mathcal{C}^\infty[0, T]$ endowed with the Whitney topology, that is, the topology for which a neighborhood base for each function $u(\cdot) \in \mathcal{C}^\infty[0, T]$ consists of the sets of the following form:

$$\mathcal{U}_{u,k,\delta} = \left\{ v \in \mathcal{C}^\infty[0, T] : \max_{0 \le i \le k,\ t \in [0, T]} |v^{(i)}(t) - u^{(i)}(t)| \le \delta \right\}$$

for some $k \ge 0$ and some $\delta > 0$. This is well known to be a Baire space (see [14]). By a generic subset of $\mathcal{C}^\infty[0, T]$ we mean a subset of $\mathcal{C}^\infty[0, T]$ containing a countable intersection of open dense sets.

THEOREM 3.1 (Sussmann's universal input theorem). *For any analytic system* (5), *and any fixed* $T > 0$, *the set of universal inputs is a generic subset of* $\mathcal{C}^\infty[0, T]$.

PROPOSITION 3.2. *There is always an analytic universal input for any analytic system.*

We will provide proofs of Theorem 3.1 and Proposition 3.2 in §4.1.

Consider the following more general class of systems:

$$x'(t) = f(x(t), u(t)), \quad y(t) = h(x(t)), \tag{15}$$

where for each t, $x(t) \in \mathcal{M}$, which is an analytic manifold of dimension n, $h : \mathcal{M} \to \mathbb{R}$ is an analytic function and $f : \mathcal{M} \times \mathbb{R}^m \to T\mathcal{M}$ is analytic and $f(x, u) \in T_x\mathcal{M}$ for each (x, u), so in particular, $f(\cdot, u)$ is an analytic vector field for each $u \in \mathbb{R}^m$. Controls are measurable essentially bounded maps: $u : [0, T] \longrightarrow \mathbb{R}^m$, for some $T = T_u > 0$. We apply the same definitions of distinguishability, observability, and universal inputs as for system (5) to system (15). One can then generalize the conclusion of Theorem 3.1 to systems of type (15) by means of the following argument. We consider the following system:

$$\begin{aligned} x'(t) &= f(x(t), z(t)), \quad z'(t) = v(t), \\ y(t) &= h(x(t)), \end{aligned} \tag{16}$$

where v is now a new control. By Proposition 5.1.11 in [30], one knows that if (x_1, x_2) is a distinguishable pair for (15), then x_1, x_2 can be distinguished by a differentiable (in fact, an analytic) control u. It then follows that for (16), the pair (ξ, ζ), where $\xi = (x_1, u(0))$ and $\zeta = (x_2, u(0))$, is distinguished by $v(t) = u'(t)$. On the other hand, if for (16) the pair $((x_1, z), (x_2, z))$ is distinguished by v, then for (15) (x_1, x_2) is distinguished by the control

$$z + \int_0^t v(s)\, ds.$$

Therefore, (x_1, x_2) is a distinguishable pair of (15) if and only if there exists some $z \in \mathbb{R}^m$ such that $((x_1, z), (x_2, z))$ is a distinguishable pair for (16) for some z. Applying Theorem 3.1 to system (16), we proved the following conclusion.

COROLLARY 3.3. *The universal inputs for system* (15) *form a generic subset of* $\mathcal{C}^\infty[0, T]$, *for any* $T > 0$.

3.1. Other notions of observability.

In what follows, we study relationships among several alternative notions of "observability" that have been proposed by various authors.

Take an open subset $\mathcal{U}$ of $\mathcal{M}$ and any two points p, $q \in \mathcal{U}$. If for every input u, $h(\varphi(t, p, u)) = h(\varphi(t, q, u))$ for each t for which $\varphi(T, p, u)$ and $\varphi(T, q, u)$ are both defined and in $\mathcal{U}$ for all $0 \leq t \leq T$, then we say that p and q are *$\mathcal{U}$-indistinguishable* (see, e.g., [29]).

Fix a point $p \in \mathcal{M}$. If for every neighborhood $\mathcal{U}_p$ there is a neighborhood $V_p \subset \mathcal{U}_p$ such that for any $q \in V_p$ the condition that q and p are $\mathcal{U}_p$-indistinguishable implies $p = q$, then we say the system (5) is *locally observable at* p. If (5) is locally observable at every point p, then we say (5) is *locally observable*. If there is an open dense set $\mathcal{U} \subset \mathcal{M}$ such that (5) is locally observable at every point p of $\mathcal{U}$, then we say (5) is *generically locally observable*. See [29] for details on local observability and related concepts such as the slightly different definition in [26]. The following fact is an immediate consequence of Lemma 2.10 and facts (2.4) and (2.8) in [29].

PROPOSITION 3.4. *An analytic system* (5) *is generically locally observable if and only if* $\max_x \dim \mathrm{d}\mathcal{F}(x) = n$.

PROPOSITION 3.5. *Let* $\mathcal{M} = \mathbb{R}^n$ *and let* (5) *be an analytic system. Then the following are equivalent*:

(1) *The system is generically locally observable.*

(2) $\dim_{K_x} \widehat{\mathrm{d}\mathfrak{F}} = n$.

(3) $\dim_{\mathbb{R}_x} \widehat{\mathrm{d}\mathcal{F}} = n$.

Proof. The maximum dimension of $\mathrm{d}\mathcal{F}(x)$ is the same as the $\dim_{\mathbb{R}_x} \widehat{\mathrm{d}\mathcal{F}}$. This shows that (1) and (3) are equivalent; (2) is equivalent to (3) by Corollary 2.5. □

For a polynomial system, the $\psi_i(x, U)$'s (as defined in (13)) are polynomial functions of both x and U. We say that a polynomial system is *weakly algebraically observable* if each coordinate x_i is algebraically over the field $K(\{\psi_i : i \geq 0\})$ $(= \mathbb{R}(\{U_{ij}, \psi_k, i = 1, \ldots, m;\ j, k \geq 0\}))$. It follows that Σ is weakly algebraically observable if and only if $\dim_{K(x)} \widehat{\mathrm{d}\mathfrak{F}} = n$, where $K(x)$ is the field of rational functions over K. (This is proved as follows: The dimension condition is equivalent, by [18, Thm. III of III.7], to the property that the transcendence degree of $K_0 = K(\{\psi_i : i \geq 0\})$ over K should be equal to n. On the other hand, we have the inclusions $K \subseteq K_0 \subseteq K(x)$, so $\mathrm{trdeg}_K K_0 + \mathrm{trdeg}_{K_0} K(x) = n$. Thus the dimension is n if and only if $\mathrm{trdeg}_{K_0} K(x) = 0$, i.e., if and only if $K(x)$ is algebraic over K_0.) By Proposition 3.5, we have the following corollary.

COROLLARY 3.6. *A polynomial system is weakly algebraically observable if and only if the system is generically locally observable.*

The notion of weakly algebraic observability used here was called "weak observability" in [28]. The same notion was used in [10] and extended to cover implicit systems as well.

3.2. Orders of i/o equations in continuous-time case.

3.2.1. State-space systems. We say that a state-space system Σ *admits an i/o equation* such as

$$A(u(t),\, u'(t),\, \ldots,\, u^{(k-1)}(t),\, y(t),\, y'(t),\, \ldots,\, y^{(k)}(t)) = 0\,, \tag{17}$$

where A is a nonzero analytic function from $\mathbb{R}^{mk} \times \mathbb{R}^{k+1}$ to $\mathbb{R}$, if (17) holds for every initial state x, every $\mathcal{C}^k$ i/o pair (u, y) of (5), and all t such that $y(t)$ is defined. The *order* of an equation (17) is defined to be the highest $r \leq k$ such that

$$\frac{\partial}{\partial \nu_r} A(\mu_0,\, \ldots,\, \mu_{k-1},\, \nu_0,\, \nu_1,\, \ldots,\, \nu_k)$$

is not a zero function.

For a given system Σ, we define $\delta(\Sigma)$ to be the lowest possible order of an i/o equation that Σ admits. In the case that there is no such i/o equation, $\delta(\Sigma)$ is defined to be $+\infty$.

THEOREM 3.7. *Assume Σ is an n-dimensional analytic system defined by* (5). *If Σ is generically locally observable, then $\delta(\Sigma) \geq n$. If, in addition, Σ is a rational system, then $\delta(\Sigma) = n$.*

Proof. Let $\mathcal{U} \subseteq \mathcal{M}$ be an open subset diffeomorphic to $\mathbb{R}^n$. We consider the restriction of Σ to $\mathcal{U}$. This system is still generically locally observable, and an equation for Σ is also an equation for the restriction. So without loss of generality, we assume from now on that $\mathcal{M} = \mathbb{R}^n$.

Assume that $\delta(\Sigma) = k < \infty$ and Σ admits i/o equation (17) of order k. For each integer $i \geq 0$, let

$$A_i = \frac{\partial^i}{\partial \nu_k^i} A(\mu_0, \ldots, \mu_{k-1}, \nu_0, \nu_1, \ldots, \nu_k).$$

Claim. There exists an i, such that A_i is not an i/o equation of Σ.

We prove the claim as follows. Assume that A_i is an i/o equation of Σ for every i. Then for any fixed i/o pair (u, y) and any fixed t, it holds that

$$A_i(u(t), \ldots, u^{(k-1)}(t), y(t), \ldots, y^{(k)}(t)) = 0$$

for all i. Thus, as a function of ν_k for these fixed values $u(t), \ldots, y^{(k-1)}(t)$, all derivatives of

$$A(u(t), \ldots, u^{(k-1)}(t), y(t), \ldots, y^{(k-1)}(t), \nu_k) \tag{18}$$

evaluated at $\nu_k = y^{(k)}(t)$ vanish. It then follows from the analyticity of A that (18) vanishes for all values of ν_k. Let $\bar{\nu}_k$ be such that the function

$$\tilde{A}(\mu_0, \ldots, \mu_{k-1}, \nu_0, \ldots, \nu_{k-1}) := A(\mu_0, \ldots, \mu_{k-1}, \nu_0, \ldots, \nu_{k-1}, \bar{\nu}_k)$$

is not a zero function. Clearly it holds that

$$\tilde{A}(u(t), \ldots, u^{(k-1)}(t), y(t), \ldots, y^{(k-1)}(t)) = 0 \tag{19}$$

for all i/o pairs of Σ. If one can show that $\tilde{A}$ does not depend on μ_{k-1}, then one concludes that $\tilde{A} = 0$ is an i/o equation for Σ. For this, we proceed as follows. First of all, (19) holds for all i/o pairs of (u, y) if and only if the following holds:

$$\tilde{A}(\mu_0, \ldots, \mu_{k-1}, \psi_0(x, \mu), \ldots, \psi_{k-1}(x, \mu)) = 0$$

for all $x \in \mathcal{M}$ and all μ. Note here that ψ_i defined by (7) does not depend on μ_j for $j \geq i$. It follows that for any $\bar{\mu}_{k-1}$,

$$\tilde{A}(\mu_0, \ldots, \mu_{k-2}, \bar{\mu}_{k-1}, \psi_0(x, \mu), \ldots, \psi_{k-1}(x, \mu)) = 0$$

for all x and all μ. Finally, pick $\bar{\mu}_{k-1}$ such that

$$\bar{A}(\mu_0, \ldots, \mu_{k-2}, \nu_0, \ldots, \nu_{k-1}) := \tilde{A}(\mu_0, \ldots, \mu_{k-2}, \bar{\mu}_{k-1}, \nu_0, \ldots, \nu_{k-1}, \bar{\nu}_k)$$

is not a zero function. Then $\bar{A} = 0$ is an i/o equation of order $k - 1$ for Σ. This contradicts the assumption that $\delta(\Sigma) = k$. The claim is thus proved.

Now let $r \geq 1$ be the smallest number for which $A_r = 0$ is not an i/o equation for Σ. Replace A in (17) by A_{r-1}. Evaluating (17) at $t = 0$, the equation implies the identity

$$A(\mu_0, \ldots, \mu_{k-1}, \psi_0(x, \mu), \ldots, \psi_k(x, \mu)) = 0. \tag{20}$$

Since $A_1 = 0$ is not an i/o equation of Σ, it follows that there exists some $\mu \in \mathbb{R}^{mk}$ such that

$$A_1(\mu_0, \ldots, \mu_{k-1}, \psi_0(x, \mu), \ldots, \psi_k(x, \mu)) \neq 0, \tag{21}$$

as a function of x, and hence, by analyticity, the complement of

$$B = \{\mu \in \mathbb{R}^{mk} : \ A_1(\mu_0, \ldots, \mu_{k-1}, \psi_0(x, \mu), \ldots, \psi_k(x, \mu)) = 0, \ \forall x\}$$

is an open dense subset of $\mathbb{R}^{mk}$.

Combining (20) and (21), one sees, for each $\mu \notin B$, that $\mathrm{d}\psi_k(\cdot, \mu)$ is a linear combination of $\mathrm{d}\psi_0(\cdot, \mu), \ldots, \mathrm{d}\psi_{k-1}(\cdot, \mu)$ over $\mathbb{R}_x$. Thus $\widehat{\mathrm{d}\mathcal{F}}_\mu^k = \widehat{\mathrm{d}\mathcal{F}}_\mu^{k-1}$, where, for each i, $\widehat{\mathrm{d}\mathcal{F}}_\mu^i$ is the subspace of $\widehat{\mathrm{d}\mathcal{F}}_\mu$ spanned by $\mathrm{d}\psi_0(\cdot, \mu), \mathrm{d}\psi_1(\cdot, \mu), \ldots, \mathrm{d}\psi_i(\cdot, \mu)$. Differentiating (17) with respect to time, one sees that for any $i > 1$ it holds that

$$\begin{aligned} &A_1(u(t), \ldots, u^{(k-1)}(t), y(t), \ldots, y^{(k)}(t))y^{(k+i)}(t) \\ &\quad = A_i(u(t), \ldots, u^{(k+i-1)}(t), \ldots, y(t), \ldots, y^{(k+i-1)}(t)) \end{aligned}$$

for every i/o pair (u, y) of Σ, where A_i is some analytic function. Thus, by induction, one can show that $\widehat{\mathrm{d}\mathcal{F}}_\mu^{k+i} = \widehat{\mathrm{d}\mathcal{F}}_\mu^{k-1}$ for all $\mu \notin B$. It then follows that $\dim_{R_x} \widehat{\mathrm{d}\mathcal{F}}_\mu \leq k$ for all $\mu \notin \widehat{B}$, where $\widehat{B} \subset \mathbb{R}^{m,\infty}$ is defined by $\widehat{B} = B \times \mathbb{R}^m \times \mathbb{R}^m \times \cdots$.

On the other hand, by Corollary 2.4 and Proposition 3.5, one knows that $\dim_{\mathbb{R}_x} \widehat{\mathrm{d}\mathcal{F}}_\mu = \dim_{\mathbb{R}_x} \widehat{\mathrm{d}\mathcal{F}} = n$ for all μ in a dense (in fact, even in a generic) subset of $\mathbb{R}^{m,\infty}$. Therefore, Σ cannot admit any i/o equation of order lower than n.

If Σ is a rational system, then an easy elimination argument (based on the fact that any set of $n+1$ rational functions in n variables must be algebraically dependent; see [40] for details) shows that it admits at least one i/o equation of order n; therefore, $\delta(\Sigma) = n$. □

3.2.2. i/o operators.

Next we consider i/o equations for i/o operators rather than for state-space systems. By an *i/o operator* we mean an i/o map given by a convergent generating series. For a detailed definition of i/o operators, we refer the reader to [41]. We say an i/o operator F satisfies an i/o equation (17) if every $\mathcal{C}^k$ i/o pair (u, y) of F satisfies (17).

For any given operator F, we define $\delta(F)$ to be the lowest possible order of an i/o equation for F. Again, in the case when there is no i/o equation for F, $\delta(F)$ is defined to be $+\infty$.

An operator F is said to be *realized* by an initialized analytic system

$$(\mathcal{M}, x_0, \{g_0, g_1, \ldots, g_m\}, h)$$

if every i/o pair (u, y) of F satisfies the equations

$$\begin{aligned} x'(t) &= g_0(x(t))u(t) + \sum_{i=1}^m g_i(x(t))u_i(t), \quad x(0) = x_0, \\ y(t) &= h(x(t)) \end{aligned}$$

for t small enough.

Let $\lambda(F)$ be the Lie rank of F, as defined in [11], [19], or [26]. It is well known that F is realizable if and only if $\lambda(F) < \infty$, and the dimension of any canonical realization for F is $\lambda(F)$; cf. [11] and [34]. Here, by a *canonical* realization we mean a realization by an accessible and generically locally observable system.

PROPOSITION 3.8. *Assume that F is an i/o operator. Then:*

(a) $\lambda(F) \leq \delta(F)$;

(b) *if there exists a rational canonical realization for F, then $\lambda(F) = \delta(F)$.*

Proof. It was shown in [41] that if $\delta(F) < \infty$, then $\lambda(F) < \infty$. Thus we may assume that $\lambda(F) < \infty$, and in this case, one knows that F is realizable by some canonical system $\Sigma = (\mathcal{M}, x_0, \{g_0, g_1, \ldots, g_m\}, h)$.

By Remark 4.2 and Lemma 4.3 in [41], one knows that F admits i/o equation (17) if and only if (17) holds at any point t at which $u^{(k-1)}(t)$ exists. Combining this fact with the accessibility of the system, one sees that F admits i/o equation (17) if and only if (20) holds for Σ for all x in an open subset $\mathcal{N}$ of $\mathcal{M}$ and for all μ. On the other hand, it can be seen that (20) holds for all $x \in \mathcal{N}$ and all μ for system Σ if and only if (17) is an i/o equation for Σ as a system restricted to $\mathcal{N}$. Applying Theorem 3.7, we obtain the desired conclusion. □

4. Proof of Lemma 2.1.

In this section, we will prove Lemma 2.1. We will show first that there exists a generic subset $\mathcal{W}_1$ of $\mathbb{R}^{m,\infty}$ so that

$$\mathcal{F}(x) = \mathcal{F}_\mu(x) \tag{22}$$

for all x and $\mu \in \mathcal{W}_1$ and then that there is a generic subset $\mathcal{W}_2$ of $\mathbb{R}^{m,\infty}$ so that

$$\mathrm{d}\mathcal{F}(x) = \mathrm{d}\mathcal{F}_\mu(x) \tag{23}$$

for all x and all $\mu \in \mathcal{W}_2$. Then we just let $\mathcal{W} = \mathcal{W}_1 \bigcap \mathcal{W}_2$.

Proof of first part (*equation* (22)). For system (5), let

$$\mathcal{B} := \{x \in \mathcal{M} : \mathcal{F}(x) = 0\}.$$

To prove (22), we consider, for each subset $\mathcal{N}$ of the open subset $\mathcal{M} \setminus \mathcal{B}$, the set

$$\mathcal{G}_{\mathcal{N}} := \{\mu \in \mathbb{R}^{m,\infty} : \Psi(x, \mu) \neq 0, \forall x \in \mathcal{N}\},$$

where $\Psi(x, \mu) = (\psi_0(x, \mu), \psi_1(x, \mu), \ldots)$.

To prove the desired conclusion, it is enough to show that $\mathcal{G}_{\mathcal{N}}$ is open dense whenever $\mathcal{N}$ is a compact subset of $\mathcal{M} \setminus \mathcal{B}$ (since $\mathcal{M} \setminus \mathcal{B}$ can be written as a countable union of such subsets). In the following we let $\mathcal{N}$ be a fixed compact subset of $\mathcal{M} \setminus \mathcal{B}$, and we just write $\mathcal{G}$ instead of $\mathcal{G}_{\mathcal{N}}$. To show that $\mathcal{G}$ is dense, we need the following fact.

Let $r > 1$ be an integer. For each fixed vector $\nu^r = (\nu_0, \nu_1, \ldots, \nu_{r-1}) \in \mathbb{R}^{mr}$, we say that $\mu = (\mu_1, \mu_1, \ldots) \in \mathbb{R}^{m,\infty}$ is an extension of ν^r if $\mu_i = \nu_i$ for each $i \in [0, r-1]$.

LEMMA 4.1. *Let $x_0 \in \mathcal{M}$ and let ν^r be a fixed vector in $\mathbb{R}^{mr}$. If $\Psi(x_0, \mu) = 0$ for every extension μ of ν^r, then $x \in \mathcal{B}$.*

The proof of the above lemma will be given in Appendix A. We now return to show that $\mathcal{G}$ is dense. Take any open subset $\mathcal{U}$ of $\mathbb{R}^{m,\infty}$; without loss of generality, we may assume that $\mathcal{U} = \mathcal{U}_0 \times \mathcal{U}_1 \times \cdots \times \mathcal{U}_l \times \cdots$, where each $\mathcal{U}_i$ is an open subset of $\mathbb{R}^m$. For each integer $r > 0$, let $\mathcal{U}^r = \prod_{i=0}^{r-1} \mathcal{U}_i$. For each $\mu^r \in \mathcal{U}^r$, define

$$\mathcal{B}_{\mu^r} := \{x \in \mathcal{N} : \Psi_r(x, \mu^r) = 0\}$$

where $\Psi_r(x, \mu^r) = (\psi_0(x, \mu), \psi_1(x, \mu), \ldots, \psi_{r-1}(x, \mu))$ for any extension μ of μ^r. Note that Ψ_r is well defined because $\psi_i(x, \mu)$ does not depend on μ_j for $j \geq i$. For each finite jet ν, $\mathcal{B}_\nu$ is an analytic subset of $\mathcal{N}$, that is, a set defined by analytic equalities. As a consequence of the Weierstrass preparation theorem (in the form given for instance in [17, Thm. 2.7, Cor. 3]), one knows that analytic subsets of a compact set satisfy a descending chain condition. That is, if $\mathfrak{A}_1 \supseteq \mathfrak{A}_2 \supseteq \cdots \supseteq \mathfrak{A}_l \supseteq \cdots$ are analytic subsets of a compact set, then there exists some $r > 0$ such that $\mathfrak{A}_j = \mathfrak{A}_r$ for all $j \geq r$. From here it follows immediately that there is a minimal element $\mathcal{B}_{\bar\nu}$ of the family $\{\mathcal{B}_\nu\}$ in the sense that $\mathcal{B}_{\bar\nu} = \mathcal{B}_\nu$ whenever $\mathcal{B}_\nu \subset \mathcal{B}_{\bar\nu}$. Assume now that $\bar\nu \in \mathcal{U}^r$ provides such a minimal element.

Claim. $\mathcal{B}_{\bar\nu} = \emptyset$.

Assume that the above claim is not true. Then there exists some $x_0 \in \mathcal{N}$ such that $\Psi_r(x_0, \bar\nu) = 0$. Pick such an x_0. By Lemma 4.1, there exists some extension μ of $\bar\nu$ such that $\Psi(x_0, \mu) \neq 0$, so there exists some $l > r$ such that $\psi_l(x_0, \mu) \neq 0$. Write

$$\mu^l = (\bar\nu_0, \bar\nu_1, \ldots, \bar\nu_{r-1}, \bar\mu_r, \ldots, \bar\mu_{l-1}) \in \mathbb{R}^{ml}.$$

Note that $(\bar\nu_0, \bar\nu_1, \ldots, \bar\nu_{r-1}) \in \mathcal{U}^r$ by construction. For these fixed $\bar\nu_0, \ldots, \bar\nu_{r-1}$ and x_0, the function $\psi_l(x_0, \mu)$ does not depend on μ_j for $j \geq l$ and is analytic in $(\mu_r, \mu_{r+1}, \ldots, \mu_{l-1})$. Since it does not vanish at $(\bar\mu_r, \ldots, \bar\mu_{l-1})$, there is also some

$$(\tilde\mu_r, \tilde\mu_{r+1}, \ldots, \tilde\mu_{l-1}) \in \mathcal{U}_r \times \mathcal{U}_{r+1} \times \cdots \times \mathcal{U}_{l-1}$$

such that, for $\tilde\nu := (\bar\nu_0, \ldots, \bar\nu_{r-1}, \tilde\mu_r, \ldots, \tilde\mu_{l-1})$, $\psi_l(x_0, \tilde\mu) \neq 0$ for any extension $\tilde\mu$ of $\tilde\nu$, and hence, $\Psi_l(x_0, \tilde\nu) \neq 0$. So $x_0 \in \mathcal{B}_{\bar\nu} \setminus \mathcal{B}_{\tilde\nu}$. Also, obviously $\mathcal{B}_{\tilde\nu} \subseteq \mathcal{B}_{\bar\nu}$, since $\tilde\nu$ is an extension of $\bar\nu$. This contradicts the minimality of $\mathcal{B}_{\bar\nu}$. So we proved that $\Psi_r(x, \bar\nu) \neq 0$ for all $x \in \mathcal{N}$, as claimed.

Take any extension $\mu \in \mathcal{U}$ of μ^r to an infinite jet. Then $\Psi(x, \mu) \neq 0$ for all $x \in \mathcal{N}$, that is, $\mathcal{G} \cap \mathcal{U}$ is not empty for any open subset $\mathcal{U}$ of $\mathbb{R}^{m,\infty}$. Since $\mathcal{U}$ was arbitrary, one concludes that $\mathcal{G}$ is dense.

To prove the openness of $\mathcal{G}$, let

$$\mathcal{G}^r = \{\mu^r \in \mathbb{R}^{mr} : \ \Psi_r(x, \mu^r) \neq 0, \ \forall x \in \mathcal{N}\}.$$

By the compactness of $\mathcal{N}$, $\mathcal{G}^r$ is open. Let $\mathcal{G}_r = \mathcal{G}^r \times \mathbb{R}^{m,\infty}$. Then $\mathcal{G}_r$ is open. Since $\mathcal{G} = \bigcup_{r=1}^\infty \mathcal{G}_r$, it follows that $\mathcal{G}$ is open.

Proof of second part (*equation* (23)). Clearly $\mathrm{d}\mathcal{F}_\mu(x) \subseteq \mathrm{d}\mathcal{F}(x)$ for all x and μ, and for each $\mu \in \mathbb{R}^{m,\infty}$, $\mathrm{d}\mathcal{F}_\mu(x) = \mathrm{d}\mathcal{F}(x)$ if and only if

$$\ker \mathrm{d}\mathcal{F}(x) = \ker \mathrm{d}\mathcal{F}_\mu(x). \tag{24}$$

We now let

$$\widehat{\mathcal{B}} = \{(x, v) \in T\mathcal{M} : \ v \in \ker \mathrm{d}\mathcal{F}(x)\}.$$

Then $T\mathcal{M} \backslash \widehat{\mathcal{B}}$ is open. Let $\widehat\Psi(x, v, \mu) = (\widehat\psi_0(x, v, \mu), \widehat\psi_1(x, v, \mu), \ldots)$, where $\widehat\psi_i(x, v, \mu) = \mathrm{d}\psi_i(x, \mu)v$. To prove the desired conclusion, it is enough to show that there exists a generic subset $\mathcal{W}$ of $\mathbb{R}^{m,\infty}$ such that for any $\mu \in \mathcal{W}$, $\widehat\Psi(x, v, \mu) \neq 0$ for all $(x, v) \notin \widehat{\mathcal{B}}$. For this, it is enough to show that for any compact subset $\widehat{\mathcal{N}}$ of $T\mathcal{M} \setminus \widehat{\mathcal{B}}$, the set

$$\widehat{\mathcal{G}}_{\widehat{\mathcal{N}}} := \{\mu \in \mathbb{R}^{m,\infty} : \ \widehat\Psi(x, v, \mu) \neq 0, \ \forall \ (x, v) \in \widehat{\mathcal{N}}\}$$

is open dense. We now fix a compact subset $\widehat{\mathcal{N}}$ of $T\mathcal{M}\setminus\widehat{\mathcal{B}}$ and write $\widehat{\mathcal{G}}$ instead of $\widehat{\mathcal{G}}_{\widehat{\mathcal{N}}}$. Similar to the proof of the first part, we need the following conclusion to prove the density property of $\widehat{\mathcal{G}}$. The proof of the conclusion will again be provided in Appendix A.

LEMMA 4.2. *For any given fixed point* $(x,\, v) \in T\mathcal{M}$, *if* $\widehat{\Psi}(x,\, v,\, \mu) = 0$ *for all extensions* μ *of* ν^r, *for some* $\nu^r \in \mathbb{R}^{mr}$, *then* $(x,\, v) \in \widehat{\mathcal{B}}$.

To show the density of $\widehat{\mathcal{G}}$, we take any open subset $\mathcal{U}$ of $\mathbb{R}^{m,\infty}$. Again, without loss of generality, we can assume that $\mathcal{U} = \mathcal{U}_0 \times \mathcal{U}_1 \times \cdots \times \mathcal{U}_l \times \cdots$, where each $\mathcal{U}_i$ is an open subset of $\mathbb{R}^m$. Using the same notions for μ^r and $\mathcal{U}^r$ as used before, we define

$$\widehat{\mathcal{B}}_{\mu^r} := \left\{(x,\, v) \in \widehat{\mathcal{N}} :\ \widehat{\Psi}_r(x,\, v,\, \mu^r) = 0\right\}$$

where $\widehat{\Psi}_r(x,\, \mu^r) = (\widehat{\psi}_0(x,\, v,\, \mu),\, \widehat{\psi}_1(x,\, v,\, \mu),\, \ldots,\, \widehat{\psi}_{r-1}(x,\, v,\, \mu))$ for any extension μ of μ^r. For each finite jet ν, $\widehat{\mathcal{B}}_\nu$ is an analytic subset of $\widehat{\mathcal{N}}$ (with the obvious analytic manifold structure on the tangent bundles). Using the same argument as before, one knows that there exists a minimal element of the family $\{\widehat{\mathcal{B}}_\nu\}$. Let $\nu \in \mathcal{U}^s$ be such that $\widehat{\mathcal{B}}_\nu$ is a minimal element.

Claim. $\widehat{\Psi}_s(x,\, v,\, \nu) \neq 0$ for all $(x,\, v) \in \widehat{\mathcal{N}}$.

Assume that the claim is not true. Then there exists some $(x_0,\, v_0) \in \widehat{\mathcal{N}}$ such that $\widehat{\Psi}_s(x_0,\, v_0,\, \nu) = 0$. By Lemma 4.2, there exists some extension μ of ν such that $\widehat{\Psi}(x_0,\, v_0,\, \mu) \neq 0$. This means there exists some $l \geq s$ such that $\widehat{\psi}_l(x_0,\, v_0,\, \mu) \neq 0$. By analyticity of $\widehat{\psi}_l$, one knows that there exists some

$$(\tilde{\mu}_s,\, \tilde{\mu}_{s+1},\, \ldots,\, \tilde{\mu}_{l-1}) \in \mathcal{U}_s \times \mathcal{U}_{s+1} \times \cdots \times \mathcal{U}_{l-1}$$

such that, for $\tilde{\mu} := (\nu_0,\, \ldots,\, \nu_{s-1},\, \tilde{\mu}_s,\, \ldots,\, \tilde{\mu}_{l-1})$, $\widehat{\Psi}_l(x_0,\, v_0,\, \tilde{\mu}) \neq 0$. So $(x_0,\, v_0) \in \widehat{\mathcal{B}}_\nu \setminus \widehat{\mathcal{B}}_{\tilde{\mu}}$. Also, obviously $\widehat{\mathcal{B}}_{\tilde{\mu}} \subseteq \widehat{\mathcal{B}}_\nu$, since $\tilde{\nu}$ is an extension of ν. This contradicts the minimality of $\widehat{\mathcal{B}}_\nu$. So we proved that $\widehat{\Psi}_r(x,\, v,\, \nu) \neq 0$ for all $(x,\, v) \in \widehat{\mathcal{N}}$. Noting then that for any extension μ of ν, $\widehat{\Psi}(x,\, v,\, \mu) \neq 0$ for any $(x,\, v) \in \widehat{\mathcal{N}}$, we conclude that $\mathcal{G} \bigcap \mathcal{U} \neq \emptyset$. This proves the density of $\mathcal{G}$.

To prove the openness of $\widehat{\mathcal{G}}$, we again let

$$\widehat{\mathcal{G}}^r = \{\mu^r \in \mathbb{R}^{mr} :\ \widehat{\Psi}_r(x,\, v,\, \mu^r) \neq 0,\ \forall (x,\, v) \in \widehat{\mathcal{N}}\}.$$

By compactness of $\widehat{\mathcal{N}}$, $\widehat{\mathcal{G}}^r$ is open. Let $\widehat{\mathcal{G}}_r = \widehat{\mathcal{G}}^r \times \mathbb{R}^{m,\infty}$. Then $\widehat{\mathcal{G}}_r$ is open. Since $\widehat{\mathcal{G}} = \bigcup_{r=1}^{\infty} \widehat{\mathcal{G}}_r$, it follows that $\mathcal{G}$ is open. The proof of Lemma 2.1 is then complete.

Finally, we remark that also with respect to the weak topology on $\mathbb{R}^{m,\infty}$, $\mathcal{G}_{\mathcal{N}}$ and $\widehat{\mathcal{G}}_{\widehat{\mathcal{N}}}$ are still open and dense. Density is obvious, as they are dense with respect to a stronger topology. The openness of $\mathcal{G}_{\mathcal{N}}$ and $\widehat{\mathcal{G}}_{\widehat{\mathcal{N}}}$ follows from the compactness of $\mathcal{N}$ and $\widehat{\mathcal{N}}$. Thus, the conclusions of Lemma 2.1 also hold with respect to the weak topology on $\mathbb{R}^{m,\infty}$.

4.1. Proof of Theorem 3.1. In this section, we provide a proof for Theorem 3.1.

To study the observability for system (5), we consider the system

$$\begin{aligned} \xi' &= \tilde{g}_0(\xi) + \sum_{i=1}^{m} \tilde{g}_i(\xi) u_i\,, \\ y &= \tilde{h}(\xi)\,, \end{aligned} \tag{25}$$

where

$$\xi = \begin{pmatrix} x \\ z \end{pmatrix} \in \mathcal{M} \times \mathcal{M}, \quad \tilde{g}_i(\xi) = \begin{pmatrix} g_i(x) \\ g_i(z) \end{pmatrix}, \; 0 \leq i \leq m,$$

and $\tilde{h}(\xi) = h(x) - h(z)$. Clearly, $x \not\sim_u z$ for system (5) if and only if $\xi \not\sim_u 0$ for system (25). Thus, to prove Theorem 3.1, it is enough to establish the following conclusion.

PROPOSITION 4.3. *Assume that for an analytic system* (5)*, the i/o map induced by the zero initial state is a zero map, that is,* $h \circ \varphi(t, 0, u) = 0$ *for all* t *and all* u*. Then for any* $T > 0$*, the set*

$$\mathfrak{G} = \{u \in \mathcal{C}^\infty[0,\ T] :\ x \not\sim_u 0 \text{ if } x \not\sim 0\}$$

is a generic subset of $\mathcal{C}^\infty[0,\ T]$.

Proof. Let $\mathcal{B} := \{x :\ x \sim 0\}$. Then $\mathcal{M} \setminus \mathcal{B}$ is an open subset of $\mathcal{M}$. To prove Proposition 4.3, it is enough to show that for every compact subset $\mathcal{N}$ of $\mathcal{M} \setminus \mathcal{B}$ the set

$$\mathfrak{G}_{\mathcal{N}} = \{u \in \mathcal{C}^\infty[0,\ T] :\ x \not\sim_u 0 \text{ for all } \ x \in \mathcal{N}\}$$

is an open dense subset of $\mathcal{C}^\infty[0,\ T]$.

Note that for $u \in \mathcal{C}^\infty$, $x \not\sim_u 0$ if $\psi_i(x,\ \mu) \neq 0$ for some i, where $\mu = (\mu_0,\ \mu_1,\ \ldots) \in \mathbb{R}^\infty$ with $\mu_i = u^{(i)}(0)$, and ψ_i is as defined in (7) for each i. Also, by Theorem 3-1.5 in [19], one knows that for $x \in \mathcal{M}$, if $\mathcal{F}(x) = 0$, then $x \in \mathcal{B}$. This means that for each $x \in \mathcal{N}$, $\mathcal{F}(x) \neq 0$. Thus, by Lemma 2.1, there exists a dense subset $\mathcal{G}$ of $\mathbb{R}^{m,\infty}$ such that $\Psi(x,\ \mu) \neq 0$ for all $x \in \mathcal{N}$ and all $\mu \in \mathcal{G}$.

To complete the proof of Proposition 4.3, we need to show that $\mathfrak{G}_{\mathcal{N}}$ is an open dense subset of $\mathcal{C}^\infty[0,\ T]$.

Take $\bar{\omega} \in \mathcal{C}^\infty[0,\ T]$, and let $\mathcal{U}$ be a neighborhood of $\bar{\omega}$. Without loss of generality when showing the density of $\mathfrak{G}_{\mathcal{N}}$, we may assume that

$$\mathcal{U} = \left\{\omega \in \mathcal{C}^\infty[0,\ T] :\ \max_{0 \leq i \leq k} |w^{(i)}(t) - \bar{\omega}^{(i)}(t)| < \delta,\ t \in [0,\ T]\right\},$$

for some integer $k \geq 0$ and some $\delta > 0$.

Let $\bar{\mu} := (\bar{\mu}_0,\ \bar{\mu}_1,\ \ldots)$, where $\bar{\mu}_i := \bar{\omega}^{(i)}(0)$, and let $\mathcal{W}$ be the open subset of $\mathbb{R}^{m,\infty}$ defined by

$$\mathcal{W} = \left\{\mu \in \mathbb{R}^{m,\infty} :\ |\mu_i - \bar{\mu}_i| < \delta e^{-T},\ i \geq 0\right\}.$$

As $\mathcal{W} \bigcap \mathcal{G} \neq \emptyset$, there exists some $\nu \in \mathcal{W}$ such that $\Psi(x,\ \nu) \neq 0$ for all $x \in \mathcal{N}$. By compactness of $\mathcal{N}$, there exists some $r > 0$ such that

$$\Psi_r(x,\ \nu^r) \neq 0, \text{ for any } x \in \mathcal{N}. \tag{26}$$

Without loss of generality, one can always assume that $r > k$.

Now let $\bar{\omega}_0(t) := \bar{\omega}(t) - \sum_{i=0}^{r-1} \frac{\bar{\mu}_i}{i!} t^i$. Note then that $\bar{\omega}_0^{(i)}(0) = 0$ for all $0 \leq i \leq r - 1$.

Finally, we define

$$\omega(t) := \bar{\omega}_0(t) + \sum_{i=0}^{r-1} \frac{\nu_i}{i!} t^i.$$

Then, for $0 \leq i \leq k$ and $0 \leq t \leq T$, we have

$$|\omega^{(i)}(t) - \bar{\omega}^{(i)}(t)| \leq \sum_{j=i}^{r-1} \frac{|\nu_j - \bar{\mu}_j|}{(j-i)!} t^{j-i} \leq \sum_{j=0}^{\infty} \frac{|\nu_{j+i} - \bar{\mu}_{j+i}|}{j!} t^j$$
$$< \delta e^{-T} e^t \leq \delta .$$

Thus, $\omega \in \mathcal{U}$.

On the other hand, (26) implies that for every $x \in \mathcal{N}$, there exists some $i \leq r-1$, such that

$$\left. \frac{d^i}{dt^i} \right|_{t=0} h(\varphi(t, x, w)) \neq 0 .$$

From here it follows that $x \not\sim_{\omega} 0$ for every $x \in \mathcal{N}$, that is, $\omega \in \mathfrak{G}_{\mathcal{N}}$. This proves that $\mathfrak{G}_{\mathcal{N}}$ is dense.

We then conclude the proof of Proposition 4.3 by noting that the openness of $\mathfrak{G}_{\mathcal{N}}$ follows from the compactness of $\mathcal{N}$. □

Remark 4.4. Note that the above proof only depends on the first half of Lemma 2.1, i.e., formula (22), and the proof of (22) is fairly straightforward (though it calls upon some notions and elementary results from the theory for generating series).

Proof of Proposition 3.2. As indicated in the beginning of this section, it is enough to show the following:

> Assume that for an analytic system (5), the i/o map induced by the zero initial state is a zero map, that is, $h \circ \varphi(t, 0, u) = 0$ for all t and all u. Then there exists some analytic input u such that $x \not\sim_u 0$ for all $x \not\sim 0$.

Proof. Consider the following open subset of $\mathbb{R}^{m,\infty}$:

$$\mathcal{U} = \mathcal{U}_0 \times \mathcal{U}_1 \times \mathcal{U}_2 \times \cdots$$

where $\mathcal{U}_i = (-1, 1)$ for all $i \geq 0$. By Lemma 2.1, there is at least one jet μ in $\mathcal{U}$ such that $\mathcal{F}_\mu(x) = \mathcal{F}(x)$, from which it follows that

$$\Psi(x, \mu) \neq 0, \quad \forall x \not\sim 0. \tag{27}$$

Now let

$$u(t) = \sum_{i=0}^{\infty} \frac{\mu_i}{i!} t^i .$$

Then u is an analytic function and $u^{(i)}(0) = \mu_i$. By (27), one knows that $x \not\sim_u 0$ for all $x \not\sim 0$. □

5. Main results for discrete-time systems.

In this section, we discuss our main results for discrete-time systems.

5.1. Basic definitions for discrete-time systems.

We consider *analytic systems* as in (4), where for each t, $x(t) \in \mathcal{M}$, an analytic manifold, and $u(t) \in \mathbb{R}^m$. We assume that $h : \mathcal{M} \to \mathbb{R}^p$ and $f : \mathcal{M} \times \mathbb{R}^n \to \mathcal{M}$ are analytic. If $\mathcal{M} = \mathbb{R}^n$ and the entries of f and h are rational functions with no (real) poles, then we call (4) a *rational system*. A system Σ will be called *reversible* if $f(\cdot, u)$ is one-to-one, for each fixed $u \in \mathbb{R}^m$. (Reversible systems are a more general class than the systems usually called *invertible* in the discrete-time controllability literature, for which one makes the stronger requirement that $f(\cdot, u)$ is a diffeomorphism of $\mathcal{M}$, for each u. Invertible systems arise naturally through the sampling of continuous-time systems in digital control, by integrating flows over a sampling period; their controllability properties were studied in, among other papers, [20], [12], [21], [24], [22], [23], [1].)

For each control sequence $\omega \in \mathrm{I\!R}^{km}$, we define $f^\omega : \mathcal{M} \to \mathcal{M}$ inductively by $f^e(x) = x$ for the empty sequence e and $f^{\omega u}(x) = f(f^\omega(x), u)$. We also let $h^\omega := h \circ f^\omega$. For $\mu = (\mu_0, \mu_1, \dots) \in \mathrm{I\!R}^{m,\infty}$, we let $H^\mu(x) := (h(x), h^{\mu_0}(x), h^{\mu_0\mu_1}(x), \dots)$.

Two states p and q are said to be distinguished by $\mu \in \mathrm{I\!R}^{m,\infty}$, denoted by $p \not\sim_\mu q$, if $H^\mu(p) \neq H^\mu(q)$. A discrete-time system is said to be *observable* if any two distinct states p and q can be distinguished by some μ. See [27] for a detailed introduction to observability and related concepts and, in particular, [25] for results on observability of discrete-time systems.

For an analytic system, we define the *observation space* of Σ as the following subspace of the space of analytic functions defined on $\mathcal{M}$:

$$\mathcal{F} = \operatorname{span}_{\mathrm{I\!R}} \{h^\omega : \ \omega \in \mathrm{I\!R}^{mr}, r \geq 0\}.$$

This space plays an important role in studying observability of discrete-time systems; see, e.g., [28] and [27]. See also [16] for related algebraic structures.

Associated with the above space, for each $x \in \mathcal{M}$ we let $\mathrm{d}\mathcal{F}(x)$ be the subspace of the cotangent space at x defined by

$$\mathrm{d}\mathcal{F}(x) = \{\mathrm{d}\alpha(x) : \ \alpha \in \mathcal{F}\} .$$

In analogy to the continuous-time case, we define, for each $\mu = (\mu_0, \mu_1, \dots) \in \mathrm{I\!R}^{m,\infty}$, the following subspace $\mathcal{F}_\mu$ of analytic functions:

$$\mathcal{F}_\mu = \operatorname{span}_{\mathrm{I\!R}} \{h, h^{\mu_0}, h^{\mu_0\mu_1}, \dots\}.$$

For each $\mu \in \mathrm{I\!R}^{m,\infty}$ and each $x \in \mathcal{M}$, we also consider

$$\mathrm{d}\mathcal{F}_\mu(x) = \{\mathrm{d}\alpha : \ \alpha \in \mathcal{F}_\mu\}.$$

Clearly, $\mathcal{F} = \sum_\mu \mathcal{F}_\mu$ and $\mathrm{d}\mathcal{F}(x) = \sum_\mu \mathrm{d}\mathcal{F}_\mu(x)$ for each x. Here we will need the following result.

LEMMA 5.1. *Assume that* (4) *is reversible and observable. Then there exists a generic subset* $\mathcal{W}$ *of* $\mathrm{I\!R}^{m,\infty}$ *such that for each* $\mu \in \mathcal{W}$,

$$\mathrm{d}\mathcal{F}(x) = \mathrm{d}\mathcal{F}_\mu(x) = \mathrm{I\!R}^n, \tag{28}$$

for all x *in an open dense subset of* $\mathcal{M}$.

The proof will be given later; it will rely on a result about universal inputs for discrete-time systems that is presented in the next section.

Assume now that $\mathcal{M} = \mathrm{I\!R}^n$. Still using the notation used in §2.2, we introduce

$$\widehat{\mathrm{d}}\mathcal{F} := \operatorname{span}_{\mathrm{I\!R}_x}\{\mathrm{d}\alpha : \ \alpha \in \mathcal{F}\}, \ \ \widehat{\mathrm{d}}\mathcal{F}_\mu := \operatorname{span}_{\mathrm{I\!R}_x}\{\mathrm{d}\alpha : \ \alpha \in \mathcal{F}_\mu\}.$$

From the lemma and using an argument analogous to that used in proving Corollary 2.4, we have the following corollary.

COROLLARY 5.2. *For an analytic, reversible, and observable system,* $\widehat{\mathrm{d}}\mathcal{F}_\mu = \widehat{\mathrm{d}}\mathcal{F}$ *for all* μ *in a generic set of* $\mathrm{I\!R}^{m,\infty}$.

5.2. Observability and universal inputs. An input sequence is said to be a *universal input* of a discrete-time system Σ if it distinguishes every distinguishable pair of Σ.

THEOREM 5.3. *Assume that* (4) *is analytic, reversible, and observable. Then the universal inputs of* (4) *form a generic subset of* $\mathbb{R}^{m,\infty}$.

Proof. First of all, we let

$$\mathcal{D} := \{(x, x) : \; x \in \mathcal{M}\} \subseteq \mathcal{M} \times \mathcal{M}.$$

By observability, every pair $(x, z) \in (\mathcal{M} \times \mathcal{M}) \setminus \mathcal{D}$ is a distinguishable pair of (4). For each $\nu \in \mathbb{R}^{mr}$, we let $\lambda_r(x, z, \nu) = h^\nu(x) - h^\nu(z)$, and we also let $\lambda_0(x, z) = h(x, z)$. For each $\mu = (\mu_0, \mu_1, \ldots)$, we define

$$\Lambda(x, z, \mu) = (\lambda_0(x, z), \lambda_1(x, z, \mu_0), \lambda_2(x, z, \mu_1\mu_0), \ldots).$$

To prove the desired conclusion, it is enough to show that for each compact subset $\mathcal{N}$ of $(\mathcal{M} \times \mathcal{M}) \setminus \mathcal{D}$, the set $\mathcal{G}_\mathcal{N}$ defined by

$$\mathcal{G}_\mathcal{N} := \{\mu \in \mathbb{R}^{m,\infty} : \Lambda(x, z, \mu) \neq 0, \; \forall (x, z) \in \mathcal{N}\}$$

is an open dense subset of $\mathbb{R}^{m,\infty}$.

For each open subset $\mathcal{U}$ of $\mathbb{R}^{m,\infty}$ given by $\mathcal{U}_0 \times \mathcal{U}_1 \times \cdots$, consider, for each $\nu \in \mathcal{U}^s = \prod_{i=0}^{s-1} \mathcal{U}_i$, the subset $\mathcal{B}_\nu$ of $\mathcal{N}$ defined by

$$\mathcal{B}_\nu = \{(x, z) \in \mathcal{N} : \; \Lambda_r(x, z, \nu) = 0\},$$

where $\Lambda_r(x, z, \nu) = (\lambda_0(x, z), \lambda_1(x, z, \nu_0), \ldots, \lambda_s(x, z, \nu))$. Using the same argument as that employed in the proof of Lemma 2.1, we know that there exists a minimal element $\mathcal{B}_{\bar\nu}$ of the family $\{\mathcal{B}_{\bar\nu}\}$. Suppose $\bar\nu \in \mathcal{U}^r$. We next show that $\bar\nu$ distinguishes every pair $(x, z) \in \mathcal{N}$. Assume that there would exist a pair $(x_0, z_0) \in \mathcal{N}$ such that $x_0 \sim_{\bar\nu} z_0$. Since (4) is reversible, $x_1 \neq z_1$, where $x_1 = f^\nu(x_0)$ and $z_1 = f^\nu(z_0)$. By observability of (4), one knows that there exists some $\tilde\nu \in \mathbb{R}^{ms}$ such that $x_1 \not\sim_{\tilde\nu} z_1$. Let $\widehat\nu = \tilde\nu\bar\nu$ (concatenation of sequences); then it follows that $\Lambda_{r+s}(x_0, z_0, \bar\nu\tilde\nu) \neq 0$. By the analyticity of Λ_{r+s} when fixing x_0, z_0 and $\bar\nu$, one knows that there exists some $\widehat\nu \in \mathcal{U}_r \times \cdots \times \mathcal{U}_{r+s-1}$ such that $\Lambda_{r+s}(x_0, z_0, \bar\nu\widehat\nu) \neq 0$. This implies that $(x_0, z_0) \in \mathcal{B}_{\bar\nu} \setminus \mathcal{B}_{\bar\nu\widehat\nu}$, which, in turn, implies that $\mathcal{B}_{\bar\nu\widehat\nu}$ is a proper subset of $\mathcal{B}_{\bar\nu}$, contradicting the assumed minimality of $\mathcal{B}_{\bar\nu}$. Thus, we showed that $\Lambda_r(x, z, \bar\nu) \neq 0$ for any $(x, z) \in \mathcal{N}$. Clearly, any extension μ of $\bar\nu$ in $\mathcal{U}$ is an element of $\mathcal{G}_\mathcal{N}$. This shows that $\mathcal{G}_\mathcal{N} \bigcap \mathcal{U} \neq \emptyset$ for any open subset $\mathcal{U}$ of $\mathbb{R}^{m,\infty}$. The density of $\mathcal{G}_\mathcal{N}$ is thus proved.

Again as in the proof of Lemma 2.1 for the continuous case, $\mathcal{G}_\mathcal{N}$ is open since $\mathcal{N}$ is compact. □

In the statement of Theorem 5.3, we assumed more than we did in its continuous counterpart, Theorem 3.1 (and also concluded slightly less). One of the extra conditions is observability. We needed to impose this because the counterpart of Lemma 4.1 is not available in the discrete-time case. The discrete case analogy would be that any distinguishable pair is again carried to a distinguishable pair by the flow of the system, no matter which input is applied. Unfortunately, this not true in general. The following example, suggested by F. Albertini, shows that distinguishable pairs can be carried to indistinguishable pairs. (Note that this can never happen with analytic continuous-time systems.)

Example 5.4. Consider the system

$$x(t+1) = x(t) + 1, \quad y(t) = h(x(t)), \tag{29}$$

where $h(x)$ is defined by

$$h(x) = \begin{cases} \frac{\sin \pi x}{\pi x} & \text{if } x \neq 0, \\ 1 & \text{if } x = 0. \end{cases}$$

Clearly the system is analytic and reversible. However, the distinguishable pair (0, 1) is carried to an indistinguishable pair after $t = 1$.

Proof of Lemma 5.1. To obtain the desired conclusion, it is enough to show that (28) holds in an open dense subset of $\mathcal{M}$ for every universal input μ (since universal inputs themselves form a generic subset).

Fix any universal input μ. By observability, one knows that $H(\cdot, \mu)$ is a one-to-one map. Let $k = \max_p \dim \mathcal{F}_\mu(p)$. It is sufficient to show that $k = n$. But this is an immediate consequence of Lemma B.1 (see Appendix B), applied to $\{h, h^{\mu_0}, h^{\mu_0\mu_1}, \ldots\}$ seen as a family of maps. □

5.2.1. Orders of i/o equations. We say that the discrete-time system (4) *admits the i/o equation* such as that if (3) holds for all input/output pairs of (4) (for $t \geq r$ and any possible initial state $x(0)$). The *order* of the equation is r if

$$\frac{\partial}{\partial \nu_r} E(\mu_0, \ldots, \mu_{k-1}, \nu_0, \nu_1, \ldots, \nu_k)$$

is not a zero function. For any given system Σ, we let $\delta(\Sigma)$ be the lowest possible order of an i/o equation that Σ admits. If there is no such equation, $\delta(\Sigma)$ is defined to be ∞. Following the same outline as in the proof of Theorem 3.7 but now using Lemma 5.1, we conclude as follows.

THEOREM 5.5. *Let Σ be an n-dimensional analytic system. Assume, further, that Σ is reversible and observable. Then $\delta(\Sigma) \geq n$. If, in addition, Σ is a rational system, then $\delta(\Sigma) = n$.*

Remark 5.6. The result in Lemma 5.1 is *false* if the assumption of reversability is dropped, as discussed in [28]. As a consequence of this, the above conclusions may be false without the invertibility assumption. To illustrate this, consider the following system of dimension 3:

$$\begin{aligned} &x_1(t+1) = u(t) \ , \quad x_2(t+1) = x_3(t) \ , \\ &x_3(t+1) = x_3(t)x_1(t) + x_1(t) + x_2(t)u(t) \ , \\ &y(t) = x_3(t). \end{aligned}$$

This is an observable polynomial system. However, it admits an equation of order 2:

$$y(t) = y(t-1)u(t-2) + y(t-2)u(t-1) + u(t-2).$$

Note that this system is not reversible.

Appendix A. Proofs of two lemmas. In this appendix, we will prove Lemmas 4.1 and 4.2. For this, we need to recall some basic definitions and properties of i/o operators defined by convergent generating series. For a detailed study of generating series and i/o operators, we refer the reader to [41].

Let m be a fixed integer and $I = \{0, 1, \ldots, m\}$. For any integer $k \geq 1$, we define I^k to be the set of all sequences $i_1 i_2 \ldots i_k$, where $i_s \in I$ for each s. We use I^0 to denote the set whose only element is the empty sequence ϕ. Let $I^* = \bigcup_{k \geq 0}^{\infty} I^k$.

A *generating series*

$$c = \sum_{\iota \in I^*} \langle c, \eta_\iota \rangle \eta_\iota$$

is a formal power series in the noncommutative variables $\eta_0, \eta_1, \dots, \eta_m$ for some fixed number m, where we use the notation $\eta_\iota = \eta_{i_1}\eta_{i_2}\dots\eta_{i_l}$ for each multiindex $\iota = i_1 i_2 \dots i_l$. The coefficients $\langle c, \eta_\iota \rangle$ are assumed to be real.

We shall say that a power series c is *convergent* if there exist K, $M \geq 0$ such that

$$|\langle c, \eta_\iota \rangle| \leq K M^k k! \quad \text{for each } \iota \in I^k \text{ and each } k \geq 0. \tag{30}$$

For any fixed real number $T > 0$, let $\mathcal{U}_T$ be the set of all essentially bounded measurable functions

$$u : [0, T] \to \mathbb{R}^m$$

endowed with the L^1 norm. We write $\|u\|_1$ for $\max\{\|u_i\|_1 :, 1 \leq i \leq m\}$ and $\|u\|_\infty$ for $\max\{\|u_i\|_\infty :, 1 \leq i \leq m\}$ where u_i is the ith component of u and $\|u_i\|_1$ is the L^1 norm of u_i, $\|u_i\|_\infty$ is the L^∞ norm of u_i. For each $u \in \mathcal{U}_T$ and each $\iota \in I^l$, we define inductively the functions

$$V_\iota = V_\iota[u] \ \in \ \mathcal{C}[0, T]$$

by

$$V_{i_1 \cdots i_{l+1}}[u](t) = \int_0^t u_{i_1}(s) V_{i_2 \dots i_{l+1}}(s)\, ds,$$

where $V_\phi = 1$ and u_i is the ith coordinate of $u(t)$ for $i = 1, 2, \dots, m$ and $u_0(t) \equiv 1$.

For each formal power series c in $\eta_0, \eta_1, \dots, \eta_m$, we define a formal operator on $\mathcal{U}_T$ in the following way:

$$F_c[u](t) = \sum \langle c, \eta_\iota \rangle V_\iota[u](t). \tag{31}$$

If the series is convergent and (30) holds, then it is known that for any

$$T < (\|u\|_\infty (Mm + M))^{-1},$$

the series (31) converges uniformly and absolutely for all $t \in [0, T]$. Let

$$\mathcal{V}_T := \{u \in L^m_\infty : \ \|u\|_\infty T < (Mm + M)^{-1}\}.$$

We refer the reader to [41] for the proof of the following lemmas.

LEMMA A.1. *Assume that c is a convergent power series. Then the operator*

$$F_c : \ \mathcal{V}_T \ \to \ \mathcal{C}[0, T]$$

is continuous with respect to the L^1 norm in $\mathcal{V}_T$ and the $\mathcal{C}^0$ norm in $\mathcal{C}[0, T]$.

LEMMA A.2. *Suppose c is a convergent series. Then $F_c[u]$ is analytic if $u \in \mathcal{V}_T$ is analytic.*

For each convergent series c, we let, for each $\mu \in \mathbb{R}^{m,\infty}$ and each integer $i \geq 0$,

$$c_i(\mu) \ = \ \left.\frac{d^i}{dt^i}\right|_{t=0} F_c[u](t)\,, \tag{32}$$

where u is any smooth input with $u^{(i)}(0) = \mu_i$. Note that $c_i(\mu)$ is a polynomial in μ and $c_i(\mu)$ doesn't depend on μ_j for $j \geq i$.

By Lemma 2.1 in [39], one knows that for a convergent series c, $F_c[u] = 0$ for every piecewise constant input u if and only if $c = 0$. On the other hand, it is not hard to see that for each piecewise constant function u, there exists a sequence $\{u_j\}$ of analytic functions such that $\|u_j\| \leq \|u\|$ and $u_j \to u$ as $j \to \infty$ in the L_1 norm. By Lemma A.2, one concludes that $F_c[u] = 0$ for every analytic input u if and only if $c = 0$. Since $F_c[u]$ is analytic if u is analytic, it then follows from (32) that for an analytic u with $u^{(i)}(0) = \mu_i$, $F_c[u] = 0$ if and only if $c_i(\mu) = 0$ for all $i \geq 0$. Thus we conclude that $c = 0$ if and only if $c_i(\mu) = 0$ for all μ and all i. To prove the desired conclusions, we need the following well-known fact.

LEMMA A.3. *Assume that f is a continuous function defined on* $[0,\ t_0]$ *for some* $t_0 > 0$. *Then for any given integer r and any vector* $(w_0, w_1, \ldots, w_r)$, *there exists a L_∞-bounded sequence of analytic functions f_j defined on* $[0,\ t_0]$, *such that $f_j^{(i)}(0) = w_i$ for all $i \leq r$ and f_j converges to f in the L_1 norm.*

Proof. For the given vector, let

$$\widehat{f}(t) = f(t) - \sum_{i=0}^{r} \frac{w_i t^i}{i!}.$$

Without loss of generality, one may assume that $\widehat{f}(0) = 0$. Otherwise, one can always choose a L_∞-bounded sequence of continuous functions $\widehat{f}_j$ converging to $\widehat{f}$ in the L_1 norm and such that $\widehat{f}_j(0) = 0$. Now one may apply Lemma 4.3 in [41] to $\widehat{f}$ to conclude that there exists a sequence $\tilde{f}_j$ converging to $\widehat{f}$ uniformly (hence also in L_1 norm) with the property that $\tilde{f}_j^{(i)}(0) = 0$ for all $i \leq r$. Then the functions

$$f_j(t) := \tilde{f}_j(t) + \sum_{i=0}^{r} \frac{w_i t^i}{i!}$$

give the desired sequence. □

Combining the above conclusion and Lemma A.1, one proves the following.

LEMMA A.4. *Assume that c is a convergent series and that r is an integer. Let $\bar{\mu}^r$ be a given vector in $\mathbb{R}^{mr}$. If for every extension μ of $\bar{\mu}^r$, $c_i(\mu) = 0$ for all i, then* $c = 0$.

Proof of Lemma 4.1. For analytic system (5) and for each $x \in \mathcal{M}$, we define a generating series by letting

$$\langle c^x, \eta_{i_1}\eta_{i_2}\cdots\eta_{i_r}\rangle = L_{g_{i_r}}\cdots L_{g_{i_2}} L_{g_{i_1}} h(x). \tag{33}$$

By Lemma 4.2 in [36], such a series is always convergent, and it follows from Theorem 3-1.5 in [19] that for any $\mu \in \mathbb{R}^{m,\infty}$,

$$\psi_i(x,\ \mu) = c_i(x,\ \mu), \tag{34}$$

where

$$c_i(x,\ \mu) = \left.\frac{d^i}{dt^i}\right|_{t=0} F_{c^x}[u](t)\,.$$

The conclusion of Lemma 4.1 then follows from Lemma A.4.

Proof of Lemma 4.2. For analytic system (5), instead of considering the series defined by (33), we consider, for each $(x, v) \in T\mathcal{M}$, the series defined by

$$(35) \qquad \langle d(x, v), \eta_{i_1}\eta_{i_2}\cdots\eta_{i_r}\rangle = \mathrm{d}L_{g_{i_r}}\cdots L_{g_{i_2}}L_{g_{i_1}}h(x)v.$$

Claim. For each (x, v), the series $d(x, v)$ is a convergent series.

First of all, by Lemma 4.2 in [36], there is some constant $M_0 > 0$ such that for $g_0(x), g_1(x), \ldots, g_m(x)$, $v \in T_x\mathcal{M}$, there exists some $M_0 > 0$ such that

$$(36) \quad \left|\mathrm{d}L_{g_{i_1}}L_{g_{i_2}}\cdots L_{g_{i_r}}h(x)v\right| = \left|L_vL_{g_{i_1}}L_{g_{i_2}}\cdots L_{g_{i_r}}h(x)v\right| \leq M_0^{r+1}(r+1)!.$$

It is then not hard to see that there exist some constants K and $M > M_0$ such that

$$\left|\mathrm{d}L_{g_{i_1}}L_{g_{i_2}}\cdots L_{g_{i_r}}h(x)v\right| \leq KM^r r!,$$

for all $r > 0$. Therefore $d(x, v)$ is a convergent series for each pair (x, v).

For each smooth input u with $u^{(i)}(0) = \mu_i$, let

$$d_i(x, v, \mu) = \left.\frac{d^i}{dt^i}\right|_{t=0} F_{d(x, v)}[u](t).$$

Then it follows from (34) that

$$\mathrm{d}\psi_i(x, \mu) = \mathrm{d}c_i(x, \mu),$$

from which it follows that

$$\mathrm{d}\psi_i(x, \mu)v = d_i(x, v, \mu).$$

Applying Lemma A.4 to the series $d(x, \mu)$, one obtains the desired conclusion of Lemma 4.2.

Appendix B. A simple consequence of the rank theorem The next result is a simple and well-known consequence of the rank theorem; we include its proof as it seems difficult to find a precise reference. (We provide a somewhat stronger form than needed, which applies in more generality, including to nonobservable systems.)

LEMMA A.5. *Assume that $\mathcal{H} = \{h_\lambda : Z \to \mathbb{R}, \lambda \in \Lambda\}$ is a family of continuously differentiable real-valued functions on an n-dimensional differentiable manifold Z, parameterized by a set Λ. Then there exists an open dense subset $Z_0 \subseteq Z$ with the following property. For each $z_0 \in Z_0$ there exist an integer $r = r(z_0)$, an open neighborhood V of z_0 in Z_0, and parameter values $\lambda_1, \ldots, \lambda_r$, so that, for each parameter $\lambda \in \Lambda$,*

$$h_\lambda(z) = F_\lambda(h_{\lambda_1,\ldots,\lambda_r}(z)) \;\forall z \in V,$$

where $h_{\lambda_1,\ldots,\lambda_r}(z) := (h_{\lambda_1}(z), \ldots, h_{\lambda_r}(z))$ and F_λ is some C^1 function from some neighborhood$\mathcal{U}$ of $h_{\lambda_1,\ldots,\lambda_r}(V)$ to $\mathbb{R}$. Moreover, the rank of the differential of $h_{\lambda_1,\ldots,\lambda_r}(z)$ is r at all $z \in V$ (so the nonempty fibers $h^{-1}_{\lambda_1,\ldots,\lambda_r}(q)$ intersect V at submanifolds of dimension $n - r$). In particular, if it is known that $z \mapsto (h_\lambda(z), \lambda \in \Lambda)$ is one-to-one on any open subset of Z, then $r(z_0) = n$ for some $z_0 \in Z_0$.

Proof. Consider for any s and any $\lambda_1, \ldots, \lambda_s$ the rank $\rho_{\lambda_1,\ldots,\lambda_s}(z)$ of the differential of $h_{\lambda_1,\ldots,\lambda_s}$ at z, and let $\rho(z)$ be the maximum possible value of this rank over all s and $\lambda_1, \ldots, \lambda_s$. A point z is *regular* if $\rho(z)$ is constant in a neighborhood of z. The regular points form an open set by definition, and it is an easy exercise to show, by induction on $n, n-1, \dots, 1$ that the set Z_0 of such points is also dense. Now pick any z_0 in Z_0, and let $\rho(\dot{z}_0) = r$. By definition of ρ, there are parameters $\lambda_1, \dots, \lambda_r$ so that $\rho_{\lambda_1,\dots,\lambda_r}(z) = r$ for all z in some neighborhood of z_0. By the rank theorem, there are local changes of coordinates in Z so that, in some neighborhood V of z_0, $h_{\lambda_i}(z) = z_i$ for $i = 1, \dots, r$, and without loss of generality one may assume that $\rho_{\lambda_1,\dots,\lambda_r}(z) = r$ for all z in this same V. Now pick any $\lambda \in \Lambda$. Let $f = h_\lambda$. If it were the case that $\frac{\partial f}{\partial z_j}(z)$ is nonzero for some $z \in V$ and some $j > r$, then the map $h_{\lambda_1,\dots,\lambda_s,\lambda}$ would have rank $r+1$ at z, contradicting the choice of V. It follows that h_λ depends only on $z_1, \dots, z_r$ on this neighborhood, as desired. □

Remark A.6. Observe that, when dealing with analytic mappings and Z connected, the rank is constant on regular points, and one could pick the elements $\lambda_1, \dots, \lambda_r$ globally on an open dense set. Also, in general this argument shows that locally there are always n control sequences that (locally) distinguish states, even in the nonanalytic case.

REFERENCES

[1] F. ALBERTINI AND E. D. SONTAG, *Discrete-time transitivity and accessibility: Analytic systems*, SIAM J. Control Optim., 31 (1993), pp. 1599–1622.

[2] Z. BARTOSIEWICZ, *Rational systems and observation fields*, Systems Control Lett., 9 (1987), pp. 379–386.

[3] ———, *Minimal polynomial realizations*, Math. Control Signals Systems, 1 (1988), pp. 227–231.

[4] G. CONTE, G. H. MOOG, AND A. PERDON, *Un théorème sur la représentation entrée-sortie d'un système non linéaire*, C. R. Acad. Sci. Paris, Sér I. Math., 307 (1988), pp. 363–366.

[5] J.-M. CORON, *Linearized control systems and applications to smooth stabilization*, SIAM J. Control Optim., 32 (1994), pp. 358–386.

[6] P. CROUCH AND F. LAMNABHI-LAGARRIGUE, *State space realizations of nonlinear systems defined by input-output differential equations*, in Proc. 8th Internat. Conf. Analysis Optimiz. Systems, Antibes, 1988, A. Bensoussan and J. L. Lions, eds., Berlin, 1988, Springer-Verlag, New York, pp. 138–149.

[7] S. DIOP, *Elimination in control theory*, Math. Control Signals Systems, 4 (1991), pp. 17–32.

[8] ———, *Closedness of morphisms of differential algebraic sets. Application to systems theory*, Forum Math., 1 (1992), pp. 1–15.

[9] S. DIOP AND M. FLIESS, *Nonlinear observability, identifiability, and persistent trajectories*, in Proc. 30th IEEE Conference on Decision and Control, Brighton, IEEE Publications, Piscataway, NJ, 1991, pp. 714–719.

[10] ———, *On nonlinear obervability*, in Proceedings of the first European Control Conference, C. Commault, D. Normand-Cyrot, L. J. M. Dion, M. Fliess, A. Titli, A. Benveniste, and I. Landau, eds., 1991, pp. 152–157.

[11] M. FLIESS, *Réalisation locale des systèmes non linéaires, algèbres de Lie filtrées transitives et séries génératrices non commutatives*, Invent. Math., 71 (1983), pp. 521–537.

[12] M. FLIESS AND D. NORMAND-CYROT, *A group-theoretic approach to discrete-time nonlinear controllability*, in Proc. IEEE Conf. Dec. Control, IEEE Publications, Piscataway, NJ, 1981, pp. 551–557.

[13] S. T. GLAD, *Nonlinear state space and input output descriptions using differential polynomials*, in New Trends in Nonlinear Control Theory, J. Descusse, M. Fliess, A. Isidori, and M. Leborgne, eds., Springer-Verlag, Heidelberg, 1989, pp. 182–189.

[14] M. GOLUBITSKY AND V. GUILLEMIN, *Stable Mapping and Their Singularities*, Springer-Verlag, New York, 1973.

[15] O. M. GRASELLI AND A. ISIDORI, *Deterministic state reconstruction and reachability of bilinear control processes*, in Proc. Joint Automatic Control Conf., San Francisco, June 22–25 1977.

[16] J. W. GRIZZLE, *Linear algebraic framework for the analysis of discrete-time nonlinear systems*, SIAM J. Control Optim., 31 (1993), pp. 1026–1044.

[17] M. HERVÉ, *Several Complex Variables, Local Theory*, Oxford University Press, London, 1963.

[18] W. HODGE AND D. PEDOE, *Methods of Algebraic Geometry*, Vol. I, Cambridge University Press, Cambridge, 1968.

[19] A. Isidori, *Nonlinear Control Systems*, second ed., Springer-Verlag, Berlin, 1989.

[20] B. Jakubczyk, *Invertible realizations of nonlinear discrete time systems*, in Proc. Princeton Conf. Info. Sci. and Syst., 1980, pp. 235–239.

[21] B. Jakubczyk and D. Normand-Cyrot, *Orbites de pseudo groupes de diffeophismes et commandabilité des systèmes non linéaires en temps discret*, in C. R. Acad. Sci. Paris Sér. I. Math., 298, (1984), pp. 257–260.

[22] B. Jakubczyk and E. D. Sontag, *Controllability of nonlinear discrete-time systems: A Lie-algebraic approach*, SIAM J. Control Optim., 28 (1990), pp. 1–33.

[23] A. Mokkadem, *Orbites de semi-groupes de morphismes réguliers et systèmes non linéaires en temps discret*, Forum Math., 1 (1989), pp. 359–376.

[24] S. Monaco and D. Normand-Cyrot, *Invariant distributions for nonlinear discrete-time systems*, Systems Control Lett., 5 (1984), pp. 191–196.

[25] H. Nijmeijer, *Observability of autonomous discrete-time nonlinear systems: A geometric approach*, Internat. J. Control, 36 (1982), pp. 867–874.

[26] H. Nijmeijer and A. van der Schaft, *Nonlinear Dynamical Control Systems*, Springer-Verlag, New York, 1990.

[27] E. D. Sontag, *On the observability of polynomial systems,* I: *Finite-time problems*, SIAM J. Control Optim., 17 (1979), pp. 139–151.

[28] ———, *Polynomial Response Maps*, Springer-Verlag, Berlin, New York, 1979.

[29] ———, *A concept of local observability*, Systems Control Lett., 5 (1984), pp. 41–47.

[30] ———, *Mathematical Control Theory, Deterministic Finite Dimensional Systems*, Springer-Verlag, New York, 1990.

[31] ———, *Universal nonsingular controls*, Systems Control Lett., 19 (1992), pp. 221–224; Errata 20 (1993), p. 77.

[32] E. D. Sontag and Y. Wang, *I/O equations for nonlinear systems and observation spaces*, in Proc. 30th IEEE Conf. Decision and Control, Brighton, UK, Dec. 1991, IEEE Publications, Piscataway, NJ, 1991, pp. 720–725.

[33] ———, *Orders of I/O equations and uniformly universal inputs*, in Proc. 33rd IEEE Conf. Decision and Control, Orlando, Dec. 1994, IEEE Publications, Piscataway, NJ, 1994, pp. 1270–1275. (Journal version in preparation.)

[34] H. J. Sussmann, *A proof of the realization theorem for convergent generating series of finite lie rank*, submitted.

[35] ———, *Single-input observability of continuous-time systems*, Math. Systems Theory, 12 (1979), pp. 371–393.

[36] ———, *Lie brackets and local controllability: A sufficient condition for scalar-input systems*, SIAM J. Control Optim., 21 (1983), pp. 686–713.

[37] A. V. van der Schaft, *On realizations of nonlinear systems described by higher-order differential equations*, Math Systems Theory, 19 (1987), pp. 239–275.

[38] ———, *Representing a nonlinear state space system as a set of higher-order differential equations in the inputs and outputs*, Systems Control Lett., 12 (1989), pp. 151–160.

[39] Y. Wang and E. D. Sontag, *On two definitions of observation spaces*, Systems Control Lett., 13 (1989), pp. 279–289.

[40] ———, *Algebraic differential equations and rational control systems*, SIAM J. Control Optim., 30 (1992), pp. 1126–1149.

[41] ———, *Generating series and nonlinear systems: Analytic aspects, local realizability and i/o represention*s, Forum Math., 4 (1992), pp. 299–322.

[42] ———, *I/O equations in discrete and continuous time*, in Proc. 31st IEEE Conf. Decision and Control, Tucson, Dec. 1992, IEEE Publications, Piscataway, NJ, 1992, pp. 3661–3662.

[43] J. C. Willems, *From time series to linear systems, Part* I: *Finite dimensional time-invariant systems, Part* II: *Exact modelling, and Part* III: *Approximate modelling*, Automatica, 22, 22, 23 (1986, 1986, 1987), pp. 561–580, 675–694, 87–115.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 33, No. 4, pp. 1127–1150, July 1995

© 1995 Society for Industrial and Applied Mathematics
010

MATRIX PAIRS IN TWO-DIMENSIONAL SYSTEMS: AN APPROACH BASED ON TRACE SERIES AND HANKEL MATRICES*

ETTORE FORNASINI† AND MARIA ELENA VALCHER†

Abstract. Two-dimensional system dynamics depends on matrix pairs that represent the shift operators along coordinate axes. The structure of a matrix pair is analysed according to its characteristic polynomial and to the traces of suitable matrices in the algebra generated by the elements of the pair. Necessary and sufficient conditions for properties L and P are provided by resorting to Hankel matrix theory. Finite memory and separable systems, as well as two-dimensional systems whose characteristic polynomials exhibit one-dimensional structures, are finally characterized in terms of spectral properties and traces.

Key words. two-dimensional systems, finite memory systems, separable systems, L property, P property

AMS subject classifications. 15A22, 93C35, 93B15, 93B20, 93B25

1. Introduction.

"Two-dimensional systems theory" connotes a fairly large collection of problems and methods held together by a central theme: to understand better the behaviour of processes and devices whose dynamics depends on two independent variables. Most of two-dimensional systems theory is concerned with quarter-plane causal models, whose state variable description essentially depends on a pair of square matrices associated with the shift operators along the coordinate axes. There is a long stream of research concerned with the problem of characterizing the structure of matrix pairs (see [16] for an extended bibliography). In spite of its relatively long history, however, several questions are far from a definite solution and stimulate further research in linear algebra.

The purpose of this paper is to highlight how purely algebraic results on matrix pairs apply to two-dimensional system modelling. Conversely, assuming a complementary point of view, we shall show that system theoretic methodologies, connected with the realization problem, lead to a satisfactory algebraic characterization of some special matrix pairs.

The two-dimensional models to which we refer are quarter-plane causal two-dimensional systems, described by the following equations [3]:

$$\begin{aligned} x(h+1,k+1) &= A_1x(h,k+1)+A_2x(h+1,k) \\ &+ B_1u(h,k+1)+B_2u(h+1,k), \\ y(h,k) &= Cx(h,k), \end{aligned} \tag{1.1}$$

where the input, state, and output sequences $u(\cdot,\cdot)$, $x(\cdot,\cdot)$, and $y(\cdot,\cdot)$ are defined on the discrete plane $\mathbf{Z}\times\mathbf{Z}$ and take values in $\mathbf{R}^m$, $\mathbf{R}^n$, and $\mathbf{R}^p$, respectively. A_1, A_2, B_1, B_2, and C are real matrices of suitable dimensions. In general, the initial conditions are assigned by specifying the local state values $x(i,-i)$, $i \in \mathbf{Z}$.

The $n\times n$ matrix pair (A_1, A_2) fully encodes the free evolution of the system, providing at the same time valuable insights into the forced motion. From this point of view, (A_1, A_2) plays the same role as the state transition matrix A in the one-

* Received by the editors March 29, 1993; accepted for publication (in revised form) January 19, 1994.

† Department of Electronics and Computer Science, University of Padova, via Gradenigo 6/a, 35131 Padova, Italy.

dimensional discrete system

$$\begin{aligned} x(t+1) &= Ax(t)+Bu(t), \\ y(t) &= Cx(t). \end{aligned} \tag{1.2}$$

In the two-dimensional case, however, no decomposition of the state space into $\{A_1, A_2\}$-invariant subspaces can be given, allowing an effective representation of the system behaviour as a superposition of elementary modes with simple structure. As the modal analysis approach to the unforced dynamics does not extend to system (1.1), we need to resort to different tools.

The characteristic polynomial of the pair (A_1, A_2),

$$\Delta_{A_1,A_2}(z_1,z_2) := \det(I - A_1z_1 - A_2z_2), \tag{1.3}$$

is probably the most useful one. Like the characteristic polynomial of A, which in general does not capture the underlying Jordan structure, Δ_{A_1,A_2} does not identify the similarity orbit of the pair. Nevertheless, several important aspects of the two-dimensional motion completely rely on it. There is, first of all, the internal stability of system (1.1), which depends only on the variety of the zeros of Δ_{A_1,A_2}. On the other hand, the two-dimensional Cayley–Hamilton theorem implies that the free-state and output evolution of (1.1) satisfies an autoregressive equation that involves only the coefficients of Δ_{A_1,A_2}. Additional insights into the structure of two-dimensional systems come from the factorization of the characteristic polynomial. Actually, properties like finite memory and separability and interesting features of the spectrum of $\alpha A_1 + \beta A_2$ such as property L can be restated as conditions on the factors of Δ_{A_1,A_2}.

A different tool is constituted by the traces of suitable matrices in the algebra generated by A_1 and A_2 and the associated formal power series T_{A_1,A_2}. As the trace series T_{A_1,A_2} biuniquely corresponds to the characteristic polynomial Δ_{A_1,A_2}, in principle both of them provide an equivalent information on the pair (A_1, A_2). On the other hand, we shall see that some properties, originally defined as constraints on the structure of Δ_{A_1,A_2}, are better understood when the trace series point of view is undertaken. This happens, for instance, when finite memory and separable pairs are considered and, more generally, when each irreducible factor of the characteristic polynomial has a support included in some straight line of $\mathbf{Z}\times\mathbf{Z}$. Interestingly enough, resorting to trace series and to well-established realization methodologies of system theory facilitates the translation of property L of the pair (A_1, A_2) into a bound on the rank of the Hankel matrix associated with T_{A_1,A_2}.

As previously mentioned, some properties of a matrix pair and, consequently, of the associated two-dimensional systems do not reduce to conditions on the structure of the characteristic polynomial. Perhaps, the most relevant example is that, according to a celebrated result of McCoy [10], property P is equivalent to simultaneous triangularizability.

Actually, when considering only characteristic polynomials, properties P and L prove to be indistinguishable because both of them correspond to linear factorizations of Δ_{A_1,A_2}. Deeper insights into the structure of a matrix pair are offered by the traces of all matrix products $A_{i_1}A_{i_2}\dots A_{i_k}$, $k \in \mathbf{N}$, $i_j \in \{1, 2\}$, and by the associated noncommutative power series. Indeed, representation methods of recognizable and exchangeable power series [2], [14], borrowed from automata and languages theory, provide a general framework for analysing property P and, what is more important, a finite criterion for deciding whether a matrix pair is endowed with it.

This paper is organized as follows. In §2 we explore the main connections existing between characteristic polynomial and traces of a matrix pair and we present a recursive algorithm for computing the coefficients of the series T_{A_1,A_2} starting from the characteristic polynomial and vice versa. Successively a partial fraction expansion of T_{A_1,A_2}, whose terms are explicitely connected with the irreducible factors of the characteristic polynomial, is provided.

Sections 3 and 4 deal with properties L and P and their characterizations in terms of commutative and noncommutative power series, respectively. Criteria for testing both properties are provided, based on the aforementioned Hankel matrix approach.

In the last section, the previous results are applied to investigate two important classes of state models, i.e., finite memory and separable two-dimensional systems. As both classes have matrix pairs (A_1, A_2) with property L, it is natural to expect that a variety of different characterizations, typical of the L property, is made available. These are based on the factorization of Δ_{A_1,A_2}, on the structure of the trace series and on the spectrum of the linear combinations $\alpha A_1 + \beta A_2$. Here, however, we follow a somewhat different approach and analyse first matrix pairs (A_1, A_2) with the property that the support of Δ_{A_1,A_2} is a subset of a straight line. The corresponding two-dimensional systems have a free state evolution that exhibits a one-dimensional pattern and provide the building blocks for synthesizing other classes of systems, in particular finite memory and separable systems, which constitute the main concern of the section. Finally, we show how a Levitzki theorem, suitably revisited, allows for a neat characterization of finite memory and separable two-dimensional systems having property P.

2. Characteristic polynomial and traces of a matrix pair.

Given an autonomous one-dimensional system

$$x(t+1) = Ax(t), \tag{2.1}$$

the motion corresponding to any initial state $x(0)$ can be represented by the power series $(I - Az)^{-1}x(0) = \sum_{t=0}^{+\infty} A^t x(0) z^t$. Thus the knowledge of the powers of A or, equivalently, of matrix $(I - Az)^{-1}$, provides a complete information on the dynamics of (2.1).

Weaker but nevertheless significant information is given by the traces of the powers of A. Actually, as shown by the following lemma, the assignment of the traces is equivalent to that of the characteristic polynomial, which constitutes an invariant, yet not complete, relative to the similarity relation.

LEMMA 2.1. *Let A be in $\mathbf{C}^{n\times n}$, and assume* $\det(I - Az) = 1 - d_1 z - d_2 z^2 - \cdots - d_n z^n$. *Then we have*

$$\text{tr}A - d_1 = 0, \quad \text{tr}A^2 - d_1\text{tr}A - 2d_2 = 0, \ldots, \text{tr}A^n - d_1\text{tr}A^{n-1} - \cdots - nd_n = 0 \tag{2.2}$$

and, for $k > 0$,

$$\text{tr}A^{n+k} - d_1\text{tr}A^{n+k-1} - \cdots - d_n\text{tr}A^k = 0. \tag{2.3}$$

Proof. Let $\lambda_1, \lambda_2, \ldots, \lambda_n$ be the eigenvalues of A, so that $\det(zI - A) = \prod_{i=1}^n (z - \lambda_i) = z^n - d_1 z^{n-1} - \cdots - d_n$. As the symmetric polynomials $s_k = \sum_{i=1}^n \lambda_i^k$ satisfy Newton's identities [7]

$$s_k - d_1 s_{k-1} - \cdots - k\, d_k = 0, \qquad k = 1, 2, \ldots,$$

and $s_k = \mathrm{tr} A^k = \sum_{i=1}^n \lambda_i^k$, then (2.2) and (2.3) follow. □

Referring to the unforced motion of system (1.1), namely

$$x(h+1,k+1) = A_1\ x(h,k+1) + A_2\ x(h+1,k), \tag{2.4}$$

the doubly indexed sequence of local states $x(h,k)$ induced by an initial global state $\mathcal{X}_0 = \sum_{\ell=-\infty}^{+\infty} x(-\ell,\ell) z_1^{-\ell} z_2^{\ell}$ is represented by the formal power series

$$X(z_1,z_2) = \sum_{h,k} x(h,k) z_1^h z_2^k = (I - A_1 z_1 - A_2 z_2)^{-1} \mathcal{X}_0 = \sum_{i,j=0}^{\infty} (A_1{}^i \sqcup\!\sqcup{}^j A_2\ z_1^i z_2^j) \mathcal{X}_0,$$

where the matrix coefficients $A_1{}^i \sqcup\!\sqcup{}^j A_2$, $i,j \in \mathbf{N}$, of the power series expansion of $(I - A_1 z_1 - A_2 z_2)^{-1}$ are inductively defined as

$$A_1{}^i \sqcup\!\sqcup{}^0 A_2 = A_1^i, \qquad A_1{}^0 \sqcup\!\sqcup{}^j A_2 = A_2^j \tag{2.5}$$

and, when i and j are both greater than zero,

$$A_1{}^i \sqcup\!\sqcup{}^j A_2 = A_1(A_1{}^{i-1} \sqcup\!\sqcup{}^j A_2) + A_2(A_1{}^i \sqcup\!\sqcup{}^{j-1} A_2). \tag{2.6}$$

Given (2.5) and (2.6), one easily sees that $A_1{}^i \sqcup\!\sqcup{}^j A_2 = \sum_{\nu_1,\nu_2,\dots,\nu_{i+j}} A_{\nu_1} A_{\nu_2} \dots A_{\nu_{i+j}}$, where the summation is extended to all matrix products that include the factors A_1 and A_2, i and j times, respectively. The above decomposition of matrices $A_1{}^i \sqcup\!\sqcup{}^j A_2$ facilitates a better understanding of how the free-state evolution depends on the transition matrices A_1 and A_2. Actually, assuming $x(i,-i) = 0$ for $i \neq 0$, the state in (h,k) is given by

$$x(h,k) = \sum_{\nu_1,\nu_2,\dots,\nu_{i+j}} A_{\nu_1} A_{\nu_2} \dots A_{\nu_{i+j}} x(0,0), \tag{2.7}$$

and it can be interpreted as the sum of the elementary contributions along all paths connecting $(0,0)$ to (h,k) in the two-dimensional grid (Fig. 1).

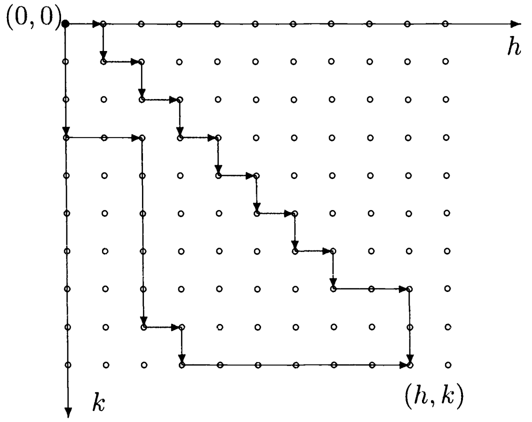


FIG. 1

The analogy between the roles played by the matrix family $\{A_1{}^i \sqcup\!\sqcup{}^j A_2\}$ and the powers of A can be further highlighted by extending both the Cayley–Hamilton theorem and Lemma 2.1 to the two-dimensional case.

PROPOSITION 2.2 (Two-dimensional Cayley–Hamilton theorem [4]). *Let*

$$\Delta_{A_1,A_2}(z_1,z_2) = 1 - \sum_{1\le i+j\le n} d_{ij} z_1^i z_2^j \tag{2.8}$$

be the characteristic polynomial of the $n\times n$ matrix pair (A_1,A_2). Then, for all pairs (h,k), with $h+k\ge n$:

$$\text{i) } A_1{}^h\shuffle^k A_2 = \sum_{1\le i+j\le n} d_{ij}\ A_1{}^{h-i}\shuffle^{k-j} A_2 \tag{2.9}$$

(where $A_1{}^i\shuffle^j A_2$; is assumed to be zero whenever i or j is negative);

$$\text{ii) } A_1{}^h\shuffle^k A_2 \in \operatorname{span}\{A_1{}^i\shuffle^j A_2 : i\le h,\ j\le k, i+j<n\}. \tag{2.10}$$

iii)$\operatorname{span}\{A_1{}^h\shuffle^k A_2 : h,k\in \mathbf{N}\} = \operatorname{span}\{A_1{}^h\shuffle^k A_2 : h,k<n\}$.

Proof. i) Since the $(n-1)$th-order minors of $I - A_1z_1 - A_2z_2$ have degrees less than n, we have

$$\operatorname{adj}(I - A_1z_1 - A_2z_2) = \sum_{0\le r+s<n} M_{rs} z_1^r z_2^s. \tag{2.11}$$

Replace (2.8) and (2.11) into $(I-A_1z_1-A_2z_2)^{-1}\Delta_{A_1,A_2}(z_1,z_2) = \operatorname{adj}(I-A_1z_1-A_2z_2)$, and use the power series expansion of $(I-A_1z_1-A_2z_2)^{-1}$, obtaining

$$\left(\sum_{h,k} A_1{}^h\shuffle^k A_2\, z_1^h z_2^k\right)\left(1 - \sum_{r+s\le n} d_{rs} z_1^r z_2^s\right) = \sum_{i+j<n} M_{ij} z_1^i z_2^j. \tag{2.12}$$

Thus, (2.9) simply states that the Cauchy product on the left-hand side of (2.12) does not include nonzero monomials with degree greater than $n-1$.

ii) If $h+k=n$, the statement is trivial. If $h+k=\nu+1>n$, assume by induction that (2.10) holds for all $(h,k)\in\mathbf{N}\times\mathbf{N}$, with $n\le h+k\le\nu$. So, for $r+s>0$, all matrices $A_1{}^{h-r}\shuffle^{k-s}A_2$ linearly depend on $\{A_1{}^i\shuffle^j A_2 : i\le h, j\le k, i+j<n\}$ and the same holds true for $A_1{}^h\shuffle^k A_2$, because of (2.9).

iii) It follows directly from ii) . □

In order to extend Lemma 2.1 to the two-dimensional case, rewrite the characteristic polynomial of the pair (A_1,A_2) as

$$\Delta_{A_1,A_2}(z_1,z_2) = 1 - \sum_{h=1}^{n}\left(\sum_{i+j=h} d_{ij} z_1^i z_2^j\right) = 1 - \sum_{h=1}^{n}\delta_h(z_1,z_2) \tag{2.13}$$

and introduce the "trace series"

$$T_{A_1,A_2}(z_1,z_2) := \sum_{h=1}^{\infty}\left(\sum_{i+j=h}\operatorname{tr}(A_1{}^i\shuffle^j A_2) z_1^i z_2^j\right) = \sum_{h=1}^{\infty}\tau_h(z_1,z_2), \tag{2.14}$$

where $\delta_h(z_1,z_2)$ and $\tau_h(z_1,z_2)$ are homogeneous forms of degree h.

PROPOSITION 2.3. *Let (A_1,A_2) be an $n\times n$ matrix pair with entries in $\mathbf{C}$, and $\Delta_{A_1,A_2}(z_1,z_2)$ and $T_{A_1,A_2}(z_1,z_2)$ its characteristic polynomial and trace series, respectively. Then:*

i) *the homogeneous components* $\delta_h(z_1,z_2)$ *and* $\tau_h(z_1,z_2)$ *satisfy*

$$\begin{aligned}&\tau_1(z_1,z_2)-\delta_1(z_1,z_2)=0,\\ &\tau_2(z_1,z_2)-\delta_1(z_1,z_2)\tau_1(z_1,z_2)-2\delta_2(z_1,z_2)=0,\\ &\dots\\ &\tau_n(z_1,z_2)-\delta_1(z_1,z_2)\tau_{n-1}(z_1,z_2)-\cdots-n\delta_n(z_1,z_2)=0\end{aligned}\tag{2.15}$$

and, for all $k>0$,

$$\tau_{n+k}(z_1,z_2)-\sum_{i=1}^{n}\tau_{n+k-i}(z_1,z_2)\delta_i(z_1,z_2)=0;\tag{2.16}$$

ii) *the traces of* $A_1{}^i\sqcup\!\sqcup^j A_2$ *and the coefficients* d_{ij} *of* $\Delta_{A_1,A_2}(z_1,z_2)$ *satisfy*

$$\operatorname{tr}(A_1{}^i\sqcup\!\sqcup^j A_2)=\sum_{0<r+s<i+j}d_{rs}\operatorname{tr}(A_1{}^{i-r}\sqcup\!\sqcup^{j-s}A_2)+(i+j)d_{ij},\tag{2.17}$$

where $d_{rs}=0$ *for* $r+s>n$ *and* $A_1{}^r\sqcup\!\sqcup^s A_2$ *is the zero matrix whenever* r *and/or* s *is negative.*

Proof. i) Let $\alpha,\beta\in\mathbf{C}$, and substitute in (2.8) z_1 and z_2 for αz and βz:

$$\det[I-(\alpha A_1+\beta A_2)z]=1-\sum_{h=1}^{n}\delta_h(\alpha,\beta)z^h.$$

Taking the traces on both sides of $(\alpha A_1+\beta A_2)^h=\sum_{i=0}^h\alpha^i\beta^{h-i}A_1{}^i\sqcup\!\sqcup^{h-i}A_2$, one gets

$$\operatorname{tr}(\alpha A_1+\beta A_2)^h=\sum_{i=0}^{h}\alpha^i\beta^{h-i}\operatorname{tr}(A_1{}^i\sqcup\!\sqcup^{h-i}A_2).\tag{2.18}$$

As (2.18) holds for all α,β in $\mathbf{C}$, it is immediate to recognize in $\operatorname{tr}(\alpha A_1+\beta A_2)^h$ the homogeneous forms $\tau_h(\alpha,\beta)$ of (2.14). Thus we can apply Lemma 2.1

$$\begin{aligned}&\tau_1(\alpha,\beta)-\delta_1(\alpha,\beta)=0,\\ &\tau_2(\alpha,\beta)-\delta_1(\alpha,\beta)\tau_1(\alpha,\beta)-2\delta_2(\alpha,\beta)=0,\\ &\dots\\ &\tau_n(\alpha,\beta)-\delta_1(\alpha,\beta)\tau_{n-1}(\alpha,\beta)-\cdots-n\delta_n(\alpha,\beta)=0\end{aligned}\tag{2.19}$$

and, for all $k>0$,

$$\tau_{n+k}(\alpha,\beta)-\sum_{i=1}^{n}\tau_{n+k-i}(\alpha,\beta)\delta_i(\alpha,\beta)=0.\tag{2.20}$$

As α and β are arbitrary, (2.15) and (2.16) follow.

ii) Substitute the expressions of $\delta_h(z_1,z_2)$ and $\tau_h(z_1,z_2)$ given in (2.13) and (2.14) into (2.15) and (2.16), and equate to zero the coefficients of all monomials on the left-hand side. □

Equation (2.15) has some simple but useful consequences. First, it provides an algorithm for recursively computing the traces of $A_1{}^i\sqcup\!\sqcup^j A_2$ from the coefficients of the characteristic polynomial. On the other hand, once the traces are given, also the converse, i.e., the computation of the coefficients of Δ, is made possible. Actually, if an upper bound $\bar{n}$ on the degree of Δ is known, assigning $\operatorname{tr}(A_1{}^{i}\sqcup\!\sqcup^{j} A_2)$ for $i+j \leq \bar{n}$ allows the recovery of both Δ and the traces of $A_1{}^{i}\sqcup\!\sqcup^{j} A_2$ for $i+j > \bar{n}$.

Consider the set of all matrix pairs $\mathcal{M} = \{(A_1, A_2) : A_1, A_2 \in \mathbf{C}^{n\times n}, n \in \mathbf{N}\}$, and introduce in $\mathcal{M}$ the equivalence relation

$$(A_1, A_2) \sim (\hat{A}_1, \hat{A}_2) \quad \Leftrightarrow \quad \Delta_{A_1 A_2}(z_1, z_2) = \Delta_{\hat{A}_1 \hat{A}_2}(z_1, z_2).$$

Corollary 2.4 below exhibits different sets of complete invariants for relation $\sim$. Actually, two matrix pairs have the same characteristic polynomial if and only if (the coefficients of) the corresponding trace series coincide. The equivalence relation on $\mathcal{M}$ can also be described in terms of spectra and traces of the linear combinations of the elements of each matrix pair.

COROLLARY 2.4. *Let A_1, A_2 be in $\mathbf{C}^{n\times n}$ and $\hat{A}_1, \hat{A}_2$ in $\mathbf{C}^{\hat{n}\times\hat{n}}$. The following statements are equivalent*:

i) $\Delta_{A_1,A_2}(z_1, z_2) = \Delta_{\hat{A}_1,\hat{A}_2}(z_1, z_2)$;

ii) *for all $\alpha, \beta \in \mathbf{C}$, $\Lambda_0(\alpha A_1 + \beta A_2) = \Lambda_0(\alpha\hat{A}_1 + \beta\hat{A}_2)$, where $\Lambda_0(M)$ denotes the set of nonzero eigenvalues of the matrix M, each of them counted according to the corresponding algebraic multiplicity*;

iii) *for all $\alpha, \beta \in \mathbf{C}$ and $k \in \mathbf{N}_+$*, $\operatorname{tr}(\alpha A_1 + \beta A_2)^k = \operatorname{tr}(\alpha\hat{A}_1 + \beta\hat{A}_2)^k$;

iv) *for all $(i, j) \neq (0, 0)$*, $\operatorname{tr}(A_1{}^{i}\sqcup\!\sqcup^{j} A_2) = \operatorname{tr}(\hat{A}_1{}^{i}\sqcup\!\sqcup^{j} \hat{A}_2)$.

Proof. i) $\Leftrightarrow$ ii) As both i) and ii) are equivalent to

$$\det[I - (\alpha A_1 + \beta A_2)z] = \det[I - (\alpha\hat{A}_1 + \beta\hat{A}_2)z] \quad \forall \alpha, \beta \in \mathbf{C},$$

they are equivalent to each other, too.

i)$\Leftrightarrow$ iii) $\Leftrightarrow$ iv) By Proposition 2.3, (A_1, A_2) and $(\hat{A}_1, \hat{A}_2)$ have the same characteristic polynomial if and only if the corresponding homogeneous forms $\operatorname{tr}(\alpha A_1 + \beta A_2)^k$ and $\operatorname{tr}(\alpha\hat{A}_1 + \beta\hat{A}_2)^k$, $k = 1, 2\ldots$, coincide. This, in turn, is equivalent to assuming $\operatorname{tr}(A_1{}^{i}\sqcup\!\sqcup^{j} A_2) = \operatorname{tr}(\hat{A}_1{}^{i}\sqcup\!\sqcup^{j} \hat{A}_2)$, for all $(i, j) \neq (0, 0)$. □

It is easy to realize that T_{A_1,A_2} has to be a rational power series, since its coefficients satisfy the recursive equations (2.17). In what remains of this section we aim to make explicit its rational structure and identify its connections with the characteristic polynomial.

PROPOSITION 2.5. *Let $\Delta(z_1, z_2) = 1 - \sum_{h=1}^{n} \delta_h(z_1, z_2)$ be the characteristic polynomial of the matrix pair (A_1, A_2). The corresponding trace series T_{A_1,A_2} can be expressed as*

$$T_{A_1,A_2}(z_1, z_2) = \frac{\delta_1(z_1, z_2) + 2\delta_2(z_1, z_2) + \cdots + n\delta_n(z_1, z_2)}{\Delta(z_1, z_2)}. \tag{2.21}$$

Proof. Consider the linear system defined on $\mathbf{C}[\alpha, \beta]$, the ring of the polynomials in the indeterminates α and β with coefficients in $\mathbf{C}$:

$$\begin{aligned} x_{i+1} &= Fx_i + gu_i, \\ y_i &= Hx_i, \end{aligned}$$

with

$$F = \begin{bmatrix} 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ & & & \cdots & \\ & & & \cdots & 1 \\ \delta_n(\alpha,\beta) & \delta_{n-1}(\alpha,\beta) & \delta_{n-2}(\alpha,\beta) & \cdots & \delta_1(\alpha,\beta) \end{bmatrix}, \qquad g = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \\ 1 \end{bmatrix},$$

$$H = [\ 0 \quad 0 \quad 0 \quad \cdots \quad 1\ ].$$

By assuming $x_0 = 0$ and

$$u_i = \begin{cases} (i+1)\delta_{i+1}(\alpha,\beta), & \mathrm{i} = 0, 1, \dots, \\ 0 & \text{otherwise,} \end{cases}$$

it is a matter of direct computation to check that the corresponding output sequence is given by $y_i = \tau_i(\alpha,\beta)$, $i = 1, 2, \dots$.

As the system transfer function is $H(I - zF)^{-1}gz = z/(1 - \sum_{h=1}^{n} \delta_h(\alpha,\beta)z^h)$, the input $U(z) = u_0 + u_1 z + \cdots + u_{n-1}z^{n-1}$ produces the output

$$Y(z) = \sum_{i=0}^{\infty} \tau_i(\alpha,\beta)z^i = \frac{\sum_{h=1}^{n} h\delta_h(\alpha,\beta)z^h}{1 - \sum_{h=1}^{n} \delta_h(\alpha,\beta)z^h}. \tag{2.22}$$

So, letting $z_1 = \alpha z$ and $z_2 = \beta z$, one gets (2.21). □

Representation (2.21) of T_{A_1,A_2} is not necessarily irreducible, but its special structure makes it quite easy to obtain an irreducible one. To this purpose, consider the injective homomorphism

$$\phi : \mathbf{C}[\alpha,\beta] \to \mathbf{C}[\alpha,\beta,z] : \sum_{i=0}^{n} \delta_i(\alpha,\beta) \mapsto \sum_{i=0}^{n} \delta_i(\alpha,\beta)z^i,$$

where, as usual, $\delta_i(\alpha,\beta)$ denotes a homogeneous polynomial of degree i, and introduce the derivation map

$$D_z : \mathbf{C}[\alpha,\beta,z] \to \mathbf{C}[\alpha,\beta,z]\ : \sum_{i=0}^{m} p_i(\alpha,\beta)z^i \mapsto \sum_{i=0}^{m} ip_i(\alpha,\beta)z^i. \tag{2.23}$$

Clearly (2.22) can be rewritten as

$$Y(z) = \frac{D_z(\phi(\Delta(\alpha,\beta)))}{\phi(\Delta(\alpha,\beta))}. \tag{2.24}$$

By assuming that Δ factorizes as $\Delta(z_1,z_2) = \prod_{i=1}^{t} \Delta_i(z_1,z_2)^{\nu_i}$, with Δ_i irreducible distinct factors, $\Delta_i(0,0) = 1$, $i = 1, 2, \dots, t$, one easily gets

$$Y(z) = \sum_{i=1}^{t} \nu_i \frac{D_z(\phi(\Delta_i(\alpha,\beta)))}{\phi(\Delta_i(\alpha,\beta))}. \tag{2.25}$$

Thus, letting $z_1 = \alpha z$ and $z_2 = \beta z$, we have proved the following proposition.

PROPOSITION 2.6. *Let $\Delta(z_1,z_2) = \prod_{i=1}^{t} \Delta_i(z_1,z_2)^{\nu_i}$ be a factorization of Δ, with $\Delta_i(z_1,z_2) = 1 - \sum_{j=1}^{r_i} \delta_j^{(i)}(z_1,z_2)$ irreducible distinct polynomials, $i = 1, 2, \dots, t$. For every matrix pair (A_1, A_2) such that $\Delta_{A_1,A_2}(z_1,z_2) = \Delta(z_1,z_2)$, the corresponding trace series is given by*

$$T_{A_1,A_2}(z_1,z_2) = \sum_{i=1}^{t} \nu_i \frac{\sum_{j=1}^{r_i} j\ \delta_j^{(i)}(z_1,z_2)}{1 - \sum_{j=1}^{r_i} \delta_j^{(i)}(z_1,z_2)}. \tag{2.26}$$

Equation (2.26) expresses the trace series $T_{A_1,A_2}(z_1,z_2)$ as a partial fraction expansion, whose ith term is the trace series of the irreducible factor $\Delta_i(z_1,z_2)$, weighted with the corresponding multiplicity ν_i. Thus the denominator of every irreducible rational function that represents a trace series factorizes into distinct irreducible factors. On the other hand, once an irreducible rational function $T(z_1,z_2)$ has been given, (2.26) suggests a quick way to check whether $T(z_1,z_2)$ can be expanded into a trace series.

3. Pairs of matrices with property L.

3. Pairs of matrices with property L. In the next two sections we focus specifically on matrix pairs endowed with property L and property P.

Pairs with property L occur quite frequently in the applications: indeed, the important classes of finite memory and separable two-dimensional systems that we are going to discuss in §5 are described by pairs with property L. A pair of $n \times n$ matrices, (A_1, A_2), with entries in $\mathbf{C}$, is said to have property L if the eigenvalues of A_1 and A_2 can be ordered into two n-tuples

$$\Lambda(A_1) = (\lambda_1, \lambda_2, \dots, \lambda_n) \quad \text{and} \quad \Lambda(A_2) = (\mu_1, \mu_2, \dots, \mu_n) \tag{3.1}$$

such that, for all α, β in $\mathbf{C}$, the spectrum of $\Lambda(\alpha A_1 + \beta A_2)$ is given by

$$\Lambda(\alpha A_1 + \beta A_2) = (\alpha\lambda_1 + \beta\mu_1, \dots, \alpha\lambda_n + \beta\mu_n). \tag{3.2}$$

It is not difficult to show that property L corresponds to the possibility of factorizing the characteristic polynomial into linear terms [11], [12]. Thus each term of the partial fraction expansion of T_{A_1,A_2} has the very special structure $(\lambda z_1 + \mu z_2)/(1 - \lambda z_1 - \mu z_2)$, which has far-reachıng consequences on the possibility of characterizing property L using the Hankel matrix theory.

PROPOSITION 3.1. *Let A_1, A_2 be in $\mathbf{C}^{n\times n}$, and consider the orderings of their spectra given in* (3.1). *The following statements are equivalent*:

L) *(A_1, A_2) has property* L *(w.r.t. the orderings* (3.1));

L$_1$) $\Delta_{A_1,A_2}(z_1,z_2) = \prod_{i=1}^n (1 - \lambda_i z_1 - \mu_i z_2)$;

L$_2$) *for all $\alpha, \beta \in \mathbf{C}$ and $k \in \mathbf{N}$,* $\operatorname{tr}(\alpha A_1 + \beta A_2)^k = \sum_{i=1}^n (\alpha\lambda_i + \beta\mu_i)^k$;

L$_3$) *for every $(h,k) \in \mathbf{N} \times \mathbf{N}$,* $\operatorname{tr}(A_1{}^h \sqcup\!\sqcup {}^k A_2) = \binom{h+k}{h} \sum_{i=1}^n \lambda_i^h \mu_i^k$;

L$_4$) $T_{A_1,A_2} = \sum_{h+k>0} \operatorname{tr}(A_1{}^h \sqcup\!\sqcup {}^k A_2) z_1^h z_2^k = \sum_{i=1}^n (\lambda_i z_1 + \mu_i z_2)/(1 - \lambda_i z_1 - \mu_i z_2)$.

Proof. Clearly matrices $\hat{A}_1 = \operatorname{diag}\{\lambda_1, \lambda_2, \dots, \lambda_n\}$ and $\hat{A}_2 = \operatorname{diag}\{\mu_1, \mu_2, \dots, \mu_n\}$ fulfill all conditions *L)–L$_4$)* of the proposition. Any other pair (A_1, A_2), of the same dimension, with property L w.r.t. the orderings (3.1), satisfies

$$\Lambda(\alpha A_1 + \beta A_2) = \Lambda(\alpha \hat{A}_1 + \beta \hat{A}_2), \tag{3.3}$$

which corresponds to ii) of Corollary 2.4. Therefore all equivalent statements in Proposition 2.3 hold true; in particular, property L of (A_1, A_2) is equivalent to any one of the following:

L$_1$) $\Delta_{A_1,A_2}(z_1,z_2) = \Delta_{\hat{A}_1,\hat{A}_2}(z_1,z_2) = \prod_{i=1}^n (1 - \lambda_i z_1 - \mu_i z_2)$;

L$_2$) $\operatorname{tr}(\alpha A_1 + \beta A_2)^k = \operatorname{tr}(\alpha \hat{A}_1 + \beta \hat{A}_2)^k = \sum_{i=1}^n (\alpha\lambda_i + \beta\mu_i)^k$;

L$_3$) $\operatorname{tr}(A_1{}^h \sqcup\!\sqcup {}^k A_2) = \operatorname{tr}(\hat{A}_1{}^h \sqcup\!\sqcup {}^k \hat{A}_2) = \binom{h+k}{k} \sum_{i+1}^n \lambda_i^h \mu_i^k$;

L$_4$) $\sum_{(h,k)\neq(0,0)} \operatorname{tr}(A_1{}^h \sqcup\!\sqcup {}^k A_2) z_1^h z_2^k$

$= \sum_{(h,k)\neq(0,0)} \operatorname{tr}(\hat{A}_1{}^h \sqcup\!\sqcup {}^k \hat{A}_2) z_1^h z_2^k = \sum_{i=1}^n (\lambda_i z_1 + \mu_i z_2)/(1 - \lambda_i z_1 - \mu_i z_2)$. □

Conditions *L)–L$_4$)* do not provide direct methods to check, in a finite number of steps, whether a given pair (A_1, A_2) is endowed with property L. To reach this goal, we shall analyse the rank of suitable matrices associated with the power series

$$R_{A_1,A_2}(z_1,z_2) := \sum_{i,j=0}^{\infty} \operatorname{tr}(A_1{}^{i}\sqcup\!\sqcup{}^{j}A_2)\binom{i+j}{i}^{-1} z_1^i z_2^j. \tag{3.4}$$

Let $\mathbf{C}[[z_1,z_2]]$ be the ring of formal power series in the commuting variables z_1, z_2, and denote by

$$s := \sum_{h,k} \langle s, z_1^h z_2^k\rangle z_1^h z_2^k$$

a generic element of the ring. We associate with s the infinite Hankel matrix [2]

$$\mathcal{H}(s) := \begin{bmatrix} \langle s,1\rangle & \langle s,z_1 & \langle s,z_2\rangle & \langle s,z_1^2\rangle & \langle s,z_1z_2\rangle & \langle s,z_2^2\rangle & \dots \\ \langle s,z_1\rangle & \langle s,z_1^2\rangle & \langle s,z_1z_2\rangle & \langle s,z_1^3\rangle & \langle s,z_1^2z_2\rangle & \dots & \dots \\ \langle s,z_2\rangle & \langle s,z_1z_2\rangle & \langle s,z_2^2\rangle & \dots & \dots & \dots & \dots \\ \langle s,z_1^2\rangle & \dots & \dots & \dots & \dots & \dots & \dots \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots \end{bmatrix},$$

whose rows and columns take indices in the multiplicative monoid of the (commutative) terms $\mathcal{T} := \{z_1^i z_2^j : i,j \in \mathbf{N}\}$.

For all $M', M'' \in \mathbf{N}$, we shall denote by $\mathcal{H}_{M' \times M''}(s)$ the submatrix, appearing in the upper left corner of $\mathcal{H}(s)$, whose rows (columns) are indexed by the terms of homogeneous degree not greater than M' (M'').

When rank $\mathcal{H}(s)$ is $\nu < \infty$, we can choose ν rows and ν columns, indexed by the terms $r_1, r_2, \dots, r_\nu$ and $c_1, c_2, \dots, c_\nu$, respectively, so that the submatrix

$$N_0 := [\langle s, r_i c_j\rangle] \tag{3.5}$$

is nonsingular. Thus, for all terms $c \in \mathcal{T}$, the ν-tuple $[\langle s, r_1 c\rangle \dots \langle s, r_\nu c\rangle]^{\mathrm{T}}$ belongs to the range space of N_0, i.e., there exists a (unique) vector $\mathbf{x}(c) \in \mathbf{C}^\nu$ such that

$$\begin{bmatrix} \langle s, r_1 c\rangle \\ \vdots \\ \langle s, r_\nu c\rangle \end{bmatrix} = N_0\ \mathbf{x}(c).$$

Moreover, the rank assumption on $\mathcal{H}(s)$ implies

$$\langle s, rc\rangle = \sum_{j=1}^{\nu} x_j(c)\langle s, rc_j\rangle \quad \forall r \in \mathcal{T}. \tag{3.6}$$

We therefore have, for all $r, c \in \mathcal{T}$,

$$\langle s, rc\rangle = \Big[\langle s, rc_1\rangle, \dots, \langle s, rc_\nu\rangle\Big] N_0^{-1} \begin{bmatrix} \langle s, r_1 c\rangle \\ \vdots \\ \langle s, r_\nu c\rangle \end{bmatrix}. \tag{3.7}$$

We are now in a position to state the following proposition.

PROPOSITION 3.2. *Let A_1, A_2 be in $\mathbf{C}^{n\times n}$. (A_1, A_2) has property* L *if and only if*

$$\bar{n} := \operatorname{rank} \mathcal{H}_{(n-1)\times(n-1)}(R_{A_1,A_2}) = \operatorname{rank} \mathcal{H}_{n\times n}(R_{A_1,A_2}) \le n. \tag{3.8}$$

Proof. For the sake of brevity, within the proof we shall drop in R_{A_1,A_2} subscripts A_1 and A_2. Assume first that (A_1, A_2) has property L. Then, by Proposition 3.1,

$$\operatorname{tr}(A_1{}^i \sqcup\!\sqcup^j A_2) = \binom{i+j}{i} \sum_{h=1}^{n} \lambda_h^i \mu_h^j$$

and, therefore,

$$R = \sum_{i,j=0}^{\infty} \left(\sum_{h=1}^{n} \lambda_h^i \mu_h^j \right) z_1^i z_2^j.$$

Since the Hankel matrix $\mathcal{H}(R)$ factorizes as

$$(3.9) \quad \mathcal{H}(R) = \begin{bmatrix} 1 & 1 & \dots & 1 \\ \lambda_1 & \lambda_2 & \dots & \lambda_n \\ \mu_1 & \mu_2 & \dots & \mu_n \\ \lambda_1^2 & \lambda_2^2 & \dots & \lambda_n^2 \\ \lambda_1\mu_1 & \lambda_2\mu_2 & \dots & \lambda_n\mu_n \\ \mu_1^2 & \mu_2^2 & \dots & \mu_n^2 \\ & & \dots & \end{bmatrix} \begin{bmatrix} 1 & \lambda_1 & \mu_1 & \lambda_1^2 & \lambda_1\mu_1 & \mu_1^2 & \dots \\ 1 & \lambda_2 & \mu_2 & \lambda_2^2 & \lambda_2\mu_2 & \mu_2^2 & \dots \\ & & & \dots & & & \\ 1 & \lambda_n & \mu_n & \lambda_n^2 & \lambda_n\mu_n & \mu_n^2 & \dots \end{bmatrix},$$

clearly rank $\mathcal{H}(R) \leq n$.

Finally, apply the two-dimensional Cayley–Hamilton theorem to the pair of matrices $Q_1 := \operatorname{diag}\{\lambda_1, \lambda_2, \dots, \lambda_n\}$ and $Q_2 := \operatorname{diag}\{\mu_1, \mu_2, \dots, \mu_n\}$. Since the matrices

$$Q_1{}^h \sqcup\!\sqcup^k Q_2 = \binom{h+k}{k} \operatorname{diag}\{\lambda_1^h\mu_1^k, \lambda_2^h\mu_2^k, \dots, \lambda_n^h\mu_n^k\}, \qquad h+k \geq n,$$

are linear combinations of

$$Q_1{}^i \sqcup\!\sqcup^j Q_2 = \binom{i+j}{i} \operatorname{diag}\{\lambda_1^i\mu_1^j, \lambda_2^i\mu_2^j, \dots, \lambda_n^i\mu_n^j\}, \quad i+j < n,$$

all rows (columns) in the first (second) factor of (3.9) that include homogeneous terms of degree greater than or equal to n, linearly depend on the previous ones. Hence rank $\mathcal{H}_{(n-1)\times(n-1)}(R) =$ rank $\mathcal{H}(R)$ and (3.8) holds.

To prove the converse, assume that (3.8) holds and select in $\mathcal{H}_{(n-1)\times(n-1)}(R)$, $\bar{n}$ rows and $\bar{n}$ columns indexed by terms $r_1, r_2, \dots, r_{\bar{n}}$ and $c_1, c_2, \dots, c_{\bar{n}}$, respectively, such that the $\bar{n} \times \bar{n}$ submatrix $N_0 := [\langle R, r_i c_j \rangle]$ is nonsingular. The following matrices

$$(3.10) \qquad M_1 := N_0^{-1}[\langle R, r_i z_1 c_j \rangle], \qquad M_2 := N_0^{-1}[\langle R, r_i z_2 c_j \rangle]$$

commute. Indeed, $r_i z_1 z_2 c_j$ are terms of degree not greater than $2n$, and consequently assumption (3.8) allows us to resort to (3.7), which gives

$$\langle R, r_i z_1 z_2 c_j \rangle = \Big[\langle R, r_i z_1 c_1 \rangle, \dots, \langle R, r_i z_1 c_{\bar{n}} \rangle \Big] N_0^{-1} \begin{bmatrix} \langle R, r_1 z_2 c_j \rangle \\ \vdots \\ \langle R, r_{\bar{n}} z_2 c_j \rangle \end{bmatrix} \quad \forall\, i,j \in \{1, 2, \dots, \bar{n}\}.$$

This implies that

$$M_1 M_2 - M_2 M_1 = N_0^{-1} \Big[[\langle R, r_i z_1 c_j \rangle] N_0^{-1} [\langle R, r_i z_2 c_j \rangle] - [\langle R, r_i z_2 c_j \rangle] N_0^{-1} [\langle R, r_i z_1 c_j \rangle] \Big] = 0.$$

Next introduce the matrices

$$(3.11)\qquad H := \Big[\langle R, c_1\rangle\ \langle R, c_2\rangle \dots \langle R, c_{\bar n}\rangle\Big] \quad\text{and}\quad G := N_0^{-1}\begin{bmatrix}\langle R, r_1\rangle\\ \vdots \\ \langle R, r_{\bar n}\rangle\end{bmatrix},$$

and consider the commutative power series

$$(3.12)\qquad S := H(I - M_1 z_1)^{-1}(I - M_2 z_2)^{-1}G = \sum_{i,j=0}^{\infty} H M_1^i M_2^j G\, z_1^i z_2^j.$$

By resorting again to (3.7), it is easy to check that $\langle R, z_1^i z_2^j\rangle = \langle S, z_1^i z_2^j\rangle$, for $i+j \le 2n$, and therefore $\mathcal{H}_{n\times n}(R) = \mathcal{H}_{n\times n}(S)$. As M_1 and M_2 commute, they are endowed with property L, and the power series

$$\bar S := \sum_{i,j=0}^{\infty} H(M_1{}^i\sqcup\!\sqcup^j M_2)G\, z_1^i z_2^j = H(I - M_1 z_1 - M_2 z_2)^{-1}G$$

can be represented as a rational function of the form

$$(3.13)\qquad \bar S := \frac{q(z_1, z_2)}{\prod_{i=1}^{\bar n}(1 - \lambda_i z_1 - \mu_i z_2)}, \qquad \deg q < \bar n \le n.$$

Note that $\bar S$ satisfies $\langle \bar S, 1\rangle = \langle S, 1\rangle = \langle R, 1\rangle = n$, and, for $0 < i + j \le 2n$

$$(3.14)\qquad \langle \bar S, z_1^i z_2^j\rangle = \binom{i+j}{i}\langle S, z_1^i z_2^j\rangle = \binom{i+j}{i}\langle R, z_1^i z_2^j\rangle = \langle T_{A_1,A_2}, z_1^i z_2^j\rangle.$$

On the other hand, being the trace series of an $n \times n$ matrix pair, T_{A_1,A_2} can be expressed as in (2.15) and, consequently we have

$$T_{A_1,A_2} + n = \frac{p(z_1, z_2)}{\Delta_{A_1,A_2}(z_1, z_2)}, \quad \deg p \le n.$$

Therefore, in the rational function $\bar S - T_{A_1,A_2} - n$ the denominator has degree not greater than $2n$ and nonzero constant term, while the numerator has degree not greater than $2n$. As all the coefficients of the power series expansion of $\bar S - T_{A_1,A_2} - n$, namely $\langle \bar S - T_{A_1,A_2} - n, z_1^i z_2^j\rangle$, are zero for $i + j \le 2n$, then $\bar S = T_{A_1,A_2} + n$.

It is clear now from (3.13) that the denominator of an irreducible representation of T_{A_1,A_2} factorizes into linear factors. Therefore, by Proposition 3.1, (A_1, A_2) has property L. □

4. Pairs of matrices with property P. Given the alphabet $\Xi = \{\xi_1, \xi_2\}$, the free monoid Ξ^* with base Ξ is the set of all words

$$w = \xi_{i_1}\xi_{i_2}\dots\xi_{i_m}, \qquad m \in \mathbf{N},\ \xi_{i_h} \in \Xi.$$

The integer m is called the length of the word w and is denoted by $|w|$, while $|w|_i$ represents the number of occurrences of ξ_i in w, $i = 1, 2$. If $v = \xi_{j_1}\xi_{j_2}\dots\xi_{j_p}$ is another element of Ξ^*, the product is defined by concatenation:

$$wv = \xi_{i_1}\xi_{i_2}\dots\xi_{i_m}\xi_{j_1}\xi_{j_2}\dots\xi_{j_p}.$$

This produces a monoid with $1 = \emptyset$, the empty word, as unit element. Clearly, $|wv| = |v| + |w|$ and $|1| = 0$.

$\mathbf{C}\langle\xi_1,\xi_2\rangle$ and $\mathbf{C}\langle\langle\xi_1,\xi_2\rangle\rangle$ are the algebras of polynomials and formal power series, respectively, in the noncommuting indeterminates ξ_1 and ξ_2. For each pair of matrices A_1, A_2 in $\mathbf{C}^{n\times n}$, the map ψ defined on $\{1,\xi_1,\xi_2\}$ by the assignments $\psi(1) = I_n$ and $\psi(\xi_i) = A_i$, $i = 1,2$, uniquely extends to an algebra morphism of $\mathbf{C}\langle\xi_1,\xi_2\rangle$ into $\mathbf{C}^{n\times n}$. The ψ-image of a polynomial $\mathcal{P}(\xi_1,\xi_2) \in \mathbf{C}\langle\xi_1,\xi_2\rangle$ is denoted by $\mathcal{P}(A_1,A_2)$.

A pair of $n \times n$ matrices (A_1,A_2) with elements in $\mathbf{C}$ is said to have property P if the eigenvalues of A_1 and A_2 can be ordered into two n-tuples

$$(4.1) \qquad \Lambda(A_1) = (\lambda_1,\lambda_2,\dots,\lambda_n), \qquad \Lambda(A_2) = (\mu_1,\mu_2,\dots,\mu_n),$$

such that, for every polynomial $\mathcal{P}(\xi_1,\xi_2) \in \mathbf{C}\langle\xi_1,\xi_2\rangle$,

$$(4.2) \qquad \Lambda(\mathcal{P}(A_1,A_2)) = (\mathcal{P}(\lambda_1,\mu_1),\mathcal{P}(\lambda_2,\mu_2),\dots,\mathcal{P}(\lambda_n,\mu_n)).$$

It is easy to check that property P implies property L, while examples can be given [11], [15] showing that the converse is not true.

Two-dimensional systems (1.1) whose transition matrices A_1, A_2 have property P are endowed with several interesting features. Indeed, property P is equivalent to simultaneous triangularizability, a feature that allows good insight into the geometric structure of the free state evolution. In particular, it implies that there exists a maximal chain of $\{A_1,A_2\}$-invariant subspaces of the local state space X

$$\{0\} = X_0 < X_1 < X_2 < \cdots < X_n = X$$

with $\dim(X_i) = i, \; i = 0,1,2,\dots,n$.

When the local states $x(-\ell,\ell)$ of the initial global state $\mathcal{X}_0 = \sum_{\ell=-\infty}^{+\infty} x(-\ell,\ell)z_1^{-\ell}z_2^{\ell}$, are in X_i, all local states $x(h,k)$, $h+k \geq 0$, are in X_i too. Correspondingly, systems (1.1) can be viewed as cascades of two-dimensional systems of dimension one.

Moreover, systems with property P constitute a class of two-dimensional systems large enough for realizing all transfer functions $p(z_1,z_2)/q(z_1,z_2)$ with denominators of the form $q(z_1,z_2) = \prod_j(1 - \lambda_j z_1 - \mu_j z_2)$ and in particular, all transfer functions with separable denominators [1]. It should be stressed that the same is not true if we consider only commutative two-dimensional systems, i.e. systems (1.1) that satisfy the (stronger) constraint $A_1A_2 - A_2A_1 = 0$.

As a consequence of Proposition 3.1, matrix pairs endowed with property L can be equivalently described as those whose characteristic polynomials factorize into a product of linear terms. This class of polynomials, however, corresponds also to matrix pairs with property P; so there is no possibility of finding an equivalent description of property P that relies only on the characteristic polynomial. Appropriate tools turn out to be certain noncommutative polynomials [12] and power series associated with the pair, as well as the corresponding Hankel matrices.

PROPOSITION 4.1. *Let A_1, A_2 be $n \times n$ matrices with entries in $\mathbf{C}$, and consider the orderings of their spectra given in* (4.1). *The following statements are equivalent*:

P) (A_1,A_2) *has property* P *w.r.t. the orderings* (4.1);

P$_1$) *for any $w \in \Xi^*$, with $|w|_1 = h$ and $|w|_2 = k$,*

$$(4.3) \qquad \operatorname{tr}(w(A_1,A_2)) = \sum_{i=1}^{n} \lambda_i^h \mu_i^k;$$

P$_2$) *the noncommutative power series, whose coefficients are the traces of the matrices* $w(A_1, A_2)$,

$$\mathcal{N} = \sum_{w \,\in\, \Xi^*} \operatorname{tr}(w(A_1, A_2))w, \tag{4.4}$$

can be represented as $\mathcal{N} = \sum_{i=1}^{n}(1-\lambda_i\xi_1-\mu_i\xi_2)^{-1}$ *and, hence, is recognizable* [2];

P$_3$) *for any* $w \in \Xi^*$, *with* $|w|_1 = h$ *and* $|w|_2 = k$,

$$\det(zI - w(A_1, A_2)) = \prod_{i=1}^{n}(z - \lambda_i^h\mu_i^k). \tag{4.5}$$

Proof. P) $\Rightarrow$ P$_1$) If $|w|_1 = h$ and $|w|_2 = k$, the definition of property P directly implies $\Lambda(w(A_1, A_2)) = (\lambda_1^h\mu_1^k, \dots, \lambda_n^h\mu_n^k)$, and therefore (4.3) holds.

P$_1$) $\Rightarrow$ P) Extend the monoid morphisms $\phi_i : \Xi^* \to \mathbf{C} : w \mapsto \lambda_i^{|w|_1}\mu_i^{|w|_2} = w(\lambda_i, \mu_i)$, $i = 1, 2, \dots, n$, to the algebra $\mathbf{C}\langle\xi_1, \xi_2\rangle$, letting $\phi_i(\mathcal{P}) = \mathcal{P}(\lambda_i, \mu_i)$, $i = 1, 2, \dots, n$, for all $\mathcal{P}(\xi_1, \xi_2) \in \mathbf{C}\langle\xi_1, \xi_2\rangle$. Then we have

$$\phi_i(\mathcal{P}^h) = \big(\phi_i(\mathcal{P})\big)^h = \big(\mathcal{P}(\lambda_i, \mu_i)\big)^h. \tag{4.6}$$

From assumption P$_1$) we deduce that $\operatorname{tr}\big(w(A_1, A_2)\big) = \sum_{i=1}^{n} \phi_i(w)$, and hence, by the linearity of the trace operator,

$$\operatorname{tr}\big(\mathcal{P}(A_1, A_2)\big) = \sum_{i=1}^{n} \phi_i(\mathcal{P}) = \sum_{i=1}^{n} \mathcal{P}(\lambda_i, \mu_i).$$

Using (4.6), for all $h \in \mathbf{N}_+$

$$\operatorname{tr}\big(\mathcal{P}(A_1, A_2)\big)^h = \sum_{i=1}^{n} \phi_i(\mathcal{P}^h) = \sum_{i=1}^{n} \mathcal{P}(\lambda_i, \mu_i)^h, \tag{4.7}$$

which gives $\Lambda(\mathcal{P}(A_1, A_2)) = (\mathcal{P}(\lambda_1, \mu_1), \dots, \mathcal{P}(\lambda_n, \mu_n))$.

P$_1$) $\Leftrightarrow$ P$_2$) Assuming P$_1$), we may write

$$\mathcal{N} = \sum_{w \,\in\, \Xi^*} \operatorname{tr}(w(A_1, A_2)w) = \sum_{i=1}^{n} \sum_{w \,\in\, \Xi^*} \lambda_i^{|w|_1}\mu_i^{|w|_2}w.$$

On the other hand, we obtain

$$\begin{aligned}\sum_{i=1}^{n}(1 - \lambda_i\xi_1 - \mu_i\xi_2)^{-1} &= \sum_{i=1}^{n}\sum_{j=0}^{+\infty}(\lambda_i\xi_1 + \mu_i\xi_2)^j \\ &= \sum_{i=1}^{n}\sum_{j=0}^{+\infty} \sum_{\substack{w \,\in\, \Xi^* \\ |w|_1+|w|_2=j}} \lambda_i^{|w|_1}\mu_i^{|w|_2}w \\ &= \sum_{i=1}^{n} \sum_{w \,\in\, \Xi^*} \lambda_i^{|w|_1}\mu_i^{|w|_2}w,\end{aligned}$$

which proves (4.6). The converse can be shown in the same way.

P$_1$) $\Rightarrow$ P$_3$) Given $w \in \Xi^*$, for all $h \in \mathbf{N}$ we have

$$\operatorname{tr}(w(A_1, A_2))^h = \sum_{i=1}^{n} \lambda_i^{h|w|_1} \mu^{h|w|_2} = \sum_{i=1}^{n} (\lambda_i^{|w|_1} \mu_i^{|w|_2})^h.$$

Thus $(\lambda_1^{|w|_1}\mu_1^{|w|_2}, \dots, \lambda_n^{|w|_1}\mu_n^{|w|_2})$ is the spectrum of $w(A_1, A_2)$, which proves (4.5).

P$_3$) $\Rightarrow$ P$_1$) This part of the proof is obvious. □

Remark. As a consequence of P$_1$), property P can be equivalently stated referring only to the words of the free monoid Ξ^* instead of the whole algebra $\mathbf{C}\langle\xi_1, \xi_2\rangle$. Indeed, (A_1, A_2) has property P if and only if for all $w \in \Xi^*$ we have

$$\Lambda\Big(w(A_1, A_2)\Big) = (\lambda_1^{|w|_1}\mu_1^{|w|_2}, \dots, \lambda_n^{|w|_1}\mu_n^{|w|_2}),$$

where $(\lambda_1, \lambda_2, \dots, \lambda_n)$ and $(\mu_1, \mu_2, \dots, \mu_n)$ are the spectra of A_1 and A_2, suitably ordered.

As for property L, we aim now to provide an effective method for testing property P that depends on the study of the noncommutative power series $\mathcal{N}$ and the associated Hankel matrix [2], [14]. By the Hankel matrix of $\mathcal{N}$ we mean the infinite matrix $\mathcal{H}(\mathcal{N})$, whose rows and columns are indexed by the words of Ξ^* and whose element with indexes u and v is equal to $\langle \mathcal{N}, uv \rangle$.

It will be convenient to order the words in Ξ^* and, consequently, the row and column indexes in $\mathcal{H}(\mathcal{N})$, according to their length; while the lexicographical order will be adopted for words of the same length. For all $M', M'' \in \mathbf{N}$, we shall denote by $\mathcal{H}_{M' \times M''}(\mathcal{N})$ the submatrix appearing in the upper left corner of $\mathcal{H}(\mathcal{N})$, whose rows (columns) are indexed by words of length not greater than M' (M'').

LEMMA 4.2. *Let A_1, A_2 be $n \times n$ matrices with entries in $\mathbf{C}$. Then*

$$\text{rank } \mathcal{H}_{(n^2-1)\times(n^2-1)}(\mathcal{N}) = \text{rank } \mathcal{H}(\mathcal{N}) \le n^2. \tag{4.8}$$

Proof. For all $w \in \Xi^*$, we have

$$\operatorname{tr}(w(A_1, A_2)) = [\,\mathbf{e}_1^{\mathrm{T}} \quad \dots \quad \mathbf{e}_n^{\mathrm{T}}\,] \operatorname{diag}\{w(A_1, A_2), \dots, w(A_1, A_2)\} \begin{bmatrix} \mathbf{e}_1 \\ \vdots \\ \mathbf{e}_n \end{bmatrix}, \tag{4.9}$$

where $\mathbf{e}_1, \dots, \mathbf{e}_n$ are the vectors of the canonical basis of $\mathbf{C}^n$. It is clear that $\mathcal{H}(\mathcal{N})$ can be expressed as $\mathcal{H}(\mathcal{N}) = \mathcal{O}\mathcal{R}$, where $\mathcal{O}$ is the $\infty \times n^2$ matrix whose row of index $v \in \Xi^*$ is given by

$$[\,\mathbf{e}_1^{\mathrm{T}} \quad \dots \quad \mathbf{e}_n^{\mathrm{T}}\,] \ \operatorname{diag} \ \{v(A_1, A_2), \dots, v(A_1, A_2)\} \tag{4.10}$$

and, similarly, $\mathcal{R}$ is the $n^2 \times \infty$ matrix whose column of index $w \in \Xi^*$ is given by

$$\operatorname{diag} \ \{w(A_1, A_2), \dots, w(A_1, A_2)\} \begin{bmatrix} \mathbf{e}_1 \\ \vdots \\ \mathbf{e}_n \end{bmatrix}. \tag{4.11}$$

This shows that rank$\mathcal{H}(\mathcal{N}) \le n^2$.

To complete the proof, suppose that all rows of $\mathcal{O}$ indexed by words of length ν linearly depend on the rows indexed in $\mathcal{I} := \{w \in \Xi^*, |w| < \nu\}$. We deduce that any row of index $v = u\xi_i$, $|u| = \nu$, also depends on the words indexed in $\mathcal{I}$, because

$$\begin{aligned}&[\mathbf{e}_1^{\mathrm{T}} \ \dots \ \mathbf{e}_n^{\mathrm{T}}] \ \mathrm{diag}\{v(A_1, A_2), \dots, v(A_1, A_2)\}\\ &\quad = [\mathbf{e}_1^{\mathrm{T}} \ \dots \ \mathbf{e}_n^{\mathrm{T}}] \ \mathrm{diag}\{u(A_1, A_2), \dots, u(A_1, A_2)\} \ \mathrm{diag}\{A_i, \dots, A_i\}\\ &\quad = \textstyle\sum_{w\in\mathcal{I}} \alpha_w [\mathbf{e}_1^{\mathrm{T}} \ \dots \ \mathbf{e}_n^{\mathrm{T}}] \ \mathrm{diag}\{w(A_1, A_2)A_i, \dots, w(A_1, A_2)A_i\}\\ &\quad = \textstyle\sum_{w\in\mathcal{I}} \beta_w [\mathbf{e}_1^{\mathrm{T}} \ \dots \ \mathbf{e}_n^{\mathrm{T}}] \ \mathrm{diag}\{w(A_1, A_2), \dots, w(A_1, A_2)\}.\end{aligned}$$

An easy inductive argument proves that all rows of $\mathcal{O}$ linearly depend on those indexed in $\mathcal{I}$. Moreover, as rank$\mathcal{O} \le n^2$, it is clear that in the definition of $\mathcal{I}$ we can assume $\nu = n^2$. The same reasoning applies to the rows of $\mathcal{R}$, showing that

$$\mathrm{rank}\ \mathcal{H}_{(n^2-1)\times(n^2-1)}(\mathcal{N}) = \mathrm{rank}\ \mathcal{H}(\mathcal{N}). \qquad \Box$$

PROPOSITION 4.3. *Let A_1, A_2 be $n \times n$ matrices with entries in $\mathbf{C}$, and consider the associated noncommutative power series $\mathcal{N} = \sum_{w \in \Xi^*} \mathrm{tr}(w(A_1, A_2))w$. The following statements are equivalent:*

i) *(A_1, A_2) has property* P;

ii) rank $\mathcal{H}_{(n^2-1)\times(n^2-1)}(\mathcal{N}) = \bar{n} \le n$ *and, for all pairs of words $w, \bar{w}$ with length not greater than $2\bar{n}$,*

$$|w|_i = |\bar{w}|_i, \quad i = 1, 2 \quad \Rightarrow \quad \mathrm{tr}(w(A_1, A_2)) = \mathrm{tr}(\bar{w}(A_1, A_2)); \tag{4.12}$$

iii) (4.12) *holds for all pairs of words $w, \bar{w}$ with length not greater than $2n^2$.*

Proof. i) $\Rightarrow$ ii) By Proposition 4.1, property P implies that

$$\mathrm{tr}(w(A_1, A_2)) = \sum_{i=1}^{n} \lambda_i^{|w|_1} \mu_i^{|w|_2} \quad \forall w \in \Xi^*, \tag{4.13}$$

for suitable orderings of the eigenvalues of A_1 and A_2. This immediately proves (4.12). Moreover, $\mathcal{H}(\mathcal{N})$ can be expressed as $\mathcal{H}(\mathcal{N}) = \mathcal{O}\mathcal{R}$, where $\mathcal{O}$ is the $\infty \times n$ matrix whose row of index $v \in \Xi^*$ is given by $[\,v(\lambda_1, \mu_1), v(\lambda_2, \mu_2), \dots, v(\lambda_n, \mu_n)\,]$, and $\mathcal{R} = \mathcal{O}^{\mathrm{T}}$.

We deduce that rank $\mathcal{H}(\mathcal{N}) \le$ rank $\mathcal{O} \le n$ and, therefore rank $\mathcal{H}_{(n^2-1)\times(n^2-1)}$ $(\mathcal{N}) \le n$.

ii) $\Rightarrow$ iii) By the previous lemma, rank $\mathcal{H}(\mathcal{N}) = \bar{n}$. So, there exist [2], [14] $M_1, M_2 \in \mathbf{C}^{\bar{n}\times\bar{n}}, H \in \mathbf{C}^{1\times\bar{n}}$, and $G \in \mathbf{C}^{\bar{n}\times 1}$ such that $\langle\mathcal{N}, w\rangle = Hw(M_1, M_2)G$, $\forall w \in \Xi^*$, and $\mathcal{H}(\mathcal{N})$ can be expressed as $\mathcal{H}(\mathcal{N}) = \mathcal{O}\mathcal{R}$, where $\mathcal{O}$ is the $\infty \times \bar{n}$ matrix whose row indexed by $v \in \Xi^*$ is $Hv(M_1, M_2)$ and $\mathcal{R}$ is the $\bar{n} \times \infty$ matrix whose column indexed by $w \in \Xi^*$ is $w(M_1, M_2)G$.

By the same argument used in the proof of Lemma 4.2, there exist $2\bar{n}$ words $r_1, r_2, \dots, r_{\bar{n}}$ and $c_1, c_2, \dots, c_{\bar{n}}$, of length less than $\bar{n}$, such that both the $\bar{n} \times \bar{n}$ matrices $\mathcal{O}_{\bar{n}}$, and whose rows are $Hr_i(M_1, M_2)$ and $\mathcal{R}_{\bar{n}}$ and whose columns are $c_j(M_1, M_2)G$, are nonsingular. Consequently, the $\bar{n} \times \bar{n}$ submatrix of $\mathcal{H}(\mathcal{N})$

$$N_0 := [\langle\mathcal{N}, r_i c_j\rangle] = \mathcal{O}_{\bar{n}}\mathcal{R}_{\bar{n}}$$

is nonsingular too.

Introduce next the matrices

$$\bar{M}_1 := N_0^{-1}[\langle\mathcal{N}, r_i\xi_1 c_j\rangle], \qquad \bar{M}_2 := N_0^{-1}[\langle\mathcal{N}, r_i\xi_2 c_j\rangle], \tag{4.14}$$

$$\bar H := [\,\langle \mathcal{N}, c_1\rangle \quad \dots \quad \langle \mathcal{N}, c_{\bar n}\rangle\,], \quad \text{and} \quad \bar G := N_0^{-1}\begin{bmatrix} \langle \mathcal{N}, r_1\rangle \\ \vdots \\ \langle \mathcal{N}, r_{\bar n}\rangle \end{bmatrix}. \tag{4.15}$$

Using the assumption on the rank of $\mathcal{H}(\mathcal{N})$, we apply the same arguments as in §3 to derive a counterpart of (3.7) for noncommutative power series. So, for all $r, c \in \Xi^*$, we have

$$\Big[\langle \mathcal{N}, rc_1\rangle \ \dots \ \langle \mathcal{N}, rc_{\bar n}\rangle\Big] N_0^{-1}\begin{bmatrix} \langle \mathcal{N}, r_1 c\rangle \\ \vdots \\ \langle \mathcal{N}, r_{\bar n} c\rangle \end{bmatrix} = \langle \mathcal{N}, rc\rangle. \tag{4.16}$$

It follows that

$$\begin{aligned}\bar M_1 \bar M_2 - \bar M_2 \bar M_1 &= N_0^{-1}\{[\langle \mathcal{N}, r_i\xi_1 c_j\rangle] N_0^{-1}[\langle \mathcal{N}, r_i\xi_2 c_j\rangle] - [\langle \mathcal{N}, r_i\xi_2 c_j\rangle] N_0^{-1}[\langle \mathcal{N}, r_i\xi_1 c_j\rangle]\} \\ &= N_0^{-1}\ [\langle \mathcal{N}, r_i\xi_1\xi_2 c_j\rangle - \langle \mathcal{N}, r_i\xi_2\xi_1 c_j\rangle] = 0,\end{aligned}$$

because of assumption (4.12). So, $\bar M_1$ and $\bar M_2$ commute.

As an immediate consequence of (4.16) and definitions (4.14)–(4.15), we get

$$M_i G = N_0^{-1}\begin{bmatrix} \langle \mathcal{N}, r_1\xi_i\rangle \\ \vdots \\ \langle \mathcal{N}, r_{\bar n}\xi_i\rangle \end{bmatrix}, \qquad i = 1, 2, \tag{4.17}$$

and, for all $v \in \Xi^*$,

(4.18)

$$[\,\langle \mathcal{N}, v\xi_i c_1\rangle \quad \dots \quad \langle \mathcal{N}, v\xi_i c_{\bar n}\rangle\,] = [\,\langle \mathcal{N}, vc_1\rangle \quad \dots \quad \langle \mathcal{N}, vc_{\bar n}\rangle\,]\ N_0^{-1} M_i, \quad i = 1, 2.$$

Finally, we propose to prove that, for all $w \in \Xi^*$

$$\langle \mathcal{N}, w\rangle = \bar H w(\bar M_1, \bar M_2)\bar G, \tag{4.19}$$

which corresponds to showing that $\operatorname{tr}(w(A_1, A_2)) = \operatorname{tr}(\bar w(A_1, A_2))$, for all $w, \bar w \in \Xi^*$, such that $|w|_i = |\bar w|_i$, $i = 1, 2$.

Equation (4.19) is easily verified for $w = 1$. For any $w = \xi_{i_1}\xi_{i_2}\dots\xi_{i_\nu}$, $\nu \geq 1$, by (4.16) and (4.17), we have

$$\langle \mathcal{N}, \xi_{i_1}\xi_{i_2}\dots\xi_{i_\nu}\rangle = [\,\langle \mathcal{N}, \xi_{i_1}\xi_{i_2}\dots\xi_{i_{\nu-1}}c_1\rangle \dots \langle \mathcal{N}, \xi_{i_1}\xi_{i_2}\dots\xi_{i_{\nu-1}}c_{\bar n}\rangle\,]\,\bar M_{i_\nu}\bar G,$$

and, by iteratively applying (4.18),

$$\langle \mathcal{N}, \xi_{i_1}\xi_{i_2}\dots\xi_{i_\nu}\rangle = [\,\langle \mathcal{N}, c_1\rangle \quad \dots \quad \langle \mathcal{N}, c_{\bar n}\rangle\,]\,\bar M_{i_1}\bar M_{i_2}\dots\bar M_{i_\nu}\bar G = \bar H w(\bar M_1, \bar M_2)\bar G.$$

iii) ⇒ i) By Lemma 4.2, $\bar n := \operatorname{rank} \mathcal{H}_{(n^2-1)\times(n^2-1)}(\mathcal{N}) = \operatorname{rank} \mathcal{H}(\mathcal{N})$. Thus, as in the proof of ii) ⇒ iii), we can represent $\mathcal{N}$ as

$$\mathcal{N} = \bar H(I - \bar M_1\xi_1 - \bar M_2\xi_2)^{-1}\bar G, \tag{4.20}$$

where $\bar M_1, \bar M_2 \in \mathbf{C}^{\bar n\times\bar n}$ commute, and (4.12) holds for all words in Ξ^*, independently of their length. Therefore, for all w in Ξ^*, with $h = |w|_1$ and $k = |w|_2$, we have

$$\operatorname{tr}(w(A_1, A_2)) = \binom{h+k}{h}^{-1}\sum_{i=1}^{n}\operatorname{tr}({A_1}^{h}{\sqcup\!\sqcup}^{k}A_2). \tag{4.21}$$

Taking the commutative images on both sides of (4.20), we obtain

$$T_{A_1,A_2} + n = \bar{H}(I - \bar{M}_1 z_1 - \bar{M}_2 z_2)^{-1}\bar{G},$$

and, consequently,

$$T_{A_1,A_2} = -n + \frac{\bar{H}\,\mathrm{adj}(I - \bar{M}_1 z_1 - \bar{M}_2 z_2)\bar{G}}{\Delta_{\bar{M}_1,\bar{M}_2}(z_1, z_2)},$$

where $\Delta_{\bar{M}_1,\bar{M}_2}(z_1, z_2)$ splits into linear factors, because of the commutativity of $\bar{M}_1$, $\bar{M}_2$. Thus the characteristic polynomial of (A_1, A_2) is given by $\Delta_{A_1,A_2}(z_1, z_2) = \prod_{i=1}^n (1 - \lambda_i z_1 - \mu_i z_2)$, and (A_1, A_2) has property L.

As Proposition 3.1 gives $\mathrm{tr}(A_1{}^h \sqcup\!\sqcup {}^k A_2) = \binom{h+k}{h} \sum_{i=1}^n \lambda_i^h \mu_i^k$, condition (4.21) immediately implies

$$\mathrm{tr}(w(A_1, A_2)) = \sum_{i=1}^n \lambda_i^{|w|_1} \mu_i^{|w|_2} \quad \forall\, w \in \Xi^*,$$

which is equivalent to property P. □

5. Special factorizations of the characteristic polynomial. In this section, we consider a further property of a matrix pair (A_1, A_2) that, like property L, can be expressed as a constraint on the factors of Δ_{A_1,A_2} as well as a condition on the spectra of the linear combinations $\alpha A_1 + \beta A_2$, $\alpha, \beta \in \mathbf{C}$. Pairs we refer to are those whose characteristic polynomials split into the product of distinct polynomials $\Delta_i(z_1, z_2)$, each of them having support included in a straight line of the plane $\mathbf{Z} \times \mathbf{Z}$, passing through the origin. The interest in this property is mostly due to the fact that, as we shall see, it constitutes an immediate generalization of finite memory and separability.

To begin with, we consider a single polynomial $\Delta(z_1, z_2)$ whose support is a subset of a straight line in $\mathbf{Z} \times \mathbf{Z}$; i.e., there exists $(\ell, m) \neq (0, 0)$ in $\mathbf{N} \times \mathbf{N}$ such that

$$\mathrm{supp}(\Delta) \subset \{(k\ell, km), k \in \mathbf{N}\}. \tag{5.1}$$

Two-dimensional systems having Δ as characteristic polynomial exhibit several features that strictly resemble those of one-dimensional systems. Indeed, the local state at $(0, 0)$ determines a free evolution that is identically zero except on a "strip" that includes the straight line $\{(k\ell, km), k \in \mathbf{Z}\}$ (see Fig. 2, for $\ell = 2$ and $m = 1$).

So, no matter how far (h, k) is from the set $\{(i, -i) : i \in \mathbf{Z}\}$ where the initial conditions are given, the local state in (h, k) is determined only by a finite subset of the initial global state, whose cardinality does not exceed a fixed integer N.

PROPOSITION 5.1. *Let (A_1, A_2) be a pair of $n \times n$ matrices with entries in $\mathbf{C}$ and $\Delta_{A_1 A_2}(z_1, z_2)$ its characteristic polynomial. Assume moreover that (ℓ, m) is a pair of nonnegative integers and $1 = \mathrm{g.c.d.}(\ell, m)$. The following statements are equivalent:*

i) $\Delta_{A_1,A_2}(z_1, z_2) = 1 - \sum_{h=1}^r d_h (z_1^\ell z_2^m)^h$; (5.2)

ii) *there exist $c_1, c_2, \dots, c_n$ in $\mathbf{C}$ such that, for every $(\alpha, \beta) \in \mathbf{C} \times \mathbf{C}$ and every $(\ell + m)$th root of $\alpha^\ell \beta^m$,*

$$\Lambda(\alpha A_1 + \beta A_2) = \left(c_1 (\alpha^\ell \beta^m)^{\frac{1}{\ell+m}}, \dots, c_n (\alpha^\ell \beta^m)^{\frac{1}{\ell+m}}\right); \tag{5.3}$$

iii) *there exist $c_1, c_2, \dots, c_n \in \mathbf{C}$ such that*

$$\Lambda(\nu^m A_1 + \nu^{-\ell} A_2) = (c_1, c_2, \dots, c_n) \quad \forall\, \nu \in \{1, 2, \dots (\ell + m)n + 1\}; \tag{5.4}$$

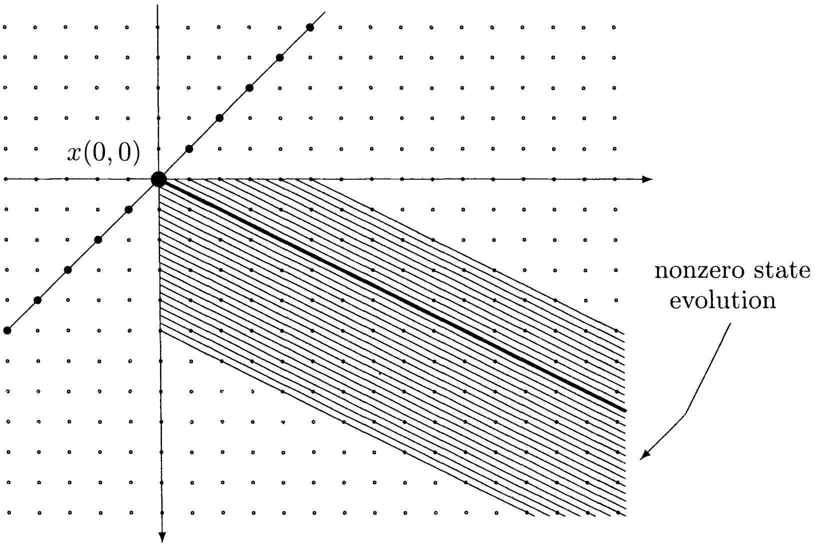


FIG. 2

iv) $(i,j) \notin \{(k\ell, km), k \in \mathbf{N}_+\}$ *implies* $\operatorname{tr}(A_1{}^i \sqcup\!\sqcup^j A_2) = 0$;
v) *for all* $\alpha, \beta \in \mathbf{C}$ *and suitable* $b_k \in \mathbf{C}$,

$$\operatorname{tr}(\alpha A_1 + \beta A_2)^k = \begin{cases} b_k(\alpha^\ell \beta^m)^\nu & \text{if } k = (\ell+m)\nu, \\ 0 & \text{otherwise;} \end{cases} \tag{5.5}$$

vi) $(i,j) \notin \mathcal{S}_n := \{(i,j) \in \mathbf{N} \times \mathbf{N} :\mid mi - \ell j \mid < \mathbf{N}\}$ *implies* $A_1{}^i \sqcup\!\sqcup^j A_2 = 0$.

Proof. i) $\Rightarrow$ ii) Since $\Delta_{A_1,A_2}(z_1,z_2) \in \mathbf{C}[z_1^\ell z_2^m]$, there exist $\lambda_1, \lambda_2, \ldots, \lambda_r \in \mathbf{C}$ such that

$$\Delta_{A_1,A_2}(z_1,z_2) = \prod_{h=1}^{r}(1 - \lambda_h z_1^\ell z_2^m)$$

and, consequently,

$$\det(zI - \alpha A_1 - \beta A_2) = z^{n-r(\ell+m)} \prod_{h=1}^{r}(z^{\ell+m} - \lambda_h \alpha^\ell \beta^m). \tag{5.6}$$

Let $(\lambda_h)^{\frac{1}{\ell+m}}$ and $(\alpha^\ell\beta^m)^{\frac{1}{\ell+m}}$ be arbitrary $(\ell{+}m)$th roots of λ_h and $\alpha^\ell\beta^m$, respectively, and ε any primitive $(\ell+m)$th root of 1. The spectrum of $(\alpha A_1 + \beta A_2)$ is given by $\Lambda(\alpha A_1 + \beta A_2) = \left(c_1(\alpha^\ell\beta^m)^{\frac{1}{\ell+m}}, \ldots, c_n(\alpha^\ell\beta^m)^{\frac{1}{\ell+m}}\right)$, where

$$\begin{aligned} c_{r\nu+h} &= (\lambda_h)^{\frac{1}{\ell+m}}\varepsilon^\nu, && h = 1,\ldots,r \quad \text{and} \quad \nu = 1,\ldots,\ell+m, \\ c_\mu &= 0, && \mu > (\ell+m)r. \end{aligned}$$

ii) $\Rightarrow$ iii) This part of the proof is obvious.

iii) $\Rightarrow$ iv) Clearly, for all $\nu \in \{1,2,\ldots,(\ell+m)n+1\}$ and $h \in \mathbf{N}_+$

$$\operatorname{tr}(\nu^m A_1 + \nu^{-\ell} A_2)^h = \sum_{i=0}^{h} \nu^{(\ell+m)i-h\ell} \operatorname{tr}(A_1{}^i \sqcup\!\sqcup^{h-i} A_2) = \sum_{i=1}^{n} c_i^h =: f_h.$$

whence $\sum_{i=0}^{h} \nu^{(\ell+m)i}\operatorname{tr}(A_1{}^{i}\sqcup\!\sqcup^{h-i}A_2) - f_h\nu^{h\ell} = 0$. As in the polynomials

$$p_h(x) := \sum_{i=0}^{h} x^{(\ell+m)i}\operatorname{tr}(A_1{}^{i}\sqcup\!\sqcup^{h-i}A_2) - f_h x^{h\ell}, \qquad h = 1, 2, \ldots, n,$$

the number of zeros exceeds the degree, all their coefficients have to be zero. We distinguish two cases.

Case 1. $k(\ell+m) = h\ell$, for some $k \in \mathbf{N}$. Since 1 is the unique common divisor of ℓ and m, there exists $t \in \mathbf{N}_+$ such that $k = \ell t$ and $h - k = mt$, and therefore

$$\operatorname{tr}(A_1{}^{i}\sqcup\!\sqcup^{h-i}A_2) = \begin{cases} f_h & \text{if } (i, h-i) = t(\ell, m), \\ 0 & \text{otherwise.} \end{cases}$$

Case 2. $k(\ell+m) \neq h\ell$ for all $h \in \mathbf{N}$. Then, for $0 \leq i \leq h$, tr $(A_1{}^{i}\sqcup\!\sqcup^{h-i}A_2) = 0$.

iv) $\Rightarrow$ v) This part of the proof is obvious.

v) $\Rightarrow$ i) Equations (2.15) and (2.16) show that the homogeneus form δ_k of the characteristic polynomial satisfies

$$\delta_k(\alpha, \beta) = \begin{cases} d_\nu(\alpha^\ell\beta^m)^\nu & \text{if } k = (\ell+m)\nu, \\ 0 & \text{otherwise.} \end{cases}$$

i) $\Rightarrow$ vi) Note that

$$\sum_{i,j=0}^{\infty} A_1{}^{i}\sqcup\!\sqcup^{j}A_2 \; z_1^i z_2^j = (I - A_1z_1 - A_2z_2)^{-1} = \frac{\operatorname{adj}(I - A_1z_1 - A_2z_2)}{1 - \sum_{h=1}^{r} d_h(z_1^\ell z_2^m)^h}.$$

As

$$\begin{aligned} \operatorname{supp}(\operatorname{adj}(I - A_1z_1 - A_2z_2)) &\subseteq \{(i,j) \in \mathbf{N}\times\mathbf{N} : i + j \leq n\}, \\ \operatorname{supp}((1 - \textstyle\sum_{h=1}^{r} d_h(z_1^\ell z_2^m)^h)^{-1}) &\subseteq \{(i,j) \in \mathbf{N}\times\mathbf{N} : mi = \ell j\}, \end{aligned}$$

it is clear that the support of $(I - A_1z_1 - A_2z_2)^{-1}$ is a subset of $\mathcal{S}_{\mathbf{N}}, \mathbf{N} = n - \max(1, m)$.

vi) $\Rightarrow$ i) Consider the injective ring homomorphism $\phi : \mathbf{C}[[z_1, z_2]] \to \mathbf{C}[[\eta, \xi, \xi^{-1}]]$ obtained by linearly extending the map that associates $z_1^i z_2^j$, $i, j \in \mathbf{N}$, with $\eta^k\xi^h$, where h, k are given by

$$\begin{bmatrix} h \\ k \end{bmatrix} = \begin{bmatrix} -m & \ell \\ 0 & 1 \end{bmatrix} \begin{bmatrix} i \\ j \end{bmatrix}.$$

ϕ maps any series in $\mathbf{C}[[z_1, z_2]]$ with support in $\mathcal{S}_{\mathbf{N}}$ into an element of $\mathbf{C}[[\eta]][\xi, \xi^{-1}]$, the ring of Laurent polynomials [12] in the indeterminate ξ, with coefficients in $\mathbf{C}[[\eta]]$. As the support of $(I - A_1z_1 - A_2z_2)^{-1}$ is included in $\mathcal{S}_{\mathbf{N}}$ and hence

$$\operatorname{supp}\Big(\det(I - A_1z_1 - A_2z_2)^{-1}\Big) \subseteq \mathcal{S}_{\mathbf{N}^2},$$

applying the map ϕ on both sides of $\det(I - A_1z_1 - A_2z_2)^{-1}\Delta_{A_1,A_1}(z_1, z_2) = 1$, one gets

$$\phi(\det(I - A_1z_1 - A_2z_2)^{-1}\phi(\Delta_{A_1,A_2}(z_1, z_2)) = 1. \tag{5.7}$$

As both factors on the left-hand side of (5.7) can be viewed as elements of $\mathbf{C}[[\eta]][\xi, \xi^{-1}]$, we have that $\phi(\Delta_{A_1,A_2}(z_1, z_2))$ is a unit of that ring, i.e. $\phi(\Delta_{A_1,A_2}(z_1, z_2)) = \xi^h s(\eta)$

for some $h \in \mathbf{Z}$ and $s(\eta) \in \mathbf{C}[[\eta]]$. Condition $\Delta_{A_1,A_2}(0,0) = 1$ implies $h = 0$ and, therefore, Δ_{A_1,A_2} is a polynomial in $z_1^\ell z_2^m$. □

An immediate consequence of property ii) in the above proposition is the following corollary.

COROLLARY 5.2. *Consider A_1, A_2 in $\mathbf{C}^{n\times n}$. If* $\operatorname{supp}(\Delta_{A_1,A_2})$ *is a subset of a straight line, different from the coordinate axes, then both A_1 and A_2 are nilpotent.*

The results of Proposition 5.1 provide a convenient framework for understanding the internal dynamics of two-dimensional finite memory state models, which arise quite naturally in several applications. For instance, when considering the realization of F.I.R. filters and dead beat regulators, the requirements on the state models that we have to use cannot be exclusively expressed as conditions on the polynomial transfer matrix that represents the input–output map. Further aspects should be taken into account, which introduce additional constraints on the choice of the matrix pair (A_1, A_2).

a) If the state–output transfer matrix $C(I - A_1z_1 - A_2z_2)^{-1}$ is not polynomial, local states $\mathbf{x}$ exist, which give rise to free output evolutions $C(I - A_1z_1 - A_2z_2)^{-1}\mathbf{x}$ with infinite supports. Clearly such states, when induced by noise, generate infinite error sequences in the output signal.

b) If the input–state transfer matrix $(I - A_1z_1 - A_2z_2)^{-1}(B_1z_1 + B_2z_2)$ is not polynomial, finite support input sequences possibly produce infinite support sequences in the state space. Therefore the system could remain indefinitely excited by a finite signal, even though the corresponding output dies out in a finite number of steps.

Both previous drawbacks can be avoided if $(I - A_1z_1 - A_2z_2)^{-1}$ is polynomial or, equivalently, if the characteristic polynomial of the system is unitary, i.e.

$$\Delta_{A_1,A_2}(z_1, z_2) = \det(I - A_1z_1 - A_2z_2) = 1. \tag{5.8}$$

Two-dimensional systems satisfying condition (5.8) are called "finite memory" [4], [6], since they reach the zero state in a finite number of steps after zeroing the input signal.

COROLLARY 5.3 (Finite memory systems [5], [9], [17]). *Let A_1, A_2 be in $\mathbf{C}^{n\times n}$. The followings are equivalent*:

FM$_1$) $\Delta_{A_1,A_2}(z_1, z_2) = 1$;

FM$_2$) $\Lambda(\alpha A_1 + \beta A_2) = (0, 0, \ldots, 0)$, $\forall \alpha, \beta \in \mathbf{C}$, *namely A_1 and A_2 are nilpotent and satisfy property* L;

FM$_3$) $\Lambda(\nu A_1 + A_2) = \Lambda(A_1 + \nu A_2) = (0, 0, \ldots, 0)$, $\nu = 1, \ldots, n + 1$;

FM$_4$) $\operatorname{tr}(A_1{}^i\sqcup\!\sqcup^j A_2) = 0$, $\forall\ (i, j) \neq (0, 0)$;

FM$_5$) $A_1{}^i\sqcup\!\sqcup^j A_2 = 0$, *for $i + j \geq n$.*

Proof. Condition FM$_1$) is equivalent to assume that the support of $\Delta_{A_1A_2}(z_1, z_2)$ is a subset of both $\{(i, 0) : i \in \mathbf{N}\}$ and $\{(0, j) : j \in \mathbf{N}\}$. Therefore finite memory systems are exactly those that satisfy properties i)–vi) of Proposition 5.1 both for $(\ell, m) = (1, 0)$ and $(\ell, m) = (0, 1)$.

FM$_2$) Choose first $(\ell, m) = (0, 1)$ and then $(\ell, m) = (1, 0)$. Then

$$\Lambda(\alpha A_1 + \beta A_2) = (c_1\beta, \ldots, c_n\beta) = (d_1\alpha, \ldots, d_n\alpha) \qquad \forall \alpha, \beta \in \mathbf{C},$$

which obviously implies $\Lambda(\alpha A_1 + \beta A_2) = (0, \ldots, 0)$.

FM$_3$) Assumptions $(\ell, m) = (0, 1)$ and $(\ell, m) = (1, 0)$ give $\Lambda(\alpha A_1 + \beta A_2) = (c_1, \ldots, c_n)$ and $\Lambda(\alpha A_1 + \beta A_2) = (d_1, \ldots, d_n)$, respectively. These imply

$$\Lambda(\alpha A_1 + \beta A_2) = (d_1\alpha, \ldots, d_n\alpha) = (c_1\beta, \ldots, c_n\beta) \qquad \forall \alpha, \beta \in \mathbf{C}.$$

Therefore $\Lambda(\nu A_1 + A_2) = \Lambda(A_1 + \nu A_2) = (0, \dots, 0)$.

FM$_4$) This part of the proof is obvious from iv) of Proposition 5.1.

FM$_5$) Assume first $(\ell, m) = (0,1)$ and then $(\ell, m) = (1,0)$. Point vi) of Proposition 5.1 gives $A_1{}^i\sqcup\!\sqcup^j A_2 = 0$, when $i \geq n$ or $j \geq n$. So $(I - A_1 z_1 - A_2 z_2)^{-1}$ is a polynomial matrix and coincides with adj $(I - A_1 z_1 - A_2 z_2)$, whose support is included in $\{(i,j) : i + j < n\}$. □

The results of Proposition 5.1 partially extend to the case of a characteristic polynomial $\Delta(z_1, z_2)$ that factorizes into irreducible factors, each of them having support on a straight line through $(0,0)$. For sake of simplicity, we confine ourselves to the case when Δ factorizes as

$$\Delta(z_1, z_2) = \Delta_1(z_1, z_2)\Delta_2(z_1, z_2), \tag{5.9}$$

with

$$\Delta_i(z_1, z_2) = 1 - \sum_{j=1}^{r_i} d_j^{(i)} (z_1^{\ell_i} z_2^{m_i})^j, \quad i = 1, 2, \tag{5.10}$$

and g.c.d.$(\ell_i, m_i) = 1$. The extension to the case of more than two factors is straightforward. If (A_1, A_2) is an $n \times n$ matrix pair with characteristic polynomial Δ, it can be easily shown that

i) there exist two positive integers ρ and σ, $\rho + \sigma \leq n$, and $\rho + \sigma$ complex numbers $c_1, \dots, c_\rho, d_1, \dots, d_\sigma$, such that, for all $\alpha, \beta \in \mathbf{C}$

$$\begin{aligned}\Lambda(\alpha A_1 + \beta A_2) = \Big(&c_1(\alpha^{\ell_1}\beta^{m_1})^{\frac{1}{\ell_1+m_1}}, \dots, c_\rho(\alpha^{\ell_1}\beta^{m_1})^{\frac{1}{\ell_1+m_1}},\\ &d_1(\alpha^{\ell_2}\beta^{m_2})^{\frac{1}{\ell_2+m_2}}, \dots, d_\sigma(\alpha^{\ell_2}\beta^{m_2})^{\frac{1}{\ell_2+m_2}}, 0, \dots, 0\Big);\end{aligned}$$

ii) $\operatorname{tr}(A_1{}^i\sqcup\!\sqcup^j A_2) \neq 0$ implies either $(i,j) = (k\ell_1, km_1)$ or $(i,j) = (h\ell_2, hm_2)$, $h, k \in \mathbf{N}_+$.

Conversely, each of the above properties guarantees that Δ factorizes as in (5.9)–(5.10).

We are now in a position to obtain a fairly complete description of two-dimensional systems whose characteristic polynomials factorize into the product of a polynomial in z_1 and a polynomial in z_2. Such systems are called "separable" [4], [5] and are usually thought of as the simplest examples of I.I.R. two-dimensional systems. Actually, many properties that one may hope to extrapolate from an understanding of one-dimensional systems carry over to separable systems. Indeed, just the knowledge that the system is separable allows one to make fairly strong statements about its behaviour; in particular, internal stability can be quickly deduced from the general theory of discrete-time one-dimensional systems, as the long-term performance of separable systems is determined by the eigenvalues of A_1 and A_2.

PROPOSITION 5.4 (Separable systems). *Let A_1, A_2 be in $\mathbf{C}^{n\times n}$. The following statements are equivalent:*

S$_1$) $\Delta_{A_1,A_2}(z_1, z_2) = r(z_1)s(z_2)$;

S$_2$) *A_1 and A_2 satisfy property* L *w.r.t. the orderings of the spectra*

$$\Lambda(A_1) = (\lambda_1, \dots, \lambda_\rho, 0, \dots, 0, 0, \dots, 0)$$

$$\Lambda(A_2) = (0, \dots, 0, \mu_1, \dots, \mu_\sigma, 0, \dots, 0),$$

so that, for every $\alpha, \beta \in \mathbf{C}$ $\Lambda(\alpha A_1 + \beta A_2) = (\alpha\lambda_1, \dots, \alpha\lambda_\rho, \beta\mu_1, \dots, \beta\mu_\sigma, 0, \dots, 0)$;

S_3) $\mathrm{tr}(A_1{}^i\sqcup\!\sqcup^j A_2) = 0$ *if both i and j are nonzero*;

S_4) $\mathrm{tr}(\alpha A_1 + \beta A_2)^k = \mathrm{tr}(\alpha A_1)^k + \mathrm{tr}(\beta A_2)^k,\ \forall \alpha, \beta \in \mathbf{C}, k \in \mathbf{N}_+$.

Property L, separability, and finite memory have been introduced by progressively strengthening the constraints on the irreducible factors of $\Delta_{A_1,A_2}(z_1, z_2)$. On the other hand, the set of matrix pairs with property L properly includes the set of pairs with property P, which in turn is strictly larger than the set of commutative pairs. So, the question naturally arises whether the above constraints on the characteristic polynomial of the pair (A_1, A_2) can be related to property P and to commutativity.

We first observe that examples can be given of commutative pairs and, hence, of pairs with property P that are not finite memory and not even separable. Actually, just by taking diagonal matrices A_1 and A_2, we easily see that commutativity and property P do not imply any particular consequence on the characteristic polynomial, except that it factorizes into first-order factors.

On the other hand, the pair

$$A_1 = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{bmatrix}, \qquad A_2 = \begin{bmatrix} 0 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & -1 & 0 \end{bmatrix}$$

is finite memory (and hence separable). Yet, it does not satisfy property P.

In view of this, no implication exists between commutativity and property P on one side and finite memory and separability on the other. What is remarkable, however, is that if we restrict our analysis to matrix pairs with property P, finite memory and separable pairs can be nicely characterized in terms of semigroups of nilpotent matrices. This is made precise in the following proposition, which provides a slight extension (and an alternative proof) of a classical Levitzki theorem [8, p. 135].

PROPOSITION 5.5. *Let A_1, A_2 be in $\mathbf{C}^{n\times n}$, and consider the multiplicative semigroups*

$$S := \{w(A_1, A_2),\ w \in \Xi^*,\ |w| \geq 1\}$$

and

$$\bar{S} := \{w(A_1, A_2),\ w \in \Xi^*,\ |w|_1 \geq 1,\ |w|_2 \geq 1\}.$$

The pair (A_1, A_2) has finite memory and property P *(respectively, separability and property* P*) if and only if all matrices in S (respectively, in $\bar{S}$) are nilpotent.*

Proof. We first remark that, if (A_1, A_2) has property P, the nilpotency of all elements of S and $\bar{S}$ is equivalent to finite memory and separability, respectively. In fact, when A_1 and A_2 are in upper triangular form, the nilpotency of the elements of S and $\bar{S}$ corresponds to the assumption that the characteristic polynomial Δ_{A_1,A_2} satisfies FM_1) of Corollary 5.3 and S_1) of Proposition 5.4, respectively.

Suppose now that the multiplicative semigroup S, generated by A_1 and A_2, is constituted by nilpotents. Since we have $\mathrm{tr}\ w(A_1, A_2) = 0$ for all $w \in \Xi^*$, $|w| \geq 1$, (4.3) is clearly fulfilled. Consequently (A_1, A_2) has property P and, by the above remark, (A_1, A_2) is a finite memory pair.

On the other hand, assume that all matrices in $\bar{S}$ are nilpotent. This implies

$$\mathrm{tr}[(A_1A_2 - A_2A_1)w(A_1, A_2)] = 0 \quad \forall w \in \Xi^*, \tag{5.11}$$

which is a necessary and sufficient condition [13] for the pair (A_1, A_2) having property P. Again, the remark at the beginning of the proof shows that (A_1, A_2) is a separable pair. □

REFERENCES

[1] M. Bisiacco, E. Fornasini, and G. Marchesini, *2D partial fraction expansions and minimal commutative realizations*, IEEE Trans. Circuits and Systems, CAS-35 (1988), pp. 1533–1538.

[2] M. Fliess, *Un outil algébrique: Les séries formelles non commutatives*, Lecture Notes in Econom. and Math. Systems 131, Springer, New York, 1975, pp.122–149.

[3] E. Fornasini and G. Marchesini, *Doubly indexed dynamical systems: State space models and structural properties*, Math. Systems Theory, 12 (1978), pp. 59–72.

[4] ———, *Properties of pairs of matrices and state models for 2D systems. Pt.* I: *State dynamics and geometry of the pairs*; *Pt.* II: *Models structure and realization problems*, in Multivariate Analysis: Future Directions, C. R. Rao, ed., North-Holland Ser. Probab. Statist., 5, North-Holland, Amsterdam, 1993, pp. 131–180.

[5] E. Fornasini, G. Marchesini, and M. E. Valcher, *On the structure of finite memory and separable 2D systems*, Automatica, 30 (1994), pp. 347–350.

[6] E. Fornasini and S. Zampieri, *A note on the state space realization of 2D FIR transfer functions*, Systems Control Letters, 16 (1990), pp. 17–22.

[7] N. Jacobson, *Basic Algebra* I, Freeman, San Francisco, 1974.

[8] I. Kaplansky, *Fields and Rings*, Chicago Univ. Press, Chicago, IL, 1972.

[9] E. Mathes, M. Omladic, and H. Radjavi, *Linear spaces of nilpotent matrices*, Linear Algebra Appl., 149 (1991), pp. 215–225.

[10] N. H. McCoy, *On the characteristic roots of matric polynomials*, Bull. Amer. Math. Soc., 42 (1936), pp. 592–600.

[11] T. S. Motzkin and O. Taussky, *Pairs of matrices with property L*, Trans. Amer. Math. Soc., 73 (1952), pp. 108–114.

[12] ———, *Pairs of matrices with property L.* (II), Trans. Amer. Math. Soc., 80 (1955), pp. 387–401.

[13] H. Radjavi, *A trace condition equivalent to simultaneous triangularizability*, Canad. J. Math., XXXVIII (1986), pp. 376–386.

[14] A. Salomaa and M. Soittola, *Automata-theoretic aspects of formal power series*, Springer-Verlag, New York, 1978.

[15] O.Taussky, *Some results concerning the transition from the L- to the P-property for pairs of finite matrices*, J. Algebra, 20 (1972), pp. 271-283.

[16] ———, *Commutativity in finite matrices*, Amer. Math. Monthly, 64 (1957), pp. 229–235.

[17] M. E.Valcher and E. Fornasini, *Polynomial inverses or two-dimensional transfer matrices and finite memory realization via inverse systems*, Multidimens. Systems Signal Process., 4 (1993), pp. 269–284.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 33, No. 4, pp. 1151–1169, July 1995

© 1995 Society for Industrial and Applied Mathematics

011

LYAPUNOV-LIKE TECHNIQUES FOR STOCHASTIC STABILITY*

PATRICK FLORCHINGER†

Abstract. The purpose of this paper is to study the stabilizability problem for control stochastic nonlinear systems driven by a Wiener process. Sufficient conditions for the existence of stabilizing feedback laws that are smooth, except possibly at the equilibrium point of the system, are provided by means of stochastic Lyapunov-like techniques. The notion of dynamic asymptotic stability in probability of control stochastic differential systems is introduced and the stabilization by means of dynamic controllers is studied.

Key words. stochastic stability, control stochastic differential equation, feedback law

AMS subject classifications. 60H10, 93C10, 93D05, 93D15, 93E15

Introduction. The stabilization of deterministic nonlinear control systems has been widely studied in the last past years by many authors (see [3], [1], [5], [17], [23], [27], and [28] for example). In these papers, sufficient conditions for the existence of smooth stabilizing feedback laws are provided using the Lyapunov machinery. Nevertheless, Sontag and Sussmann [26] have proved that in general, controllable deterministic nonlinear systems cannot be stabilized by means of continuous feedback laws. Therefore, if one does not assume the smoothness of the stabilizing feedback laws, some extensions of the results listed above have been obtained (see [3], [15], [22], and [23] for example).

Different types of stabilizing feedback laws have been studied by different authors. Sontag [22] has studied piecewise linear feedback laws, whereas Artstein has studied relaxed feedback laws [3]. Furthermore, note than in the case where the system is affine in the control, stabilizing feedback laws that are continuous for every $x \neq 0$, in a neighbourhood of the origin, can be computed by means of Lyapunov-like techniques (see [17], [24], [27], or [28]).

Actually, only few results on the stabilization of nonlinear stochastic systems can be found in the literature. The stabilization of linear stochastic control systems has been adressed by Willems and Willems [29] and by Gao and Ahmed [12]. In [29], sufficient conditions based on the properties of the solution of the algebraic Riccati equation are provided. The existence of stabilizing feedback laws for a class of nonlinear stochastic control systems has been discussed by Gao and Ahmed [13]. The procedure used in [13] is based on the properties of the solution of the stochastic algebraic Riccati equation introduced by Wonham [30]. More recently, control nonlinear stochastic systems, the drift of which is affine in the control, have been studied from the point of view of the stabilization by means of the stochastic Lyapunov theorem developed by Khasminskii [18] (see [6], [8], [10], [11] for different types of nonlinear stochastic systems and [7], [9] for stochastic bilinear systems).

The aim of this paper is to extend the stabilization results proved in [28] by Tsinias to control stochastic differential systems driven by a Wiener process.

This paper is divided in four parts and is organized as follows. In §1, we recall some definitions and results, proved by Khasminskii [18], on the Lyapunov stability

* Received by the editors July 19, 1993; accepted for publication (in revised form) February 1, 1994.

† URA CNRS No 399, Département de Mathématiques, UFR MIM, Université de Metz, Ile du Saulcy, F 57045 Metz Cedex, France, and INRIA Lorraine, Projet Congé, CESCOM, Technopôle de Metz 2000, 4 Rue Marconi, F 57070 Metz, France.

in probability of the equilibrium solution of a control stochastic differential equation. In §2, we introduce the class of control stochastic differential systems and the associated notions of stochastic stabilizability with which we deal in this paper. In §3, we study stochastic differential systems, the drift of which is affine in the control. The main results proved in this section extend the Artstein [3] and Jurdjevic–Quinn [17] theorems. Moreover, a stabilization result is also proved for a class of stochastic bilinear systems and for some nonlinear stochastic differential systems by means of a slight extension of the stochastic Jurdjevic–Quinn theorem obtained previously. In §4, we deal with the stabilization of nonlinear stochastic differential systems by means of dynamic feedback laws. Necessary and sufficient conditions for the dynamic stabilization of control nonlinear stochastic systems are provided.

1. Some elements of stochastic stability. The purpose of this section is to recall some basic facts about the Lyapunov functional approach of stochastic stability theory that we need in the sequel. For a more detailed exposition of this subject refer to Khasminskii [18], Arnold [2], and Mao [21].

Consider a complete probability space $(\Omega, \mathcal{F}, P)$ and a standard $\mathbb{R}^m$-valued Wiener process w defined on this space. Denote by $\{\mathcal{F}_t\}_{t\in\mathbb{R}_+}$ the complete right-continuous filtration generated by the Wiener process w; i.e., for any $t \in \mathbb{R}_+$,

$$\mathcal{F}_t = \sigma(w_s; 0 \leq s \leq t) \vee \mathcal{N}$$

where $\mathcal{N}$ is the class of all P-negligible sets.

In the rest of this paper, if x_t is a semimartingale on the probability space $(\Omega, \mathcal{F}, \{\mathcal{F}_t\}_{t\in\mathbb{R}_+}, P)$, the term $\circ dx_t$ (respectively, dx_t) will denote its differential in the sense of Stratonovitch (respectively, Itô) (see for example Ikeda–Watanabe [16]).

Let b and σ_k, $1 \leq k \leq m$, be $(m+1)$ functionals mapping $\mathbb{R}^n$ into $\mathbb{R}^n$ such that:

1. $b(0) = 0$ and $\sigma_k(0) = 0$ for any $k \in \{1, \dots, m\}$.
2. There exists a nonnegative constant K for which

$$\forall x \in \mathbb{R}^n, \qquad |b(x)|^2 + \sum_{k=1}^{m} |\sigma_k(x)|^2 \leq K(1 + |x|^2),$$

$$\forall x, y \in \mathbb{R}^n, \qquad |b(x) - b(y)| + \sum_{k=1}^{m} |\sigma_k(x) - \sigma_k(y)| \leq |x - y|.$$

Consider the stochastic process solution $x_t \in \mathbb{R}^n$ of the stochastic differential equation

$$x_t = x_0 + \int_0^t b(x_s)\, ds + \sum_{k=1}^{m} \int_0^t \sigma_k(x_s)\, dw_s^k \tag{1}$$

where x_0 is given in $\mathbb{R}^n$. For any $s \in \mathbb{R}_+$ and $x \in \mathbb{R}^n$, denote by $x_t^{s,x}$, $s \leq t$, the solution at time t of the stochastic differential equation (1) starting form the state x at time s.

Next, we introduce the notions of stochastic stability used in this paper.

DEFINITION 1.1. 1) *The equilibrium solution $x_t \equiv 0$ of the stochastic differential equation* (1) *is said to be stable in probability if for any $s \geq 0$ and $\epsilon > 0$,*

$$\lim_{x\to 0} P\left(\sup_{s\leq t} |x_t^{s,x}| > \epsilon\right) = 0.$$

2) *The equilibrium solution* $x_t \equiv 0$ *of the stochastic differential equation* (1) *is said to be locally asymptotically stable in probability if it is stable in probability and for any* $s \geq 0$,

$$\lim_{x \to 0} P\left(\lim_{t \to +\infty} |x_t^{s,x}| = 0\right) = 1.$$

Remark 1.2. It should be noted that, in the case $\sigma_k \equiv 0$, $1 \leq k \leq m$, these definitions reduce to the corresponding deterministic ones.

Denoting by L the infinitesimal generator of the stochastic process solution x_t of the stochastic differential equation (1), that is, L is the second-order differential operator defined for any function Ψ in $C^2(\mathbb{R}^n)$ by

$$L\Psi(x) = \sum_{i=1}^{n} b^i(x) \frac{\partial \Psi}{\partial x_i}(x) + \frac{1}{2} \sum_{i,j=1}^{n} a_{ij}(x) \frac{\partial^2 \Psi}{\partial x_i \partial x_j}(x) \tag{2}$$

where $a_{ij}(x) = \sum_{k=1}^{m} \sigma_k^i(x) \sigma_k^j(x)$, $1 \leq i, j \leq n$, one can prove the following stochastic version of the Lyapunov theorem.

THEOREM 1.3. *Assume that there exists a Lyapunov function* V *defined in a bounded open neighbourhood* D *of* $x = 0$ (*i.e., a proper function* V *of class* C^2 *mapping* D *into* $\mathbb{R}$ *that is positive definite*) *such that*

$$LV(x) \leq 0 \quad (\textit{respectively}, \;\; LV(x) < 0)$$

for any $x \in D$, $x \neq 0$; *then the equilibrium solution* $x_t \equiv 0$ *of the stochastic differential equation* (1) *is stable* (*respectively, locally asymptotically stable*) *in probability.*

Remark 1.4. If one assumes that $\sigma_k \equiv 0$, $1 \leq k \leq m$, Theorem 1.3 reduces to the well-known Lyapunov theorem for deterministic systems (see, for example, [25] or [14]).

For a detailed proof of Theorem 1.3, we refer the reader to Khasminskii [18, Chap. V, pp. 156–171] or Arnold [2, Chap. XI, pp. 176–187].

2. Setting of the problem. The purpose of this section is to introduce the class of control stochastic differential systems with which we are concerned in this paper.

Denote by $(\Omega, \mathcal{F}, P)$ a probability space and by w a standard $\mathbb{R}^m$-valued Wiener process defined on this space.

Consider the multi-input stochastic differential system in $\mathbb{R}^n$

$$x_t = x_0 + \int_0^t f(x_s, u)\, ds + \int_0^t g(x_s)\, dw_s \tag{3}$$

where

1. x_0 is given in $\mathbb{R}^n$.
2. u is an $\mathbb{R}^p$-valued control law.
3. f and g are Lipschitz functionals mapping $\mathbb{R}^n \times \mathbb{R}^p$ (respectively, $\mathbb{R}^n$) into $\mathbb{R}^n$ (respectively, $\mathbb{R}^n \times \mathbb{R}^m$) such that $f(0,0) = 0$ and $g(0) = 0$, and there exists a nonnegative constant K such that for any $x \in \mathbb{R}^n$ and $u \in \mathbb{R}^p$,

$$|f(x,u)| + |g(x)| \leq K(1 + |x| + |u|).$$

The stochastic differential system (3) is said to be locally feedback stabilizable in probability at the origin if there exist a neighbourhood D of the origin in $\mathbb{R}^n$ and a functional ϕ mapping D into $\mathbb{R}^p$ such that:

1. $\phi(0) = 0$.
2. For every $x \in D$, the solution $x_t^{0,x}$ of the closed-loop system

$$x_t = x + \int_0^t f(x_s, \phi(x_s))\, ds + \int_0^t g(x_s)\, dw_s \tag{4}$$

is uniquely defined.

3. The equilibrium solution $x_t \equiv 0$ of the closed-loop system (4) is asymptotically stable in probability.

The concept of dynamic stabilization for deterministic nonlinear control systems was introduced by Sontag and Sussmann [26]. Here, we introduce an extension of this concept to stochastic differential system as follows.

The stochastic differential system (3) is said to be locally dynamic asymptotically stabilizable in probability at the origin if the stochastic differential system

$$\begin{pmatrix} x_t \\ z_t \end{pmatrix} = \begin{pmatrix} x_0 \\ 0 \end{pmatrix} + \int_0^t \begin{pmatrix} f(x_s, z_s) \\ v \end{pmatrix} ds + \int_0^t \begin{pmatrix} g(x_s) \\ 0 \end{pmatrix} dw_s \tag{5}$$

is locally feedback stabilizable in probability at the origin.

To illustrate the main ideas of this paper, we introduce the following class of nonlinear stochastic differential systems that are affine in the control

$$x_t = x_0 + \int_0^t \left(X_0(x_s) + \sum_{j=1}^p u^j Y_j(x_s) \right) ds + \sum_{i=1}^m \int_0^t X_i(x_s) \circ dw_s^i \tag{6}$$

where $X_0, X_1, \ldots, X_m, Y_1, \ldots, Y_p$ are $(m+p+1)$ vector fields in $C_b^1(\mathbb{R}^n, \mathbb{R}^n)$, which we write for any $x \in \mathbb{R}^n$ as

$$X_i(x) = \sum_{k=1}^n X_i^k(x) \frac{\partial}{\partial x_k}\ , \quad 0 \le i \le m,$$

$$Y_j(x) = \sum_{k=1}^n Y_j^k(x) \frac{\partial}{\partial x_k}\ , \quad 1 \le j \le p,$$

and such that the vector fields X_i, $0 \le i \le m$, vanish at the origin.

Introduce the second-order differential operator L associated with the uncontrolled part of the stochastic differential system (6) defined by

$$L = X_0 + \frac{1}{2} \sum_{i=1}^m X_i^2.$$

Assume that there exists a Lyapunov function V defined in a neighbourhood D of the origin in $\mathbb{R}^n$, such that for any $x \in D$, $x \neq 0$, there exists at least one $i \in \{1, \ldots, p\}$ such that $(Y_i V)(x) \neq 0$ (here, $Y_i V$ is the Lie derivative of the functional V in the direction of the vector field Y_i). Then the feedback law ϕ whose ith component $(1 \leq i \leq p)$ is the functional ϕ_i defined on D by

$$\phi_i(x) = \begin{cases} \left(\frac{-LV(x)}{|(Y_1V(x),\ldots,Y_pV(x))|} - 1\right) Y_iV(x) & \text{if } x \neq 0, \\ 0 & \text{if } x = 0 \end{cases}$$

is smooth for $x \neq 0$ and renders the stochastic differential system (6) locally asymptotically stable in probability. Indeed, denoting by $\mathcal{L}$ the infinitesimal generator associated with the resulting closed-loop system yields

$$\mathcal{L}V(x) = -\sum_{i=1}^{p} |Y_iV(x)|^2$$

for any $x \in D$, $x \neq 0$. So, according to Theorem 1.3 the equilibrium solution $x_t \equiv 0$ of the resulting closed-loop system is locally asymptotically stable in probability. This result extends a well-known result from [4] (see also [28]) to a stochastic differential system.

A simpler formula for the feedback law may be obtained if one assumes that there exists a Lyapunov function V defined in a neighborhood D of the origin such that $LV(x) \leq 0$ for all $x \in D$ and $LV(x) < 0$ for all $x \in D \setminus \{0\}$ such that $Y_iV(x) = 0$ for all $i \in \{1,\ldots,p\}$. Then, the feedback law $\phi = -(Y_1V,\ldots,Y_pV)^\star$ (where $\star$ denotes the transpose of matrices) renders the stochastic differential system (6) locally asymptotically stable in probability at the origin.

Denoting by $\mathcal{L}$ the infinitesimal generator of the resulting closed-loop system yields

$$\mathcal{L}V(x) = LV(x) - \sum_{i=1}^{p} |Y_iV(x)|^2$$

for any $x \in D$. Therefore, $\mathcal{L}V(x) < 0$ for any $x \in D$, and according to Theorem 1.3 the equilibrium state of the closed-loop system is locally asymptotically stable in probability. A deterministic version of this result is due to Tsinias [27].

The aim of this paper is to extend the ideas developed in the above examples to compute stabilizing feedback laws for a larger class of stochastic differential systems.

3. Affine control stochastic differential systems. In order to state some sufficient conditions for the local asymptotic stability in probability of stochastic differential systems that are affine in the control, we introduce the following definition.

DEFINITION 3.1. 1) *The stochastic differential system* (6) *is said to satisfy a stochastic Lyapunov condition at the origin if there exists a Lyapunov function V defined in a neighbourhood D of the origin in $\mathbb{R}^n$ such that*

a. *For every $x \in D \setminus \{0\}$ such that $Y_iV(x) = 0$ for all $i \in \{1,\ldots,p\}$ one has $LV(x) < 0$.*

2) *The stochastic differential system* (6) *is said to satisfy a strong stochastic Lyapunov condition at the origin if there exists a Lyapunov function V defined in a neighbourhood D of the origin in $\mathbb{R}^n$ satisfying condition* 1 *above and real functions a and b defined in D such that*:

b. *The function a is smooth and nonnegative on D.*

c. *The function b is continuous and nonnegative on D.*

d. *The inequality*

$$|LV(x) + a(x)| \leq b(x)\ |(Y_1 V(x), \dots, Y_p V(x))|$$

holds for every $x \in D$.

Remark 3.2. Since the differential operator L appears in the conditions stated in Definition 3.1, the computations in the stochastic case are more tedious than in the deterministic one.

Then, one can prove the following result, which extends Artstein's theorem (see [3], [24]) to the feedback stabilization of stochastic differential systems.

THEOREM 3.3. *If the stochastic differential system* (6) *satisfies a stochastic Lyapunov condition at the origin, then it is locally feedback stabilizable in probability at the origin by means of a feedback law that is smooth in a neighbourhood D of the origin except possibly in* 0. *Moreover, if the stochastic differential system* (6) *satisfies a strong stochastic Lyapunov condition at the origin, then the stabilizing feedback law is bounded on D and is continuous at the origin if* $b(0) = 0$.

Proof of Theorem 3.3. 1) Assume that the stochastic differential system (6) satisfies a stochastic Lyapunov condition at the origin. Then, applying in (6) the control law u defined by

$$u = (\nu Y_1 V, \dots, \nu Y_p V)^\star$$

where ν is a new real-valued control law, one gets the following control stochastic differential system:

$$x_t = x_0 + \int_0^t (X_0(x_s) + \nu\, Y(x_s))\, ds + \sum_{i=1}^m \int_0^t X_i(x_s) \circ dw_s^i \tag{7}$$

where Y denotes the vector field on $\mathbb{R}^n$ defined by

$$Y = \sum_{i=1}^p (Y_i V) Y_i.$$

Then, for any $x \in D \backslash \{0\}$ such that $(YV)(x) = 0$, one can deduce easily from condition 1 in Definition 3.1 that $LV(x) < 0$. Therefore, the stochastic differential system (7) satisfies a stochastic Lyapunov condition at the origin.

To define a stabilizing feedback law for the stochastic differential system (7) we make use of the partition of unity theorem. With this aim, denote by C_1 and C_2 the two relatively closed subsets of $D \setminus \{0\}$ defined by

$$C_1 = \{x \in D \setminus \{0\} : YV(x) = 0\}$$

and

$$C_2 = \{x \in D \setminus \{0\} : LV(x) \geq 0\}.$$

Then, $C_1 \cap C_2 = \emptyset$ and there exists a C^∞ function ψ mapping $D \setminus \{0\}$ into $[0, 1]$ such that $\psi \equiv 0$ on a neighbourhood of C_1 in $D \setminus \{0\}$ and $\psi \equiv 1$ on C_2.

Consider the real-valued function H defined on D by

$$H(x) = \begin{cases} -\psi(x) \left(\dfrac{LV + a}{YV} \right)(x) - 1 & \text{if } x \in D \setminus \{0\} \text{ s.t. } (YV)(x) \neq 0, \\ 0 & \text{if } x \in \{0\} \cup C_1 \end{cases}$$

where a is any smooth and positive real-valued function defined on D.

Then, applying the feedback law $\nu = H$ in (7) yields

$$x_t = x_0 + \int_0^t (X_0(x_s) + H(x_s)Y(x_s))\, ds + \sum_{i=1}^{m} \int_0^t X_i(x_s) \circ dw_s^i. \tag{8}$$

Moreover, one can easily prove that the feedback law u given by

$$u = (H.(Y_1V), \ldots, H.(Y_pV))^\star \tag{9}$$

is smooth on $D \setminus \{0\}$, and denoting by $\mathcal{L}$ the infinitesimal generator associated with the stochastic differential system (8) yields

$$\mathcal{L}V = (1 - \psi)\ LV - YV - \psi a.$$

Hence, $\mathcal{L}V(x) < 0$ for any $x \in D\setminus\{0\}$ and, according to Theorem 1.3, the equilibrium solution $x_t \equiv 0$ of the stochastic differential system (8) is locally asymptotically stable in probability. Therefore, the control law u given by (9) is a stabilizing feedback law for the stochastic differential system (6).

2) Assume that the stochastic differential system (6) satisfies a strong stochastic Lyapunov condition at the origin. Then, one can construct a stabilizing feedback law for the stochastic differential system (6) as above where the function a in the definition of the functional H is given by assumption 2 in Definition 3.1.

On the other hand, by condition d in Definition 3.1, for any $x \in D$, it holds that

$$|LV(x) + a(x)| \leq b(x)\ |(YV)(x)|^{1/2}$$

and, since $|\psi| \leq 1$ and $|Y_iV| \leq (YV)^{1/2}$ for any $i \in \{1, \ldots, p\}$, one has

$$|H(Y_iV)| \leq |b| + |Y_iV|$$

for any $i \in \{1, \ldots, p\}$.

Therefore, by condition c in Definition 3.1, one can deduce easily that the stabilizing feedback law u defined in (9) is bounded on D.

Furthermore, if $b(0) = 0$, it is easy to prove that the feedback law u is continuous at the origin.

This completes the proof of Theorem 3.3.

Remark 3.4. The stabilizing feedback law obtained in Theorem 3.3 depends explicitly on the system coefficients and the Lyapunov function given by the hypothesis. However, the presence of the function ψ, obtained by means of the partition of unity theorem, in formula (9) makes the result rely on nonconstructive techniques.

Note that under a slightly different hypothesis (the small control property introduced by Sontag in [24]) one can obtain a more easily computable stabilizing feedback law for the stochastic differential system (6) (see [6]).

To conclude, note that a serious drawback of both approaches is the lack of information about the construction of the Lyapunov function V in the assumptions of the theorem.

To illustrate the results stated in Theorem 3.3, consider the following two examples.

Example 1. Let x_0 be given in $\mathbb{R}^3$ and denote by $x_t \in \mathbb{R}^3$ the solution of the stochastic differential system

$$dx_t = \begin{pmatrix} \frac{1}{2}x_{1,t} \\ \frac{1}{2}x_{2,t} \\ -\frac{9}{2}x_{3,t} \end{pmatrix} dt + u_1 \begin{pmatrix} x_{2,t}^2 \\ 0 \\ -x_{2,t}x_{3,t} \end{pmatrix} dt + u_2 \begin{pmatrix} x_{1,t}^2 \\ -x_{2,t}^2 \\ 0 \end{pmatrix} dt + \begin{pmatrix} x_{1,t} \\ x_{2,t} \\ x_{3,t} \end{pmatrix} \circ dw_t.$$

Then, using the Lyapunov functional V defined on $\mathbb{R}^3$ by

$$V(x) = \frac{1}{2}(x_1^2 + x_2^2 + x_3^2),$$

one can prove that the above stochastic differential system satisfies a stochastic Lyapunov condition at the origin and thus by Theorem 3.3 is feedback stabilizable at the origin by means of a feedback law that is smooth in a neighbourhood of the origin except possibly in zero.

Example 2. Consider the stochastic differential system defined in $\mathbb{R}^3$ by

$$dx_t = \begin{pmatrix} -2x_{1,t} \\ -2x_{2,t} \\ -\frac{1}{2}x_{1,t}x_{3,t} + x_{1,t}^2x_{3,t} \end{pmatrix} dt + u \begin{pmatrix} x_{2,t} \\ -x_{1,t} \\ x_{1,t}^2 + x_{3,t}^2 \end{pmatrix} dt + \begin{pmatrix} x_{1,t} \\ x_{2,t} \\ x_{1,t}x_{3,t} \end{pmatrix} \circ dw_t$$

where x_0 is given in $\mathbb{R}^3$. Then, denoting by V the Lyapunov functional defined on $\mathbb{R}^3$ by

$$V(x) = \frac{1}{2}(x_1^2 + x_2^2 + x_3^2)$$

and by a, b the two functionals mapping $\mathbb{R}^3$ into $\mathbb{R}$ defined by

$$a(x) = x_1^2 + x_2^2,$$

$$b(x) = |x_1|,$$

one can prove that the stochastic differential system introduced above satisfies a strong stochastic Lyapunov condition at the origin.

Therefore, by Theorem 3.3 this stochastic differential system is feedback stabilizable in probability at the origin by means of a feedback law that is smooth in a neighbourhood of the origin except possibly in zero and is continuous at the origin.

The method used in the proof of Theorem 3.3 leads to the following result that extends the well-known theorem of Jurdjevic and Quinn [17] for the stabilization of deterministic control systems to stochastic differential system.

DEFINITION 3.5. *The stochastic differential system* (6) *is said to satisfy a stochastic Jurdjevic–Quinn condition at the origin if there exists a Lyapunov function V defined in a neighbourhood D of the origin in $\mathbb{R}^n$ such that*:

1) *For every $x \in D \setminus \{0\}$ such that $Y_iV(x) = 0$ for all $i \in \{1, \dots, p\}$ one has $LV(z) \leq 0$ for all z in some neighbourhood D_x of x in $\mathbb{R}^n$.*

2) *For every $x \in D \setminus \{0\}$ such that $LV(x) = 0$ and $Y_iV(x) = 0$ for all $i \in \{1, \ldots, p\}$ there exist integers $k \in \mathbb{N}^\star$ and $i \in \{1, \ldots, p\}$ such that $(L^kY_i)V(x) \neq 0$.*

Remark 3.6. Condition 2) in Definition 3.5 is not as easy to check as in the deterministic case since one has to compute iterates of the differential operator L.

Then one can prove the following stabilization result.

THEOREM 3.7. *If the stochastic differential system* (6) *satisfies a stochastic Jurdjevic–Quinn condition at the origin, then it is locally feedback stabilizable in probability at the origin by means of a feedback law that is smooth in a neighbourhood of the origin except possibly at zero.*

Proof of Theorem 3.7. Under the hypothesis of the theorem, one can construct, as in the proof of Theorem 3.3, a function ψ mapping $D \setminus \{0\}$ into $[0, 1]$ except that now C_2 will denote the closure relatively to $D \setminus \{0\}$ of the set

$$\{x \in D \setminus \{0\} : LV(x) > 0\}.$$

Here, one has $C_1 \cap C_2 = \emptyset$ by condition 1) in Definition 3.5.

Therefore, the feedback law u given by

$$u = (H(Y_1V), \ldots, H(Y_pV))^\star \tag{10}$$

where the function H is defined as in the proof of Theorem 3.3, is smooth on $D \setminus \{0\}$, and denoting by $\mathcal{L}$ the infinitesimal generator of the closed-loop system deduced from (6) when the control law u is given by (10) yields

$$\mathcal{L}V = (1 - \psi)\, LV - YV - \psi a$$

where Y is the vector field defined on $\mathbb{R}^n$ by

$$Y = \sum_{i=1}^{p} (Y_iV)Y_i.$$

Hence, $\mathcal{L}V(x) \leq 0$ for any $x \in D \setminus \{0\}$, and according to Theorem 1.3, the equilibrium solution $x_t \equiv 0$ of the closed-loop system deduced from (6) when the control law u is given by (10) is stable in probability.

Moreover, according to the stochastic version of La Salle's theorem (see [19]), the stochastic process x_t tends in probability to the largest positively invariant set whose support is contained in the locus $\mathcal{L}V(x_t) = 0$ for all $t \in \mathbb{R}_+$. On the other hand, it is obvious that $\mathcal{L}V(x) = 0$ for an $x \neq 0$ if and only if $LV(x) = 0$ and $Y_iV(x) = 0$ for all $i \in \{1, \ldots, p\}$.

Then the successive application of Itô's formula yields $L^kY_iV(x) = 0$ for $i = 1, \ldots, p$, and $k \in \mathbb{N}$, which contradicts the second condition in Definition 3.5.

Thus, the equilibrium solution $x_t \equiv 0$ of the closed-loop system deduced from (6) when the control law u is given by (10) is asymptotically stable in probability. Therefore, the control law u given by (10) is a stabilizing feedback law for the stochastic differential system (6).

This completes the proof of Theorem 3.7.

Remark 3.8. If instead of condition 1) in Definition 3.5 one assumes that $LV(x) \leq 0$ for all $x \in D$, then one can choose $\psi \equiv 0$ and the stabilizing feedback law reads

$$u = (-Y_1V, \ldots, -Y_pV)^\star,$$

which is the same stabilizing formula as the one proposed in [17] for the stabilization of deterministic control systems (see also [8]). In the latter case, note that the equilibrium solution of the uncontrolled part of the stochastic differential system (6) is stable in probability.

As has been noted previously, a serious drawback for practical applications of Theorem 3.7 is the presence of the functional ψ, obtained by means of a nonconstructive technique, in the design of the stabilizing feedback law (10). Successive application of Itô's formula to the conditions deduced from the stochastic version of La Salle's theorem leads to $L^{k+1}V(x) = 0$ and $L^kY_iV(x) = 0$ for $i = 1, \dots, p$ and $k \in \mathbb{N}$. Hence, one can improve condition 2) in Definition 3.5 by assuming that the set

$$\mathcal{K} = \{x \in \mathbb{R}^n / L^{k+1}V(x) = L^kY_jV(x) = 0 \ ; \ k \in \mathbb{N} \ ; \ j = 1, \dots, p\}$$

is reduced to $\{0\}$ (see [8]).

Example 3. Let x_0 be given in $\mathbb{R}^2$, and denote by $x_t \in \mathbb{R}^2$ the solution of the stochastic differential system

$$dx_t = \begin{pmatrix} -x_{1,t} - x_{2,t} \\ \\ x_{1,t} - x_{2,t} + x_{1,t}^2\phi(x_{1,t}, x_{2,t}) \end{pmatrix} dt + u \begin{pmatrix} x_{1,t}^2 \\ \\ 0 \end{pmatrix} dt + \begin{pmatrix} x_{1,t} \\ \\ x_{2,t} \end{pmatrix} \circ dw_t$$

where ϕ is a smooth functional mapping $\mathbb{R}^2$ into $\mathbb{R}$ such that $x_2\phi(0, x_2) < 0$ for any $x_2 \in \mathbb{R}$, $x_2 \neq 0$. Then this stochastic differential system satisfies a stochastic Jurdjevic–Quinn condition at the origin with the Lyapunov function V defined on $\mathbb{R}^2$ by

$$V(x) = \frac{1}{2}(x_1^2 + x_2^2).$$

Indeed, if Y denotes the vector field defined on $\mathbb{R}^2$ by $Y(x) = \left(\begin{smallmatrix} x_1^2 \\ 0 \end{smallmatrix}\right)$, one has $LV(0, x_2) = 0$ for all $x \in \mathbb{R}^2$, $x \neq 0$, with $YV(x) = x_1^3 = 0$. Therefore, there exists a neighbourhood D_x of the point x such that $LV(z) = z_1^2z_2\phi(z_1, z_2) \leq 0$ for all $z \in D_x$. Furthermore, by means of easy computations, one can prove that $L^3YV(0, x_2) = -6x_2^3 \neq 0$ if $x_2 \neq 0$.

On the other hand, for some particular stochastic control systems, one can state, by means of stochastic Lyapunov-like approaches different from those of Theorems 3.3 and 3.7, more easily computable assumptions on the system coefficients that lead to the existence of stabilizing feedback laws.

For instance, consider the stochastic process solution $x_t \in \mathbb{R}^n$ of the single-input stochastic bilinear differential system

$$x_t = x_0 + \int_0^t (Ax_s + uBx_s)\, ds + \sum_{i=1}^{m} \int_0^t C_ix_s\, dw_s^i \tag{11}$$

where A, B, C_i, $1 \leq i \leq m$, are matrices in $\mathcal{M}_{n\times n}(\mathbb{R})$ and assume that (11) satisfies a stochastic Lyapunov condition at the origin where the Lyapunov function is quadratic. Then the stochastic Lyapunov condition at the origin leads to the following result.

PROPOSITION 3.9 (see also [9]). *If there exists a symmetric and positive definite matrix P in $\mathcal{M}_{n\times n}(\mathbb{R})$ such that*

$$\ker(PB + B^\star P) \setminus \{0\} \subset \{x \in \mathbb{R}^n \ / \ \langle \bar{P}x, x\rangle \ < 0\}$$

where $\bar{P}$ is the matrix given by $\bar{P} = A^\star P + PA + C^\star PC$, then there exists a positive constant c such that the stochastic differential system (11) *is feedback stabilizable in probability at the origin by means of the feedback law u defined by*

$$u(x) = \begin{cases} -c \, \frac{\langle PBx, x\rangle}{|x|^2} & \text{if } x \neq 0, \\ 0 & \text{if } x = 0, \end{cases} \tag{12}$$

which is smooth for $x \neq 0$ and bounded.

Proof of Proposition 3.9. Since the matrix P is symmetric and positive definite, the function V defined on $\mathbb{R}^n$ by

$$V(x) = \langle Px, x\rangle$$

is a Lyapunov function. Denoting by L the infinitesimal generator of the closed-loop system deduced from (11) when the control law u is given by (12), one has for any $x \in \mathbb{R}^n$,

$$LV(x) = \langle \bar{P}x, x\rangle - \frac{c}{||x||^2}\langle PBx, x\rangle^2. \tag{13}$$

Therefore, since for any $x \in \mathbb{R}^n \setminus \{0\}$ there exists a unique pair $(r, z) \in \mathbb{R}_+ \times S^{n-1}$ (where S^{n-1} denotes the unit sphere in $\mathbb{R}^n$) such that $x = rz$, equality (13) reads

$$LV(x) = r^2 \left(\langle \bar{P}z, z\rangle - c\langle PBz, z\rangle^2\right). \tag{14}$$

Hence, for any $x \in \mathbb{R}^n$, $x \neq 0$, one has

$$LV(x) < 0$$

provided that

$$c > \frac{\max_{z \in S^{n-1}} |\langle \bar{P}z, z\rangle|}{\min_{z \in S^{n-1}, \langle \bar{P}z, z\rangle \geq 0} \langle PBz, z\rangle^2}.$$

Then, according to Theorem 1.3, the equilibrium solution $x_t \equiv 0$ of the closed-loop system deduced from (11) when the control law u is given by (12) is asymptotically stable in probability.

This completes the proof of Proposition 3.9.

Remark 3.10. The matrix $\bar{P}$ in Proposition 3.9 appears in the stochastic Lyapunov equation that gives a necessary and sufficient condition for the exponential stability in mean square of the equilibrium solution of a linear stochastic differential equation (see Arnold [2]). Conditions for the existence and the uniqueness of the solution to the stochastic Lyapunov equation are given by Wonham [30].

For more general stochastic differential systems that are not necessarily linear in the control we are unable at this time to formulate a proper Lyapunov condition and to apply one of the above methods to achieve the feedback stabilization. Only some specific cases can be solved properly (see [9] for stochastic bilinear systems or [10] for homogeneous stochastic systems). In the following, we propose a slight extension of the stochastic version of Jurdjevic–Quinn theorem to a class of stochastic differential system given by (3) that are not linear in the control.

PROPOSITION 3.11. *Assume that there exists a Lyapunov function V defined on a neighbourhood D of the origin in $\mathbb{R}^n$ such that $LV(x) \leq 0$ for $x \in D$ where L denotes the second-order differential operator given by*

$$L = X_0 + \frac{1}{2}\sum_{i=1}^{m} X_i^2$$

and X_0, X_i, $1 \leq i \leq m$, are the vector fields on $\mathbb{R}^n$ defined for any $x \in \mathbb{R}^n$ as

$$X_0(x) = \sum_{j=1}^{n}\left(f_j(x,0) + \sum_{k=1}^{n} g_{ki}(x)\frac{\partial g_{ji}(x)}{\partial x_k}\right)\frac{\partial}{\partial x_j}$$

and

$$X_i(x) = \sum_{j=1}^{n} g_{ji}(x)\frac{\partial}{\partial x_j}.$$

For any $i \in \{1,\dots,p\}$, let $Y_i(x) = (\frac{\partial f}{\partial u_i})(x,0)$; and assume that condition 2 *of Definition* 3.5 *is fulfilled. Then the stochastic differential system* (3) *is locally feedback stabilizable in probability at the origin by means of the smooth control law u defined on $\mathbb{R}^n$ by*

$$u_i(x) = -Y_iV(x),\ i = 1,\dots,p. \tag{15}$$

Proof of Proposition 3.11. The hypothesis on the system coefficients implies that there exist a function R mapping $\mathbb{R}^n \times \mathbb{R}^p$ into $\mathbb{R}^n$ and a positive constant C such that one can write the stochastic differential system (3) as

$$x_t = x_0 + \int_0^t \left(X_0(x_s) + \sum_{j=1}^{p} u^j Y_j(x_s) + R(x_s,u)\right) ds + \sum_{i=1}^{m}\int_0^t X_i(x_s)\circ dw_s^i \tag{16}$$

where $||R(x,u)|| \leq C||u||^2$ for any (x,u) in a neighbourhood of the origin in $\mathbb{R}^n \times \mathbb{R}^p$. Then, applying the feedback law u given by (15) in equation (16) yields

$$\begin{aligned} x_t = x_0 &+ \int_0^t \left(X_0(x_s) - \sum_{j=1}^{p}(Y_jV)(x_s)Y_j(x_s)\right) ds \\ &+ \int_0^t R(x_s,(Y_1V)(x_s),\dots,(Y_pV)(x_s))ds + \sum_{i=1}^{m}\int_0^t X_i(x_s)\circ dw_s^i. \end{aligned} \tag{17}$$

Denoting by $\mathcal{L}$ the infinitesimal generator of the closed-loop system (17), one has for any $x \in D$, $x \neq 0$,

$$\mathcal{L}V(x) = LV(x) - \sum_{j=1}^{p}((Y_jV)(x))^2(1 + O(x))$$

where

$$O(x) = -\frac{\nabla V(x)R(x,(Y_1V)(x),\dots,(Y_pV)(x))}{\sum_{j=1}^{p}((Y_jV)(x))^2}$$

and $O(x)$ tends to zero when x tends to zero.

Therefore, according to the hypothesis on the function V, one can deduce that

$$\mathcal{L}V(x) \leq 0$$

for any $x \in D$, $x \neq 0$. Furthermore, one has $\mathcal{L}V(x) = 0$ for any $x \in D$, $x \neq 0$, if and only if $LV(x) = 0$ and $(Y_j V)(x) = 0$ for all $j \in \{1, \ldots, p\}$. Then, arguing as in the proof of Theorem 3.7, one can prove that the equilibrium solution $x_t \equiv 0$ of the closed-loop, system (17) is asymptotically stable in probability. Therefore, the control law u given by (15) is a stabilizing feedback law for the stochastic differential system (16). This completes the proof of Proposition 3.11.

4. Stochastic dynamic stabilization. The purpose of this section is to study the stabilization of the nonlinear stochastic differential system (3) by means of a dynamic feedback control law. With this aim, one has to extend the concept of a stochastic Lyapunov condition introduced in Definition 3.1 as follows.

DEFINITION 4.1. *The stochastic differential system* (3) *is said to satisfy a dynamic stochastic Lyapunov condition at the origin if there exists a Lyapunov functional V defined on a neighbourhhood D of the origin in $\mathbb{R}^n \times \mathbb{R}^p$ such that:*

1) *The function V is smooth on $D \setminus \{0\}$.*

2) *For any $(x,z) \in D \setminus \{0\}$ such that $\frac{\partial V}{\partial z}(x,z) = 0$ one has*

$$\sum_{i=1}^{n} f_i(x,z) \frac{\partial V}{\partial x_i}(x,z) + \frac{1}{2} \sum_{i,j=1}^{n} (g(x)g(x)^{\star})_{ij} \frac{\partial^2 V}{\partial x_i \partial x_j}(x,z) < 0.$$

Note that according to Definitions 3.1 and 4.1 the stochastic differential system (3) satisfies a dynamic stochastic Lyapunov condition at the origin if and only if the stochastic differential system (5) satisfies a stochastic Lyapunov condition at the origin.

The following result underlines the relationship between stabilization and dynamic stabilization of the stochastic differential system (3) and the dynamic stochastic Lyapunov condition.

THEOREM 4.2. 1) *If the stochastic differential system* (3) *satifies a dynamic stochastic Lyapunov condition at the origin, then it is dynamic asymptotically stabilizable in probability at the origin.*

2) *If the stochastic differential system* (3) *satisfies a dynamic stochastic Lyapunov condition at the origin and the equation $\frac{\partial V}{\partial z}(x,z) = 0$ has a solution $z = \phi(x)$, $\phi(0) = 0$, that is continuous on a neighbourhood S of the origin in $\mathbb{R}^n$, smooth on $S \setminus \{0\}$, and such that $\frac{\partial^2 V}{\partial x \partial z}(x, \phi(x)) = 0$ for any $x \in S \setminus \{0\}$, then the stochastic differential system* (3) *is locally feedback stabilizable in probability at the origin by means of the control law $u = \phi(x)$.*

3) *If the stochastic differential system* (3) *is locally feedback stabilizable in probability at the origin by means of a locally smooth feedback law $u = \phi(x)$ with $\phi(0) = 0$, then it satisfies a dynamic stochastic Lyapunov condition at the origin with V smooth in a neighbourhood of the origin.*

4) *If the stochastic differential system* (3) *satisfies a dynamic stochastic Lyapunov condition at the origin with V smooth in a neighbourhood S of the origin such that* $\det \frac{\partial^2 V}{\partial z^2}(0,0) \neq 0$ *and $\frac{\partial^2 V}{\partial x \partial z}(x,z) = 0$ for any $(x,z) \in S$ with $\frac{\partial V}{\partial z}(x,z) = 0$, then it is locally feedback stabilizable in probability at the origin by means of a locally smooth feedback law $u = \phi(x)$ with $\phi(0) = 0$.*

Proof of Theorem 4.2. 1) The first assertion of the theorem is an immediate consequence of Theorem 3.3. Indeed, if the stochastic differential system (3) satisfies a dynamic stochastic Lyapunov condition at the origin, then the stochastic differential system (5) satisfies a stochastic Lyapunov condition at the origin. Hence, according to Theorem 3.3, there exists a stabilizing feedback law $v = r(x,z)$ for the stochastic differential system (5) that is smooth for $(x,z) \neq 0$ in a neighbourhood of the origin in $\mathbb{R}^n \times \mathbb{R}^p$.

2) Let ϕ be a function mapping $\mathbb{R}^n$ into $\mathbb{R}^p$ satisfying the hypothesis of assertion 2) in Theorem 4.2. Then the closed-loop system deduced from the control stochastic differential system (3) when the control law u is given by $u(x) = \phi(x)$ reads

$$x_t = x_0 + \int_0^t f(x_s, \phi(x_s))ds + \int_0^t g(x_s)\ dw_s. \tag{18}$$

Denote by $\hat{V}$ the function mapping S into $\mathbb{R}$ defined for any $x \in S$ by

$$\hat{V}(x) = V(x, \phi(x))$$

where V is the Lyapunov function given by Definition 4.1. Then $\hat{V}$ is a Lyapunov function; and denoting by $\mathcal{L}$ the infinitesimal generator of the stochastic process x_t solution of the closed-loop system (18), one has

$$\begin{aligned}\mathcal{L}\hat{V}(x) &= \nabla V(x,\phi(x))f(x,\phi(x)) + \frac{1}{2}\mathrm{Tr}\left[g(x)g(x)^\star \nabla^2 V(x,\phi(x))\right] \\ &= \left(\frac{\partial V}{\partial x}(x,\phi(x)) + \frac{\partial V}{\partial z}(x,\phi(x))\nabla\phi(x)\right) f(x,\phi(x)) \\ &\quad + \frac{1}{2}\mathrm{Tr}\left(g(x)g(x)^\star\left(\frac{\partial^2 V}{\partial x^2}(x,\phi(x)) + 2\frac{\partial^2 V}{\partial x \partial z}(x,\phi(x))\nabla\phi(x)\right.\right. \\ &\quad \left.\left. + \frac{\partial^2 V}{\partial z^2}(x,\phi(x))(\nabla\phi(x))^2 + \frac{\partial V}{\partial z}(x,\phi(x))\nabla^2\phi(x)\right)\right)\end{aligned}$$

for any $x \in S$, $x \neq 0$.

Then since, by assumptions, one has $\frac{\partial V}{\partial z}(x,\phi(x)) = 0$ and $\frac{\partial^2 V}{\partial x \partial z}(x,\phi(x)) = 0$ for any $x \in S$, $x \neq 0$, one gets

$$\mathcal{L}\hat{V}(x) = \sum_{i=1}^n f_i(x,\phi(x))\frac{\partial V}{\partial x_i}(x,\phi(x)) + \frac{1}{2}\sum_{i,j=1}^n (g(x)g(x)^\star)_{ij}\frac{\partial^2 V}{\partial x_i \partial x_j}(x,\phi(x)).$$

Hence, condition 2) in Definition 4.1 yields

$$\mathcal{L}\hat{V}(x) < 0$$

for any $x \in S$, $x \neq 0$.

Therefore, according to Theorem 1.3 the equilibrium solution $x_t \equiv 0$ of the closed-loop system (18) is locally asymptotically stable in probability at the origin, which implies that the control law u given by $u(x) = \phi(x)$ is a stabilizing feedback law for the stochastic differential system (3).

3) Let ϕ be a smooth functional defined on a neighbourhood S of the origin in $\mathbb{R}^n$ with values in $\mathbb{R}^p$ such that $\phi(0) = 0$ and the equilibrium solution $x_t \equiv 0$ of the closed-loop system

$$x_t = x_0 + \int_0^t f(x_s, \phi(x_s))ds + \int_0^t g(x_s)dw_s \tag{19}$$

deduced from the stochastic differential system (3) when the control law u is given by $u(x) = \phi(x)$ is locally asymptotically stable in probability. Then, by the converse Lyapunov theorem proved by Kushner [20], there exists a Lyapunov function V defined on S such that

$$\sum_{i=1}^{n} f_i(x,\phi(x))\frac{\partial V}{\partial x_i}(x) + \frac{1}{2}\sum_{i,j=1}^{n} (g(x)g(x)^\star)_{ij}\frac{\partial^2 V}{\partial x_i \partial x_j}(x) < 0 \tag{20}$$

for any $x \in S$, $x \neq 0$.

Define the function $\tilde{V}$ mapping $S \times \mathbb{R}^p$ into $\mathbb{R}$ by

$$\tilde{V}(x,z) = V(x) + \frac{1}{4}||z - \phi(x)||^4. \tag{21}$$

Then, $\tilde{V}$ is smooth and positive definite on a neighbourhood D of the origin in $\mathbb{R}^n \times \mathbb{R}^p$. Furthermore, for any $(x,z) \in D \setminus \{0\}$ such that $\frac{\partial \tilde{V}}{\partial z}(x,z) = 0$, one has $z = \phi(x)$ and since (20) holds, yields

$$\sum_{i=1}^{n} f_i(x,\phi(x))\frac{\partial \tilde{V}}{\partial x_i}(x) + \frac{1}{2}\sum_{i,j=1}^{n} (g(x)g(x)^\star)_{ij}\frac{\partial^2 \tilde{V}}{\partial x_i \partial x_j}(x) < 0$$

for any $x \in S$, $x \neq 0$.

Hence, the stochastic differential system (3) satisfies a dynamic stochastic Lyapunov condition at the origin. Note that for any $x \in S$, $x \neq 0$, one has $\frac{\partial^2 \tilde{V}}{\partial x \partial z}(x,\phi(x)) = 0$.

4) If the stochastic differential system (3) satisfies a dynamic stochastic Lyapunov condition at the origin with V smooth and such that det $\frac{\partial^2 V}{\partial z^2}(0,0) \neq 0$ one can deduce, applying the implicit function theorem, that there exists a smooth functional ϕ defined on a neighbourhood S of the origin in $\mathbb{R}^n$ with values in $\mathbb{R}^p$ such that $\phi(0) = 0$ and $\frac{\partial V}{\partial x}(x,\phi(x)) = 0$ for any $x \in S$.

Moreover, since by assumptions one has $\frac{\partial^2 V}{\partial x \partial z}(x,\phi(x)) = 0$ for any $x \in S$, one can deduce by application of assertion 2) in Theorem 4.2 that the stochastic differential system (3) is locally feedback stabilizable in probability by means of the control law u defined for any $x \in S$ by $u(x) = \phi(x)$.

This completes the proof of Theorem 4.2.

Remark 4.3. In assertion 2) in Theorem 4.2 one has to assume conditions on the second derivative of the Lyapunov function V which are not needed in the deterministic case. The hypotheses stated in assertion 3) in Theorem 4.2 do not lead, as in the deterministic case, to det $\frac{\partial^2 V}{\partial z^2}(0,0) \neq 0$. Therefore, it seems that assertion (c) in Theorem 3 from [28] cannot be easily generalized to stochastic differential systems.

Example 4. Let x_0 be given in $\mathbb{R}$ and denote by $x_t \in \mathbb{R}$ the solution of the stochastic differential system

$$x_t = x_0 + \int_0^t (-x_s^2 - x_s^3 + u^4 - u^8)\, ds + \int_0^t x_s^3\, dw_s. \tag{22}$$

Then the feedback law u defined by $u(x) = ((1+\sqrt{1-4x^2})/2)^{1/4}$ which is smooth for $x \neq 0$, x in a neighbourhood of 0, and continuous at the origin renders the stochastic differential system (22) asymptotically stable in probability. Indeed, the equilibrium solution of the resulting closed-loop system

$$x_t = x_0 - \int_0^t x_s^3\, ds + \int_0^t x_s^3\, dw_s \tag{23}$$

is asymptotically stable in probability.

Furthermore, the stochastic differential system (22) can be dynamically asymptotically stabilized in probability. With this aim, one has to prove that the stochastic differential system (22) satisfies a dynamic stochastic Lyapunov condition at the origin or, equivalently, that the stochastic differential system

$$\begin{pmatrix} x_t \\ z_t \end{pmatrix} = \begin{pmatrix} x_0 \\ 0 \end{pmatrix} + \int_0^t \begin{pmatrix} -x_s^2 - x_s^3 + z_s^4 - z_s^8 \\ v \end{pmatrix} ds + \int_0^t \begin{pmatrix} x_s^3 \\ 0 \end{pmatrix} dw_s \tag{24}$$

satisfies a Lyapunov condition at the origin.

Let V be the functional mapping $\mathbb{R}^2$ into $\mathbb{R}$ defined by

$$V(x,z) = \begin{cases} x^2 + \left(z - (x^2 + z^8)^{1/4}\right)^4 & \text{if } (x,z) \neq 0, \\ 0 & \text{if } (x,z) = 0. \end{cases}$$

Then the function V is positive definite, smooth for any $(x,z) \neq 0$ on a neighbourhood S of the origin in $\mathbb{R}^2$, and continuous at 0. Moreover, for any $(x,z) \in S$, $(x,z) \neq 0$, such that $\frac{\partial V}{\partial z}(x,z) = 0$ one has $z^4 = x^2 + z^8$ and so

$$f(x,z)\frac{\partial V}{\partial z}(x,z) + \frac{1}{2}g(x)^2\frac{\partial^2 V}{\partial x^2}(x,z) = -x^4 + x^6 < 0,$$

which implies, according to Theorem 4.2, that the equilibrium solution $(x_t, z_t) = 0$ of the stochastic differential system (24) is asymptotically stable in probability.

Using a different stochastic Lyapunov approach from that of Theorems 3.3 and 4.2, one can prove the following result.

PROPOSITION 4.4. *Assume that the stochastic differential system* (3) *is asymptotically stable in probability by means of a smooth feedback law* $u(x) = \phi(x)$ *such that* $\phi(0) = 0$ (*or equivalently that condition* 4) *in Theorem* 4.2 *is fulfilled*). *Denote by* V *a Lyapunov function associated with the closed-loop system deduced from* (3) *when the control law* u *is given by* $u(x) = \phi(x)$, *and assume that there exist smooth functions* h_i, $1 \leq i \leq p$, *mapping* $\mathbb{R}^n \times \mathbb{R}^p \times \mathbb{R}^p$ *into* $\mathbb{R}$ *such that*

$$\begin{aligned} &\sum_{i=1}^n (f_i(x,z) - f_i(x,\phi(x)))\frac{\partial V}{\partial x_i}(x) + \sum_{i,j=1}^n (g(x)g(x)^\star)_{ij}\langle\frac{\partial \phi}{\partial x_i}(x), \frac{\partial \phi}{\partial x_j}(x)\rangle \\ &\qquad = \sum_{i=1}^p (z_i - \phi_i(x))h_i(x,\phi(x),z) \end{aligned} \tag{25}$$

for any x *in a neighbourhood of the origin in* $\mathbb{R}^n$. *Then the stochastic differential system* (3) *is dynamic asymptotically stable in probability by means of a smooth feedback law.*

Remark 4.5. The existence of a Lyapunov function V in the hypothesis of the theorem is given by the converse stochastic Lyapunov theorem proved by Kushner [20].

Proof of Proposition 4.4. Denote by $\bar{V}$ the function mapping $\mathbb{R}^n \times \mathbb{R}^p$ into $\mathbb{R}$ defined by

$$\bar{V}(x,z) = V(x) + \frac{1}{2}||z - \phi(x)||^2.$$

Then $\bar{V}$ is smooth and positive definite on a neighbourhood S of the origin on $\mathbb{R}^n \times \mathbb{R}^p$. Let r be the smooth feedback law mapping $\mathbb{R}^n \times \mathbb{R}^p$ into $\mathbb{R}^p$ defined by

$$\begin{aligned} r_i(x,z) = \sum_{j=1}^{n} f_j(x,z)\frac{\partial \phi_i}{\partial x_j}(x) + \sum_{j,k=1}^{n} (g(x)g(x)^\star)_{jk}\frac{\partial^2 \phi_i}{\partial x_j \partial x_k}(x) \\ -h_i(x,\phi(x),z) - (z_i - \phi_i(x)) \end{aligned} \tag{26}$$

for $i = 1,\dots,p$.

Denoting by $\mathcal{L}$ the infinitesimal generator associated with the closed-loop system deduced from (5) when the control law u is given by (26) one has

$$\begin{aligned} \mathcal{L}\bar{V}(x,z) = {} & \sum_{i=1}^{n} f_i(x,z)\frac{\partial V}{\partial x_i}(x) + \sum_{j,k=1}^{n} (g(x)g(x)^\star)_{jk}\frac{\partial^2 V}{\partial x_j \partial x_k}(x) \\ & + \sum_{j,k=1}^{n} (g(x)g(x)^\star)_{jk}\langle\frac{\partial \phi}{\partial x_j}(x), \frac{\partial \phi}{\partial x_k}(x)\rangle \\ & - \sum_{i=1}^{n}(z_i - \phi_i(x)) \left(\sum_{j=1}^{n} f_j(x,z)\frac{\partial \phi_i}{\partial x_j}(x) - r_i(x,z)\right) \\ & - \sum_{i=1}^{n}(z_i - \phi_i(x)) \left(\sum_{j,k=1}^{n} (g(x)g(x)^\star)_{jk}\frac{\partial^2 \phi}{\partial x_j \partial x_k}(x)\right). \end{aligned} \tag{27}$$

Then, denoting by $\mathcal{L}_\phi$ the infinitesimal generator associated with the closed-loop system deduced from (3) when the control law u is given by ϕ, one can deduce from (27) and (26) that

$$\mathcal{L}\bar{V}(x,z) = \mathcal{L}_\phi V(x,\phi(x)) - ||z - \phi(x)||^2 \tag{28}$$

for any $(x,z) \in \mathbb{R}^n \times \mathbb{R}^p$.

Thus, since V is a Lyapunov function associated with the closed-loop system deduced from (3) when the control law u is given by ϕ, one can deduce from (28) that for any $(x,z) \in \mathbb{R}^n \times \mathbb{R}^p$, one has

$$\mathcal{L}\bar{V}(x,z) < 0.$$

Therefore, according to Theorem 1.3, the equilibrium solution $(x_t, z_t) \equiv (0,0)$ of the stochastic differential system (5) is locally asymptotically stable in probability, which implies that the stochastic differential system (3) is dynamically asymptotically stable in probability.

This completes the proof of Proposition 4.4.

Remark 4.6. One has to assume that equality (25) is fulfilled in the hypotheses of Proposition 4.4. Indeed, equality (25) cannot be obtained by easy computations, as in its deterministic version computed in the proof of Theorem 4 in [28], since the second term in the left-hand side of this equality does not depend on $(x,z) \in \mathbb{R}^n \times \mathbb{R}^p$ but on $x \in \mathbb{R}^n$ only.

Example 5. Let x_0 be given in $\mathbb{R}^2$, and denote by $x_t \in \mathbb{R}^2$ the solution of the following composite stochastic differential system:

$$dx_t = \begin{pmatrix} x_{2,t} - \frac{1}{2}x_{1,t} \\ -x_{1,t} - \frac{1}{2}x_{2,t} \end{pmatrix} dt + \begin{pmatrix} 0 \\ r\left(x_{1,t}, x_{2,t}, u\right) \end{pmatrix} dt + \begin{pmatrix} x_{1,t} \\ x_{2,t} \end{pmatrix} dw_t \tag{29}$$

$$u = z \quad , \quad \dot{z} = v \tag{30}$$

where r is the functional defined on $\mathbb{R}^3$ by

$$r(x_1, x_2, z) = z(ax_1 + bx_2) + z^2\hat{r}(x_1, x_2, z), \tag{31}$$

$\hat{r}$ being a smooth functional defined on $\mathbb{R}^3$ and a, b real numbers such that $a^2 + b^2 \neq 0$.

Our aim is to prove that the composite stochastic differential system of (29) and (30) is asymptotically stabilizable in probability (or equivalently that the stochastic differential system (29) is dynamic asymptotically stabilizable in probability) by means of a smooth feedback law.

Denote by X_0, X_1, and Y the vector fields defined on $\mathbb{R}^2$ by

$$X_0(x) = \begin{pmatrix} x_2 \\ -x_1 \end{pmatrix}, \quad X_1(x) = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}, \quad Y(x) = \begin{pmatrix} 0 \\ ax_1 + bx_2 \end{pmatrix};$$

and let V be the Lyapunov function defined on $\mathbb{R}^2$ by

$$V(x) = \frac{1}{2}(x_1^2 + x_2^2).$$

Then, denoting by L the second-order differential operator defined by

$$L = X_0 + \frac{1}{2}X_1^2,$$

one can prove, by means of easy computations, that for any $x \in \mathbb{R}^2$, the following holds:

$$LV(x) = 0, \quad YV(x) = x_2(ax_1 + bx_2),$$

$$LYV(x) = a(x_2^2 - x_1^2) - 2bx_1x_2,$$

and

$$L^2YV(x) = -2b(x_2^2 - x_1^2) - 4ax_1x_2.$$

Therefore, since

$$\det \begin{pmatrix} a & -2b \\ -2b & -4a \end{pmatrix} = -4(a^2 + b^2) \neq 0,$$

one can deduce that for any $x \in \mathbb{R}^2$, $x \neq 0$, $LYV(x)$ and $L^2YV(x)$ are not both zero. Hence, by Proposition 3.11, the smooth feedback law $z = -YV(x) = -ax_1x_2 - bx_2^2$ renders the stochastic differential system (29) asymptotically stable in probability at the origin. Furthermore, applying Proposition 4.4, one can prove that the stochastic differential system (29) is dynamic asymptotically stable in probability at the origin by means of a smooth feedback law.

Remark 4.7. The result proved in Example 5 is an extension to stochastic differential systems of a stabilization result studied by Aeyels [1] for deterministic control systems.

REFERENCES

[1] D. Aeyels, *Stabilization of a class of nonlinear systems by a smooth feedback control*, Systems Control Lett., 5 (1985), pp. 289–294.

[2] L. Arnold, *Stochastic Differential Equations : Theory and Applications*, Wiley, New York, 1974.

[3] Z. Artstein, *Stabilization with relaxed controls*, Nonlinear Analysis Theory Methods Appl., 7 (1983), pp. 1163–1173.

[4] N. P. Bhatia and G. P. Szegö, *Stability Theory of Dynamical Systems*, Springer-Verlag, Berlin, Heidelberg, New York, 1970.

[5] R. W. Brockett, *Asymptotic stability and feedback stabilization*, in Differential Geometric Control Theory, R. W. Brockett, R. S. Millman, and H. J. Sussmann, eds., Birkhauser, Boston, 1983, pp. 181–191.

[6] P. Florchinger, *A universal formula for the stabilization of control stochastic differential equations*, Stochastic Anal. Appl., 11 (1993), pp. 155–162.

[7] ———, *Stabilization of control stochastic bilinear systems by linear feedback laws*, Proceedings of SINS '92, Dallas Fort Worth, TX, 1992.

[8] ———, *A stochastic version of Jurdjevic–Quinn theorem*, Stochastic Anal. Appl., 12 (1994), pp. 473–480.

[9] ———, *Feedback stabilization of stochastic bilinear systems and of some nonlinear stochastic systems*, Stochastic Anal. Appl., 12 (1994), pp. 527–542.

[10] ———, *On the stabilization of homogeneous control stochastic systems*, Proceedings of the 32nd IEEE Conference on Decision and Control, San Antonio, TX, 1993, pp. 855–856.

[11] P. Florchinger, A. Iggidr, and G. Sallet, *Stabilization of a class of nonlinear stochastic systems*, Stochastic Processes and Appl., 50 (1994), pp. 235–243.

[12] Z. Y. Gao and N. U. Ahmed, *Stabilizability of certain stochastic systems*, Internat. J. Systems Sci., 17 (1986), pp. 1175–1185.

[13] ———, *Feedback stabilizability of nonlinear stochastic systems with state–dependent noise*, International Journal of Control, 45 (1987), pp. 729–737.

[14] J. P. Gauthier, *Structure des systèmes non linéaires*, CNRS, 1984.

[15] H. Hermes, *On the synthesis of a stabilizing feedback control via Lie algebraic methods*, SIAM J. Control Optim., 18 (1980), pp. 352–361.

[16] N. Ikeda and S. Watanabe, *Stochastic differential equations and diffusion processes*, North-Holland Kodansha, Amsterdam, 1981.

[17] V. Jurdjevic and J. P. Quinn, *Controllability and stability*, J. Differential Equations, 28 (1978), pp. 381–389.

[18] R. Z. Khasminskii, *Stochastic stability of differential equations*, Sijthoff & Noordhoff, Alphen aan den Rijn, 1980.

[19] H. J. Kushner, *Stochastic stability*, in Stability of Stochastic Dynamical Systems, R. Curtain, ed., Lecture Notes in Math. 294, Springer-Verlag, Berlin, Heidelberg, New York, 1972, pp. 97–124.

[20] ———, *Converse theorems for stochastic Liapunov functions*, SIAM J. Control, 5 (1967), pp. 228–233.

[21] X. Mao, *Stability of stochastic differential equations with respect to semimartingales*, Pitman Res. Notes in Math. 251, Longman Scientific and Technical, Essex, 1991.

[22] E. D. Sontag, *Nonlinear regulation : the piecewise linear approach*, IEEE Trans. Automat. Control, 26 (1981), pp. 346–358.

[23] ———, *A Lyapunov-like characterization of asymptotic controllability*, SIAM J. Control Optim., 21 (1983), pp. 462–471.

[24] ———, *A universal construction of Artstein's theorem on nonlinear stabilization*, Systems and Control Lett., 13 (1989) pp. 117–123.

[25] ———, *Mathematical control theory*, Texts in Appl. Math., 6, Springer-Verlag, Berlin, Heidelberg, New York, 1990.

[26] E. D. Sontag and H. J. Sussmann, *Remarks on continuous feedback*, Proceedings of the 19th IEEE Conference on Decision and Control, Albuquerque, NM, 1980, pp. 916–921.

[27] J. Tsinias, *Stabilization of affine in control nonlinear systems*, Nonlinear Analysis, Theory, Methods Appl., 12 (1988), pp. 1283–1296.

[28] ———, *Sufficient Lyapunov-like conditions for stabilization*, Math. of Control Signals Systems, 2 (1989), pp. 343–357.

[29] J. L. Willems and J. C. Willems, *Feedback stabilizability for stochastic systems with state and control dependent noise*, Automatica, 12 (1976), pp. 277–283.

[30] W. M. Wonham, *On a matrix Riccati equation of stochastic control*, SIAM J. Control Optim., 6 (1968), pp. 681–697.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 33, No. 4, pp. 1170–1207, July 1995

© 1995 Society for Industrial and Applied Mathematics
012

ON FEEDBACK EQUIVALENCE OF A PARAMETERIZED FAMILY OF NONLINEAR SYSTEMS*

J.-B. POMET† AND I. A. K. KUPKA‡

Abstract. The following question is considered for a smoothly parameterized family of control systems: Does there exist a smooth family of transformations (feedback only, feedback plus diffeomorphism, or feedback plus diffeomorphism with some restrictions on the parameter dependence of the diffeomorphism, called "matching conditions") changing the systems of the family into a single one? Some abstract necessary and sufficient conditions are given, under which an explicit construction of the transformation is proposed. Both local and global results are obtained. No constant rank assumption is needed for the general conditions, but they can be translated more explicitly under such assumptions.

Key words. nonlinear systems, feedback equivalence, family of systems, nonlinear adaptive control

AMS subject classifications. 93B17, 93B29, 58A30

1. Introduction and problem statement.

We consider a family of nonlinear systems, indexed by a parameter vector $p = (p_1, \ldots, p_l) \in I\!R^l$. Although §6 briefly outlines the generalization of some results to nonaffine systems, most of this paper is devoted to systems that are affine in the control variables. The system $\mathcal{S}_p$ corresponding to a given value of p is therefore described by

$$\mathcal{S}_p\colon \quad \dot{x} = f_o(p,x) + g(p,x)\,u \tag{1}$$

$$\triangleq f_o(p,x) + \sum_{k=1}^{m} u_k\, f_k(p,x) \tag{2}$$

where x exists in an n-dimensional C^∞ manifold M^n, the input $u = (u_1, \ldots, u_m)$ is in $I\!R^m$, and the f_k's are $\mathcal{C}^\infty$ vector fields $\mathcal{C}^\infty$-ly depending on the parameter p.

The problem we are addressing here is finding, if possible, a family of transformations (feedback, feedback + diffeomorphism) that transforms the family of systems $(\mathcal{S}_p)$ into a family of systems that are all identical.

This implies in particular that any two systems of the family are equivalent via feedback or via feedback and diffeomorphism. This equivalence has been explored (see [3], [1], [2], and references therein) but as an equivalence between two systems, possibly on different manifolds, and one of the important questions is of course to find a set of invariants for this equivalence. Looking at the problem under the point of view of families of systems is rather new. Note that we ask the family of transformations to depend on the parameters as smoothly as the systems themselves do.

* Received by the editors January 2, 1992; accepted for publication (in revised form) February 3, 1994. This research was partially supported by Natural Sciences and Engineering Research Council of Canada grant OG0036498.

† Institut National de Recherche en Informatique et en Automatique, 2004 rte des lucioles, B.P. 93, 06902 Sophia Antipolis Cédex, France (pomet@sophia.inria.fr). This research was done while this author was with the Department of Mathematics, University of Toronto, Toronto, Ontario, M5S 1A1 Canada, partially supported by a scholarship from the Institut National de Recherche en Informatique et en Automatique, 78153 Le Chesnay Cédex, France, and then with the Department of Mathematics and Statistics, Queen's University, Kingston, Ontario, K7L 3N6, Canada, partially supported by Natural Sciences and Engineering Research Council of Canada grant OGP0009224.

‡ Department of Mathematics, University of Toronto, Toronto, Ontario, M5S 1A1 Canada, (kupka@math.toronto.edu).

For the most general problem (feedback and diffeomorphism), we give some rather abstract infinitesimal necessary and sufficient conditions. We also consider a stronger type of equivalence, where some "matching conditions" on the diffeomorphisms are imposed. These matching conditions come from nonlinear adaptive control and are also related to the possibility of controlling the time-varying system that is obtained when p is a function of time with control laws designed for p constant. It turns out that a much more explicit characterization of this equivalence is possible since, basically, the partial differential equations that characterize the general problem degenerate here into algebraic linear equations. Actually two different kinds of matching conditions are considered. One leads to a naturally integrable set of conditions and can be thoroughly characterized in terms of some system of algebraic linear equations having some solutions; the other involves some integrability conditions, but these are automatically satisfied in small dimensions (at the most four controls). Finally, pure feedback equivalence is an even more restrictive equivalence, since no diffeomorphism is allowed (or one independent of the parameters), and it may be completely characterized.

The paper is organized as follows. Section 2 is devoted to some definitions, preliminary results, and notation. Section 3 deals with pure feedback equivalence. Necessary and sufficient conditions are given for both local and global equivalence. Section 4 deals with general feedback and diffeomorphism equivalence. Some abstract necessary and sufficient conditions are given, both for local and global equivalence, and it is explained how they translate into some systems of linear partial differential equations. Section 5 is devoted to feedback and diffeomorphism equivalence with some "matching conditions" on the family of diffeomorphisms. After giving some abstract necessary and sufficient conditions, both for the local and the global case, which are of the same kind as those for the general problem, we give an explicit characterization of these, at least for the "first type" of matching conditions considered. We find as a consequence of the present result some known characterizations for matching conditions in the case of linearizable systems ([4], [5]). The two different sets of matching conditions that are considered are similar in a lot of cases; an example is given on which we have evidence that they are not equivalent, that some integrability conditions are actually necessary, and in which it is illustrated how constructive our method is, all the transformations being explicitly written. Finally §6 presents an extension of our result to systems that are not affine in the control, and §7 gives a brief conclusion.

2. Families of systems: Definitions and basic remarks.

This section is devoted to some notation and basic preliminary results. The main notation is summed up in §2.7.

2.1. The product manifold $I\!R^l \times M^n$.

We will relate parameterized families of systems, or of vector fields, on M^n to some more ordinary systems or vector fields on the product manifold $I\!R^l \times M^n$.

$I\!R^l \times M^n$ being a product, we may define the natural projections π_1 and π_2 by

$$\pi_1(p,x) = p; \quad \pi_2(p,x) = x. \tag{3}$$

We may also write the tangent space at any point (p,x) as a product:

$$T_{(p,x)}\left(I\!R^l \times M^n\right) = T_p I\!R^l \times T_x M^n = I\!R^l \times T_x M^n, \tag{4}$$

which allows us to define the "vector" projections Π_1 and Π_2: for a vector field Y,

$$\Pi_1(Y) + \Pi_2(Y) = Y, \tag{5}$$

$$\Pi_1(Y)\,(p,x) \in T_p\mathbb{R}^l \times \{0\}, \tag{6}$$

$$\Pi_2(Y)\,(p,x) \in \{0\} \times T_xM^n. \tag{7}$$

We refer to $\Pi_1(Y)$ and $\Pi_2(Y)$ as the p-component and x-component of Y, respectively. Finally, we denote by $\frac{\partial}{\partial p_1}, \ldots, \frac{\partial}{\partial p_l}$ the natural coordinate vector fields in $\mathbb{R}^l$, and in $\mathbb{R}^l \times M^n$:

$$\frac{\partial}{\partial p_i}(p,x) = (e_i \ , \ 0) \tag{8}$$

for any (p,x), $\{e_1, \ldots, e_l\}$ being the canonical basis of the vector space $T_p\mathbb{R}^l = \mathbb{R}^l$.

2.2. Vector fields, systems. A $\mathcal{C}^\infty$ *parameterized family of vector fields* is an f that maps any $(p,x) \in \mathbb{R}^l \times M^n$ into a vector $f(p,x)$ of T_xM^n such that $f(p,x)$ is a $\mathcal{C}^\infty$ function of (p,x). In particular, for any p, f_p, also denoted by $f(p,.)$, and defined by

$$f_p(x) = f(p,x), \tag{9}$$

is a $\mathcal{C}^\infty$ vector field on M^n. A $\mathcal{C}^\infty$ *family of systems* (affine in the controls) $\mathcal{S}$ with m inputs is defined by $m+1$ $\mathcal{C}^\infty$ parameterized families of vector fields f_o, $f_1, \ldots, f_m$, and we simply write $\mathcal{S} = (f_o, f_1, \ldots, f_m)$. In this family, the control system that corresponds to the value p of the parameter is described by (2).

We say that a *vector field* on $\mathbb{R}^l \times M^n$ is *parameter preserving* if it has a zero p-component. We also say that a *control system* $(F_o, F_1, \ldots, F_m)$ on $\mathbb{R}^l \times M^n$ is *parameter preserving* if $F_o, F_1, \ldots, F_m$ are parameter-preserving vector fields.

Clearly, a "$\mathcal{C}^\infty$ parameterized family of vector fields" f on M^n may be identified with the "parameter-preserving $\mathcal{C}^\infty$ vector field" F on $\mathbb{R}^l \times M^n$ that is defined by

$$F(p,x) = (\ 0 \ , \ f(p,x)\ ). \tag{10}$$

Also, for a parameter-preserving system $(F_o, F_1, \ldots, F_m)$, the submanifolds $\pi_1^{-1}(\{p\})$ (i.e., $\{p = \text{constant}\}$) are invariant: for any p, identifying the submanifold $\pi_1^{-1}(\{p\})$ with M^n, we obtain a control system on M^n, which is exactly $\mathcal{S}_p$ defined in (1) if the f_i's are given by $F_i(p,x) = (\ 0 \ , \ f_i(p,x)\ )$. Therefore, a family of systems on M^n and a parameter-preserving system on $\mathbb{R}^l \times M^n$ describe the same object.

2.3. Lie brackets. If f and f' are two $\mathcal{C}^\infty$ parameterized families of vector fields, we define their Lie bracket $[f, f']$ to be the $\mathcal{C}^\infty$ parameterized family of vector fields such that for any p, $[f, f'](p,.)$ is the usual Lie bracket of $f(p,.)$ and $f'(p,.)$:

$$[f,f'](p,x) \ \triangleq \ [f(p,.), f'(p,.)](x). \tag{11}$$

We have the following obvious relation, if $F = (0, f)$ and $F' = (0, f')$, where the right-hand side is the usual Lie bracket on $\mathbb{R}^l \times M^n$:

$$(0\,, [f,f']) = [F\,,\, F']\,. \tag{12}$$

2.4. Derivative with respect to the parameters. Considering a $\mathcal{C}^\infty$ parameterized family of vector fields f and noticing that, for a fixed x, when p varies, $f(p,x)$ remains in the same vector space T_xM^n, one may define the $\mathcal{C}^\infty$ parameterized families of vector fields $\frac{\partial f}{\partial p_i}$ $(i = 1, \ldots, l)$, $\frac{\partial f}{\partial p_i}(.,x)$ being just the ith partial derivative of the map $p \longmapsto f(p,x)$ from $\mathbb{R}^l$ to T_xM^n. A straightforward computation provides, if f and F are related by (10) ($F = (0,f)$), the relation

$$\left(0\,,\frac{\partial f}{\partial p_i}\right) = \left[\frac{\partial}{\partial p_i}\,,\,F\right]. \tag{13}$$

2.5. Diffeomorphisms. A $\mathcal{C}^\infty$ *parameterized family of diffeomorphisms* φ on the manifold M^n is a $\mathcal{C}^\infty$ map

$$\varphi\colon \mathbb{R}^l \times M^n \longrightarrow M^n$$

such that for any p, φ_p (defined by $\varphi_p(x) = \varphi(p,x)$) is a diffeomorphism of M^n.

A $\mathcal{C}^\infty$ *diffeomorphism* Φ of $\mathbb{R}^l \times M^n$ is called *parameter preserving* if

$$\pi_1 \circ \Phi \;=\; \pi_1, \tag{14}$$

which just means that for any (p,x) the p-component of $\Phi(p,x)$ is p. There is a one-to-one correspondence between parameter-preserving diffeomorphisms on M^n and parameterized families of diffeomorphisms on $\mathbb{R}^l \times M^n$, relating Φ and φ by

$$\Phi(p,x) = (\;p\;,\;\varphi(p,x)\;) = (\;p\;,\;\varphi_p(x)\;). \tag{15}$$

It is clear that Φ is a $\mathcal{C}^\infty$ parameter-preserving diffeomorphism on $\mathbb{R}^l \times M^n$ if and only if φ is a $\mathcal{C}^\infty$ parameterized family of diffeomorphisms on M^n because Φ is one-to-one onto if and only if φ_p is so for all p, and the differential of Φ is given by

$$\Phi'(p,x) = \begin{pmatrix} I & 0 \\ \dfrac{\partial\varphi}{\partial p}(p,x) & \dfrac{\partial\varphi}{\partial x}(p,x) \end{pmatrix}. \tag{16}$$

We define the $\mathcal{C}^\infty$ parameterized family of diffeomorphisms φ^{-1} by $\varphi^{-1}(p,.) = \varphi_p^{-1}$. With this natural definition of φ^{-1} and Φ defined according to (15), we have

$$\Phi^{-1}(p,x) = (\;p\;,\;\varphi_p^{-1}(x)\;) = (\;p\;,\;\varphi^{-1}(p,x)\;). \tag{17}$$

Let f be a parameterized family of vector fields and φ be a parameterized family of diffeomorphisms. For any fixed p, f_p is a vector field on M^n and φ_p is a diffeomorphism from M^n to M^n. We may therefore define the vector field transformed of f_p by φ_p, $\varphi_{p\,*}\,f_p$, whose value at x is

$$\varphi_{p\,*}\,f_p\,(x) = \frac{\partial\varphi}{\partial x}(p,\varphi^{-1}(p,x)).f(p,\varphi^{-1}(p,x)) \tag{18}$$

(with the above definition of φ^{-1}) and whose flow at time t is

$$\varphi_p \circ \phi^t \circ \varphi_p^{-1} \tag{19}$$

if ϕ^t is the flow of $f(p,.)$ at time t. We then *define* the parameterized family of vector fields $\varphi_* f$ by

$$(\varphi_* f)\,(p,.) = \varphi_{p\,*}\,f_p. \tag{20}$$

If F and f are related according to (10), i.e., $F = (0,f)$, and Φ and φ are related according to (15), we have the following obvious relation:

$$(0\,,\,\varphi_* f) = \Phi_* F \tag{21}$$

where Φ_* is defined by $(\Phi_* F)\,(\Phi(p,x)) = \Phi'(p,x)\,.\,F(p,x)$, i.e., it is the usual transformation on vector fields induced by Φ.

Of course, all of this is also meaningful *locally.* If φ is defined only on an open subset U of $I\!R^l \times M^n$, it is a $\mathcal{C}^\infty$ family of diffeomorphism if and only if it is a $\mathcal{C}^\infty$ map and it induces, for any p (such that $p \in \pi_1(U)$), a diffeomorphism from $\{p\} \times \pi_1^{-1}(p)$ to $\varphi(\{p\} \times \pi_1^{-1}(p))$. The corresponding Φ (defined according to (5)) is simply a parameter-preserving diffeomorphism from U to $\Phi(U)$.

2.6. Modules of vector fields, distributions. Since we wish to give some results without assuming that certain ranks are constant, we will not use distributions but modules of vector fields. A module (over the ring of $\mathcal{C}^\infty$ functions) of vector fields is a set of vector fields closed under addition and multiplication by $\mathcal{C}^\infty$ functions. A module $\mathcal{F}$ of vector fields on a manifold X defines a distribution $\mathcal{D}$ on X: $\mathcal{D}$ is the subset.of TX defined as follows : for any x on X, $\mathcal{D}_x = T_xX \cap \mathcal{D} = \{F(x) \,|\, F \in \mathcal{F}\}$; it is clear that for all $x \in X$, $\mathcal{D}_x$ is a vector subspace of T_xX. If the rank of $\mathcal{D}$ is locally constant, i.e., if the function $x \in X \mapsto \dim\mathcal{D}_x \in Z\!\!\!Z_+$ is locally constant, then $\mathcal{D}$ determines $\mathcal{F}$, but otherwise there are in general several distinct modules that define the same distribution. In this sense, modules of vector fields are more precise a tool than distributions.

Of course, one may speak of modules, over the ring of $\mathcal{C}^\infty$ functions of p and x, of parameterized families of vector fields on M^n, and of modules, over the same ring, of parameter-preserving vector fields on $I\!R^l \times M^n$. If $f_1, \ldots, f_r$ are some parameterized families of vector fields on M^n (resp., $F_1, \ldots, F_r$ are some parameter-preserving vector fields on $I\!R^l \times M^n$);

$$\text{Span}\{f_1 \ldots f_r\} \quad (\text{resp.,} \quad \text{Span}\{F_1 \ldots F_r\}) \tag{22}$$

stands for the module of parameterized families of vector fields generated by $f_1, \ldots, f_r$ (resp., the module of parameter-preserving vector fields generated by $F_1, \ldots, F_r$), which is composed of all the linear combinations—on the module of $\mathcal{C}^\infty$ functions of (p,x)—of $f_1, \ldots, f_r$ (resp., $F_1, \ldots, F_r$). Therefore, if f is a parameterized family of vector fields,

$$f \in \text{Span}\{f_1, \ldots, f_r\} \tag{23}$$

means that there exists some $\mathcal{C}^\infty$ functions $a^1 \ldots a^r$ such that

$$f(p,x) = \sum_{k=1}^{m} a^k(p,x) f_k(p,x); \tag{24}$$

and if F is a parameter-preserving vector field,

$$F \in \text{Span}\{F_1, \ldots, F_r\} \tag{25}$$

means that there exists some $\mathcal{C}^\infty$ functions $a^1 \ldots a^r$ such that

$$F(p,x) = \sum_{k=1}^{m} a^k(p,x) F_k(p,x). \tag{26}$$

Of course, if f and the f_i's are related to F and the F_i's according to (10), then (23) is equivalent to (25) and one may take the same coefficients a^k in (24) and (26) if they hold.

Remark 1. A parameterized family of vector fields on M^n defines, for any p, a vector field on M^n. Therefore, a module $\mathcal{M}$ of parameterized families of vector fields on M^n (or equivalently a module of parameter-preserving vector fields on $\mathbb{R}^l \times M^n$) defines, for any p, a module $\mathcal{M}_p$ of vector fields on M^n. Then one might expect that a parameterized family of vector fields f belongs to a module of parameterized families of vector fields $\mathcal{M}$ if and only if, for any p, the corresponding vector field f_p belongs to the corresponding module of vector fields $\mathcal{M}_p$.

This is true if $\mathcal{M}$ (or the associated module of parameter-preserving vector fields $\mathcal{M}$) spans a distribution of constant rank, for both of these conditions are then equivalent to some vectors, being pointwise in some vector spaces. If the distribution spanned by the module $\mathcal{M}$ does not have constant rank, this is usually false, even if $\mathcal{M}$ is finitely generated, as shown by the following example. Let $\mathbb{R}^l$ and M^n both be $\mathbb{R}$ and f_o be the $\mathcal{C}^\infty$ parameterized family of vector fields defined by

$$f_o(p,x) = p^2 \frac{\partial}{\partial x}, \tag{27}$$

and let $\mathcal{M}$ be the module of parameterized families of vector fields generated by f_o. Consider the $\mathcal{C}^\infty$ parameterized family of vector fields f defined by

$$f(p,x) = p \frac{\partial}{\partial x}. \tag{28}$$

It does not belong to $\mathcal{M}$; i.e., there exists no $\mathcal{C}^\infty$ function a such that $f(p,x) = a(p,x) f_o(p,x)$. However, for any p (even $p = 0$), the vector field f_p defined by $f_p(x) = f(p,x)$ belongs to the module $\mathcal{M}_p$ spanned by the vector field $f_{0,p}$ ($f_{0,p}(x) = f_o(p,x)$) because there exists for any p a $\mathcal{C}^\infty$ function a_p of x such that $f_p(x) = a_p(x) f_{0,p}(x)$, given by $a_p(x) \equiv 1$ if $p = 0$ and $a_p(x) \equiv \frac{1}{p}$ if $p \neq 0$.

2.7. Notation. We use the following notation throughout this paper:

- We denote parameterized families of vector fields on M^n by lower-case letters, and by the corresponding upper-case letter the corresponding parameter-preserving vector field on $\mathbb{R}^l \times M^n$ according to (10).
- The parameterized family of systems that we are studying is $\mathcal{S} = (f_0, f_1, \dots, f_m)$; see (1)–(2). We do not distinguish it from the corresponding parameter-preserving system on $\mathbb{R}^l \times M^n$, and we write indifferently $\mathcal{S} = (f_o, f_1, \dots, f_m)$ or $\mathcal{S} = (F_o, F_1, \dots, F_m)$.
- $\mathcal{g}$ is the module of parameterized families of vector fields on M^n defined by

$$\mathcal{g} = \text{Span}\,\{f_1, \dots, f_m\}, \tag{29}$$

and $\mathcal{G}$ is the module of parameter-preserving vector fields on $\mathbb{R}^l \times M^n$ defined by

$$\mathcal{G} = \text{Span}\{\ F_1\ , \dots,\ F_m\ \}. \tag{30}$$

Of course, the "parameterized family of distributions" associated with $\mathcal{g}$ maps (p,x) into the vector subspace Range $g(p,x)$ (image of the linear mapping $g(p,x)$, spanned by the vector $f_1(p,x), \dots, f_m(p,x)$) of T_xM^n. If the rank of $g(p,x)$ is constant, $Z \in \mathcal{G}$ or $z \in \mathcal{g}$ is equivalent to $z(p,x) \in$ Range $g(p,x)$ for all (p,x).

- If Z is a parameter-preserving vector field (resp., z is a parameterized family of vector fields), $Z \in \mathcal{G}$ (resp., $z \in \mathcal{g}$) *means* that there exists some $\mathcal{C}^\infty$ functions $a^1 \dots a^m$ such that $Z = \sum_{k=1}^m a^k F_k$ (resp., such that $z = \sum_{k=1}^m a^k f_k$).

3. Pure feedback equivalence. In this section, we deal with equivalence of the systems $\mathcal{S}_p$ via feedback transformations only. We establish both local and global conditions.

As made precise by the following definitions, we say that the family of systems $\mathcal{S}$ is pure feedback equivalent (FE) if a feedback transformation that depends smoothly on the parameters transforms it into a "constant" family.

DEFINITION 3.1 (Constant). *A family $\mathcal{S} = (f_o, f_1, \dots, f_m)$ of systems is* constant *on an open subset U of $\mathbb{R}^l \times M^n$ if for all x, p^1, and p^2 such that (p^1, x) and (p^2, x) are in U, we have*

$$f_k(p^1, x) = f_k(p^2, x), \;\; k = 0, 1, \dots, m. \tag{31}$$

DEFINITION 3.2 (FE). *A family $\mathcal{S} = (f_o, f_1, \dots, f_m)$ is* FE (pure feedback equivalent) *on an open subset U of $\mathbb{R}^l \times M^n$ if there exist two C^∞ maps α and β*

$$\begin{aligned} \alpha : \;\; & U \longrightarrow \mathbb{R}^m, \\ \beta : \;\; & U \longrightarrow \mathrm{M}_{m\times m}(\mathbb{R}) \end{aligned}$$

such that $\beta(p, x)$ is invertible for any (p, x) in U and the family $\widetilde{\mathcal{S}} = (\widetilde{f_o}, \widetilde{f_1}, \dots, \widetilde{f_m})$ defined by (with g (resp., $\tilde{g}$) related to $f_1 \dots f_m$ (resp., $\tilde{f_1} \dots \tilde{f_m}$) according to (1)–(2))

$$\widetilde{f_o}(p, x) = f_o(p, x) \; + \; g(p, x) \; \alpha(p, x), \tag{32}$$

$$\widetilde{g}(p, x) = g(p, x) \; \beta(p, x) \tag{33}$$

is constant on U.

It is locally FE *at $(\bar{p}, \bar{x})$ if it is FE on a certain open neighborhood U of $(\bar{p}, \bar{x})$ in $\mathbb{R}^l \times M^n$. It is* globally FE *if it is FE on $\mathbb{R}^l \times M^n$.*

Pure feedback equivalence may be completely characterized, as seen on the following theorems. A discussion and a comparison with other results is given below.

PROPOSITION 3.3. *Let $\mathcal{S} = \{F_0, F_1, \dots, F_m\}$ be a parameterized family of systems on M^n and U be an open subset of $\mathbb{R}^l_. \times M^n$ of the form*

$$U = C \times V \tag{34}$$

where V is an open subset of M^n and $C \subset \mathbb{R}^l$ is a product of open intervals. Then the following three properties are equivalent:

1. *$\mathcal{S}$ is FE on U.*
2. *On U,*

$$\left[\frac{\partial}{\partial p_i}, F_k\right] \in \mathcal{G}, \qquad \begin{array}{l} k = 0, 1, \dots, m, \\ i = 1, \dots, l. \end{array} \tag{35}$$

3. *On U,*

$$\frac{\partial f_k}{\partial p_i} \in \mathcal{G}, \qquad \begin{array}{l} k = 0, 1, \dots, m, \\ i = 1, \dots, l. \end{array} \tag{36}$$

THEOREM 3.4. *A family $\mathcal{S} = (f_0, f_1, \dots, f_m)$ is* locally FE *at $(\bar{p}, \bar{x})$ if and only if conditions* 2 *or* 3 *hold for a certain neighborhood U of $(\bar{p}, \bar{x})$. A family $\mathcal{S} = (f_0, f_1, \dots, f_m)$ is* globally FE *if and only if conditions* 2 *or* 3 *hold with $U = \mathbb{R}^l \times M^n$.*

Theorem 3.4 is a straightforward corollary of Proposition 3.3. Proposition 3.3 is proved further.

Note that we are able to obtain a global result, which would be false if we were considering two a priori completely independent systems instead of a continuous family. Actually, we give in [7] an example of two systems with the same control module and a zero drift vector field such that no global feedback transformation can transform one into the other.

Note also that it was proved in [1] for the case of a constant rank control distribution and in [3] for the general case that two systems (without parameters) are locally pure feedback equivalent if and only if the module spanned by the control vector fields is the same for the two systems and the difference between the drift vector fields of the two systems belongs to this module. From the lemma given in the appendix, applied to the case $X = \frac{\partial}{\partial p_i}$, $\{G_1, \dots, G_s\} = \{F_1, \dots, F_m\}$, condition 2 of the above proposition implies existence of an $m \times m$ invertible matrix transforming $f_1(p,x), \dots, f_m(p,x)$ into $f_1(q,x), \dots, f_m(q,x)$ for any p and q, this matrix depending smoothly on x, p, and q; this implies that the module of vector fields generated by the control vector fields $f_1(p,.), \dots, f_m(p,.)$ of the system $\mathcal{S}_p$ does not depend on p and that for two different p, the difference between the two drift vector fields $f_0(p,.)$ belongs to this module and hence implies the necessary and sufficient condition quoted above for any two systems obtained for different values of the parameters to be pure feedback equivalent. However, the converse is not true: on one hand, condition 2 is stronger since it implies smooth dependence on the parameters of the matrix quoted above; on the other hand, FE is stronger than any two systems in the family that are pure feedback equivalent to one another, as seen on the following example.

Consider the family of systems in $\mathbb{R}$ that depends on one parameter ($\mathbb{R}^l = M^n = \mathbb{R}$) and is defined by

$$f_0(p,x) = e^{-\frac{1+p^2}{x^2}}, \quad f_1(p,x) \;=\; e^{-\frac{1}{x^2}}; \tag{37}$$

i.e., the system $\mathcal{S}_p$ is

$$\dot{x} = e^{-\frac{1+p^2}{x^2}} + e^{-\frac{1}{x^2}}\, u. \tag{38}$$

Note that, despite the vanishing denominators, f_0 and f_1 are C^∞ on $\mathbb{R}^2$. All these systems are feedback equivalent to one another. The transformation that turns $\mathcal{S}_p$ into $\mathcal{S}_0$ is $u = v + 1 - e^{-p^2/x^2}$ (i.e., $\alpha = e^{-1/x^2}$ and $\beta = 1$) if $p \neq 0$ and $u = v$ (i.e., $\alpha = 0$ and $\beta = 1$) if $p = 0$. However, it is impossible to define α as a continuous function of (p,x) since the one given here is the only solution at all points where $x \neq 0$ and cannot be prolonged at $(0,0)$. The family is therefore not FE in our sense, and one may verify that condition (35), for example, is not met.

Proof of Proposition 3.3. Points 2 and 3 are equivalent, as an obvious consequence of (13) and the facts delineated in (23)–(26).

Point 1 implies 3 because, on one hand, (33) may be written

$$f_k(p,x) = \sum_{j=1}^{m} \beta^j_{-1,k}(p,x)\, \widetilde{f}_j(p,x), \quad k = 1, \dots, m, \tag{39}$$

where the $\beta^j_{-1,k}$'s are the coefficients of $\beta(p,x)^{-1}$, and the fact that $\widetilde{\mathcal{S}}$ is constant implies that $\frac{\partial \widetilde{f}_k}{\partial p_i} = 0$, so taking the derivative of both members in (39) with respect to p_i gives (36) for $k = 1, \ldots, m$; on the other hand, (32) may be rewritten, with $\widetilde{\alpha}$ an obvious function of α and the invertible β:

$$\widetilde{f_0}(p,x) \;=\; f_0(p,x) \;+\; \sum_{j=1}^{m} \widetilde{\alpha}^j(p,x)\,\widetilde{f_j}(p,x), \tag{40}$$

which implies (36) for $k = 0$ by taking the derivative of both members in (39) with respect to p_i.

Let us prove that 2 implies 1 to conclude this proof. By applying the lemma given in the appendix to the case where $\mathcal{D}$ is $\mathcal{G}$, the G_k's are the F_k's, F is F_0, X is $\frac{\partial}{\partial p_i}$, and χ is (p,x). We get, from (35), that there exists some $\mathcal{C}^\infty$ functions $c^j_{i,k}$ such that

$$\left(\phi^{t_i}_{\frac{\partial}{\partial p_i}}\right)_* F_k\,(p,x) = \sum_{i=1}^{s} c^j_k(p,x,t_i) F_j(p,x), \tag{41}$$

$$\left(\phi^{t_i}_{\frac{\partial}{\partial p_i}}\right)_* F_0\,(p,x) = F_0(p,x) \;+\; \sum_{i=1}^{s} c^j_k(p,x,t_i) G_j(p,x), \tag{42}$$

which means

$$f_k(p_1, \ldots, p_i - t_i, p_{i+1}, \ldots, p_l, x) = \sum_{i=1}^{s} c^j_k(p,x,t_i) f_j(p,x), \tag{43}$$

$$f_0(p_1, \ldots, p_i - t_i, p_{i+1}, \ldots, p_l, x) = f_0(p,x) \;+\; \sum_{i=1}^{s} c^j_0(p,x,t_i) G_j(p,x); \tag{44}$$

and the lemma given in the appendix says, in addition, that the matrix

$$C_i(p,x,t_i) \;=\; \begin{pmatrix} 1 & 0 & \ldots & 0 \\ c^1_{0,i}(p,x,t) & c^1_{1,i}(p,x,t) & \ldots & c^1_{m,i}(p,x,t) \\ \vdots & \vdots & & \vdots \\ c^m_{0,i}(p,x,t) & c^m_{1,i}(p,x,t) & \ldots & c^m_{m,i}(p,x,t) \end{pmatrix} \tag{45}$$

is invertible for any (p,x,t) (to be precise, the lemma states that the matrix obtained by removing the first line and first column in $C_i(p,x,t)$ is invertible). Let us define the invertible $l \times l$ matrix $\beta(p,x)$ and the l-column vector $\alpha(p,x)$ by

$$\begin{pmatrix} 1 & 0 \\ \alpha(p,x) & \beta(p,x) \end{pmatrix} = B_1(p,x)\, B_2(p,x)\, \ldots\, B_l(p,x) \tag{46}$$

where

$$B_i(p_1, \ldots, p_i, p_{i+1}, \ldots, p_l, x) = C_i(p_1, \ldots, p_i, 0, \ldots, 0, x, p_i). \tag{47}$$

Using relation (43) for $i = l, \ldots, 1$, one gets

$$\begin{aligned} f_0(0, \ldots, 0, x) \;&=\; f_0(p,x) \;+\; \textstyle\sum_{i=1}^{s} \beta^j_0(p,x) f_j(p,x), \\ f_k(0, \ldots, 0, x) \;&=\; \textstyle\sum_{i=1}^{s} \beta^j_k(p,x) f_j(p,x), \quad k = 1, \ldots, m, \end{aligned} \tag{48}$$

which implies that the family $\widetilde{\mathcal{S}}$, defined by (32)–(33), is constant because $\widetilde{f_k}(p,x) = \widetilde{f_k}(0,x)$. Note that, since C is a product of intervals, this construction does define α and β over $U = C \times V$. □

4. Feedback and diffeomorphism equivalence.

4.1. Definition. Feedback and diffeomorphism equivalence, for a parameterized family of systems, means that the systems of the family are equivalent to each other via feedback and diffeomorphism, where both the feedback transformations and the diffeomorphisms are smooth functions of both the points and the parameters.

DEFINITION 4.1 (FDE). *A family $\mathcal{S} = (f_o, f_1, \ldots, f_m)$ is* locally FDE (feedback and diffeomorphism equivalent) *at $(\bar{p}, \bar{x})$ if and only if there exists a neighborhood U of $(\bar{p}, \bar{x})$ and a family of diffeomorphisms φ, defined on U, such that the family $\widetilde{\mathcal{S}} = (\widetilde{f}_o, \widetilde{f}_1, \ldots, \widetilde{f}_m)$ defined, in a neighborhood of $(\bar{p}, \varphi(\bar{p}, \bar{x}))$, by*

$$\widetilde{f}_k = \varphi_* f_k \tag{49}$$

is locally FE at $(\bar{p}, \varphi(\bar{p}, \bar{x}))$.

It is globally FDE *if and only if there exists a family of diffeomorphisms φ defined all over $I\!R^l \times M^n$ such that the family $\widetilde{\mathcal{S}} = (\widetilde{f}_o, \widetilde{f}_1, \ldots, \widetilde{f}_m)$ defined on $I\!R^l \times M^n$ by (49) is globally FE.*

This exactly means that there exist some φ, α, β and $\widetilde{f}_0, \widetilde{g}$ (or $\widetilde{f}_0, \ldots, \widetilde{f}_m$) such that if $y = \varphi(p, x)$ and $u = \alpha(p, x) + \beta(p, x)v$, (1) reads

$$\dot{y} = \tilde{f}(y) + \tilde{g}(y)\, v. \tag{50}$$

4.2. Necessary and sufficient conditions for FDE. The two following theorems give some necessary and sufficient conditions for a family of systems to be FDE, locally or globally. Each of them give two equivalent necessary and sufficient conditions, labeled 2 and 3. These conditions are given in terms of the existence of some vector fields that satisfy some relations. Their existence is necessary, and when they exist φ may be computed explicitly from them. The difference between conditions 2 and 3 is that 3 asks for vector fields $Y_1, \ldots, Y_l$ to satisfy some decoupled relations (each Y_i has to satisfy a set of relations not involving any other Y_j), whereas in 2 the vector fields $Z_1, \ldots, Z_l$ have to satisfy some individual relations *plus* commutation relations involving all of them. Condition 3 is much more practical a tool than 2 to check FDE. Actually the practical content of the theorems is 1 $\Longleftrightarrow$ 3.

These vector fields can be interpreted as elements of the "Lie Algebra" of the infinite-dimensional "Lie group" of diffeomorphisms of M^n and φ as a submanifold of this infinite dimensional Lie group, parameterized by $p = (p_1, \ldots, p_l)$, i.e., for instance, a curve in the case of one parameter.

THEOREM 4.2. *Let $\mathcal{S} = (F_0, F_1, \ldots, F_m)$ be a parameterized family of systems.*

Local FDE. The following three propositions are equivalent:

1. *$\mathcal{S}$ is locally FDE at $(\bar{p}, \bar{x})$ (resp., is globally FDE).*

2. *There exist some C^∞ vector fields $Z_1, \ldots, Z_l$, defined on a certain neighborhood of $(\bar{p}, \bar{x})$, such that, on their domain of definition,*

$$Z_i \textit{ is parameter preserving}, \qquad i = 1, \ldots, l, \tag{51}$$

$$\left[\frac{\partial}{\partial p_i} + Z_i \;,\; F_k \right] \in \mathcal{G}, \qquad i = 1, \ldots, l, \quad k = 0, 1, \ldots, m, \tag{52}$$

$$\left[\frac{\partial}{\partial p_i} + Z_i, \; \frac{\partial}{\partial p_j} + Z_j \right] = 0, \qquad i, j = 1, \ldots, l. \tag{53}$$

3. *There exist some C^∞ vector fields $Y_1, \ldots, Y_l$, defined on a certain neighborhood of $(\bar{p}, \bar{x})$, such that, on their domain of definition,*

$$Y_i \text{ is parameter preserving}, \qquad i = 1, \ldots, l, \tag{54}$$

$$\left[\frac{\partial}{\partial p_i} + Y_i \ , \ F_k \right] \in \mathcal{G}, \qquad i = 1, \ldots, l \ , \ k = 0, 1, \ldots, m. \tag{55}$$

Moreover, if these conditions are satisfied, an explicit expression of φ on a neighborhood of $(\bar{p}, \bar{x})$ is

$$\varphi(p, x) = \pi_2 \left(\phi^{\bar{p} - p}(p, x) \right), \tag{56}$$

where π_2 is defined by (3) *and ϕ by*

$$\phi^{(t_1, \ldots, t_l)} = \phi_1^{t_1} \circ \cdots \circ \phi_l^{t_l}, \tag{57}$$

ϕ_i being the flow of the vector field $\frac{\partial}{\partial p_i} + Y_i$:

$$\phi_i^0(p, x) \equiv (p, x), \tag{58}$$

$$\frac{\partial}{\partial t}[\phi_i^t(p, x)] = \frac{\partial}{\partial p_i} + Y_i(\phi_i^t(p, x)). \tag{59}$$

Global FDE. *If one adds, in* 2 *and* 3, *the condition that the vector fields $\frac{\partial}{\partial p_i} + Z_i$ and $\frac{\partial}{\partial p_i} + Y_i$ be defined all over $\mathbb{R}^l \times M^n$ and complete, then the result holds globally; i.e.,* 1 *can be replaced by "S is globally FDE" and φ may then be computed according to* (56) *all over $\mathbb{R}^l \times M^n$, choosing any $\bar{p}$ in $\mathbb{R}^l$.*

Proof of Theorem 4.2. We write the proof for the global result; the proof of the local result is similar, replacing $\mathbb{R}^l \times M^n$ with some neighborhood of $(\bar{p}, \bar{x})$ and omitting the parts concerned with the completeness of $\frac{\partial}{\partial p_i} + Z_i$ or φ being defined everywhere.

1 ⇒ 2. Let Φ be the parameter-preserving diffeomorphism on $\mathbb{R}^l \times M^n$ associated with the family of diffeomorphisms φ (see (15)). We have, from (49) and (21),

$$\widetilde{F}_k = \Phi_* F_k. \tag{60}$$

Since $\widetilde{S}$ is FE, we have, from Theorem 3.4,

$$\left[\frac{\partial}{\partial p_i} \ , \ \widetilde{F}_k \right] \in \widetilde{\mathcal{G}}, \qquad i = 1, \ldots, l \ , \ k = 0, 1, \ldots, m. \tag{61}$$

Applying $(\Phi_*)^{-1}$ to this relation between vector fields, i.e., writing its inverse image by Φ, one gets

$$\left[(\Phi_*)^{-1} \frac{\partial}{\partial p_i} \ , \ F_k \right] \in \mathcal{G}, \qquad i = 1, \ldots, l. \tag{62}$$

Therefore, if we define the Z_i's by

$$\frac{\partial}{\partial p_i} + Z_i = (\Phi_*)^{-1} \frac{\partial}{\partial p_i}, \tag{63}$$

(52) is obviously satisfied. (53) is also satisfied because

$$\left[\frac{\partial}{\partial p_i} + Z_i, \ \frac{\partial}{\partial p_j} + Z_j \right] = (\Phi_*)^{-1} \left[\frac{\partial}{\partial p_i}, \ \frac{\partial}{\partial p_j} \right] = 0; \tag{64}$$

(51) is also satisfied, i.e., the p-component of Z_i defined by (63) is zero, because, from (16),

$$\left((\Phi_*)^{-1}\frac{\partial}{\partial p_i}\right)(p,x)=\left(\frac{\partial}{\partial p_i},\ -\frac{\partial\varphi}{\partial x}(p,x)^{-1}\frac{\partial\varphi}{\partial p_i}(p,x)\right). \tag{65}$$

If φ (and therefore Φ) is defined all over $I\!R^l\times M^n$, (63) defines the Z_i's all over $I\!R^l\times M^n$, and, for any i, the vector field $\frac{\partial}{\partial p_i}+Z_i$ is complete because it is the diffeomorphic image of the complete vector field $\frac{\partial}{\partial p_i}$.

2 ⇒ 3. The Y_i's defined by $Y_i=Z_i$ obviously work.

3 ⇒ 1. *Without loss of generality, we suppose, in this part of the proof, that $\bar{p}$ is* 0 (*replace p with $p-\bar{p}$*).

Let us define φ by (56). This does define φ over $I\!R^l\times M^n$ if the Y_i's are defined on $I\!R^l\times M^n$ and are complete. If the Y_i's are only defined on a neighborhood of $(\bar{p},\bar{x})$ or are not complete, this defines φ on a neighborhood of $(\bar{p},\bar{x})$.

We now have to prove that the family $\widetilde{\mathcal{S}}$ given by (60), where Φ is the parameter-preserving diffeomorphism defined from φ according to (15), is FE on $I\!R^l\times M^n$ (on a neighborhood of $(\bar{p},\bar{x})$ for the local case). From Theorem 3.4, it is (locally or globally) FE if and only if, on a neighborhood of $(\bar{p},\bar{x})$ or on $I\!R^l\times M^n$, $[\frac{\partial}{\partial p_i},\Phi_*F_k]\in\Phi_*\mathcal{G}$ or, equivalently,

$$\left[(\Phi_*)^{-1}\frac{\partial}{\partial p_i}\,,\,F_k\right]\in\mathcal{G},\qquad i=1,\dots,l,\ k=0,1,\dots,m. \tag{66}$$

Only (66) remains to be proved. Defining φ according to (56) is equivalent to defining Φ by (remember that we suppose in this proof that $\bar{p}=0$)

$$\Phi(p,x)=\left(p\,,\,\pi_2\left(\phi^{-p}(p,x)\right)\right). \tag{67}$$

From (57), considering the fact that the p-component of $\phi^{-p}(p,x)$ is 0 and defining for any $i=1,\dots,l$ and any $t\in I\!R$ the translation τ_i^t to be the flow at time t of the vector field $\frac{\partial}{\partial p_i}$:

$$\tau_i^t(p_1,\dots,p_l,x)=(p_1,\dots,p_i+t,\dots,p_l,x), \tag{68}$$

(67) may be rewritten as

$$\Phi(p_1,\dots,p_l,x)=\tau_l^{p_l}\circ\cdots\circ\tau_1^{p_1}\circ\phi_1^{-p_1}\circ\cdots\circ\phi_l^{-p_l}\ (p_1,\dots,p_l,x), \tag{69}$$

which implies

$$\begin{aligned}\Phi^{-1}(p,x)&=\phi_l^{p_l}\circ\cdots\circ\phi_1^{p_1}\circ\tau_1^{-p_1}\circ\cdots\circ\tau_l^{-p_l}\ (p_1,\dots,p_l,x)\\&=\phi_l^{p_l}\circ\cdots\circ\phi_1^{p_1}\ (0,x).\end{aligned} \tag{70}$$

Therefore,

$$\begin{aligned}\left(\Phi^{-1}{}_*\frac{\partial}{\partial p_i}\right)(p,x)&=\left(\Phi^{-1}\right)'\left(\Phi(p,x)\right)\cdot\frac{\partial}{\partial p_i}(\Phi(p,x))\\&=\left(\phi_l^{p_l}\circ\cdots\circ\phi_{i+1}^{p_{i+1}}\right)'\left(\phi_{i+1}^{-p_{i+1}}\circ\cdots\circ\phi_l^{-p_l}\ (p,x)\right)\\&\qquad\cdot\left[\frac{\partial}{\partial p_i}+Y_i(\phi_{i+1}^{-p_{i+1}}\circ\cdots\circ\phi_l^{-p_l}\ (p,x))\right]\end{aligned} \tag{71}$$

which reads

$$
\begin{aligned}
(\Phi_*)^{-1}\frac{\partial}{\partial p_i} &= (\phi_l^{p_l} \circ \cdots \circ \phi_{i+1}^{p_{i+1}})_* \left(\frac{\partial}{\partial p_i} + Y_i\right) \\
&= \phi_l^{p_l}{}_* \cdots {}_*\phi_{i+1}^{p_{i+1}}{}_* \left(\frac{\partial}{\partial p_i} + Y_i\right)
\end{aligned}
\tag{72}
$$

so that (66) may be rewritten as

$$
\left[\frac{\partial}{\partial p_i} + Y_i \,,\, \phi_{i+1}^{-p_{i+1}}{}_* \cdots {}_*\phi_l^{-p_l}{}_* F_k\right] \in \phi_{i+1}^{-p_{i+1}}{}_* \cdots {}_*\phi_l^{-p_l}{}_*\mathcal{G}, \qquad \begin{array}{l} i = 1, \ldots, l, \\ k = 0, 1, \ldots, m. \end{array}
\tag{73}
$$

But a classical expression for the Lie bracket is

$$
\left[\frac{\partial}{\partial p_i} + Y_i \,,\, \phi_{i+1}^{-p_{i+1}}{}_* \cdots {}_*\phi_l^{-p_l}{}_* F_k\right] = \frac{d}{dt}\,\phi_i^{-t}{}_*\phi_{i+1}^{-p_{i+1}}{}_* \cdots {}_*\phi_l^{-p_l}{}_* F_k \bigg|_{t=0},
\tag{74}
$$

and from (55) and the lemma given in the appendix,

$$
\begin{aligned}
\phi_i^{-t}{}_*\phi_{i+1}^{-p_{i+1}}{}_* \cdots {}_*\phi_l^{-p_l}{}_* F_k &\in \mathcal{G}, \qquad k = 1, \ldots, m, \\
F_0 - \phi_i^{-t}{}_*\phi_{i+1}^{-p_{i+1}}{}_* \cdots {}_*\phi_l^{-p_l}{}_* F_0 &\in \mathcal{G}, \\
\phi_{i+1}^{-p_{i+1}}{}_* \cdots {}_*\phi_l^{-p_l}{}_*\mathcal{G} &= \mathcal{G},
\end{aligned}
\tag{75}
$$

which proves that (73), and therefore (66), is true. □

4.3. How to use these necessary and sufficient conditions. Given a family of systems defined by certain F_0, $F_1, \ldots, F_m$, the previous theorems tell us that the systems of the family are feedback and diffeomorphism equivalent (i.e., the family is FDE) if and only if one may find some vector fields $Y_1, \ldots, Y_l$ satisfying (54) and (55) and give a way to build the diffeomorphisms from these vector fields. The interesting question then is: how to determine whether these vector fields exist and how actually to compute them. We will not give a general answer to this question but take a look at the form of the equations for $Y_1, \ldots, Y_l$.

Locally, let us use a system of coordinates $x_1, \ldots, x_n$ on M^n (then $p_1, \ldots, p_l$, $x_1, \ldots, x_n$ is a system of coordinates on $\mathbb{R}^l \times M^n$). (54) means that the Y_i's have no component on the coordinate vector fields $\frac{\partial}{\partial p_1}, \ldots, \frac{\partial}{\partial p_l}$, so we may write

$$
Y_i(p, x) = b_i^1(p, x)\frac{\partial}{\partial x_1} + \cdots + b_i^n(p, x)\frac{\partial}{\partial x_n}.
\tag{76}
$$

We may now translate condition (55) into some equations that involve the functions b_i^j. Suppose that $\mathcal{G}$ has constant rank m' ($m' \leq m$), and let $\eta_{m'+1}, \ldots, \eta_n$ be $n - m'$ independent differential forms of the form

$$
\eta_j(p, x) = c_j^1(p, x)\,\mathrm{d}x_1 + \cdots + c_j^n(p, x)\,\mathrm{d}x_n
\tag{77}
$$

vanishing on $\mathcal{G}$ so that the equation of $\mathcal{G}$ is $\mathrm{d}p_1 = \cdots = \mathrm{d}p_l = \eta_{m'+1} = \cdots = \eta_n = 0$. (55) may then be written, for each i, $1 \leq i \leq l$, and all k, $0 \leq k \leq m$, as

$$
\begin{aligned}
\left\langle \mathrm{d}p_j, \left[\frac{\partial}{\partial p_i} + Y_i \,,\, F_k\right]\right\rangle &= 0, \qquad j = 1, \ldots, l, \\
\left\langle \eta_j, \left[\frac{\partial}{\partial p_i} + Y_i, F_k\right]\right\rangle &= 0, \qquad j = m' + 1, \ldots, n.
\end{aligned}
\tag{78}
$$

The first part of (78) is always satisfied if Y_i has no p-component (see (76)), and the second part may be rewritten as

$$(79) \qquad L_{Y_i}\langle \eta_j, F_k\rangle - L_{F_k}\langle \eta_j, Y_i\rangle + \mathrm{d}\eta_j(Y_i, F_k) = \left\langle \eta_j, \left[\frac{\partial}{\partial p_i}, F_k\right]\right\rangle.$$

This may be written as $l(n - m')(m + 1)$ equations in the b_i^j's. The first term in the left-hand side depends linearly on the b_i^j's and their first partial derivatives, the two other terms in the left-hand side depend linearly on the the b_i^j's only, and the right-hand side does not depend on the b_i^j's at all. Therefore (79) gives, for each $i = 1, \dots, l$, a set of $(n - m')(m + 1)$ linear partial differential equations in $b_i^1, \dots, b_i^n$, whose satisfaction is equivalent to the Y_i given by (76) satisfying (55). Note that if $\mathcal{G}$ does not have constant rank, it is usually not possible to find a finite number of differential forms describing $\mathcal{G}$ and, therefore, to translate (55) into a finite number of equations in the coordinates of the vector fields Y_i.

The obtained set of PDEs may provide a practical way of checking whether or not a given family is FDE, and, if a solution of these PDEs is available; building φ from this solution is an systematic process according to (58)–(59).

We understand here the superiority of Theorem 4.2 over a theorem that would only contain the set of conditions labeled 2. Looking for $Z_1, \dots, Z_l$ meeting 2, i.e., (51), (52), and (53), would mean, if we define Z_i, instead of Y_i, by (76), that the b_i^j have to satisfy not only the set of linear PDEs (79), which may be decomposed into l decoupled systems each involving $b_i^1, \dots, b_i^n$ for a different value of i, but also the commutation relation (53), which can be translated into a set of PDEs that are nonlinear and involve all the different b_i^j's (actually each involves two different i's).

In this paper, we go no further in the characterization of FDE, i.e., we make no attempt to characterize the cases where the above-described set of partial differential equations admit some solutions. However, in the following sections, we will specialize the notion of feedback and diffeomorphism equivalence by restricting the p-dependence of φ on the parameter p. This will yield some conditions similar to those given here but with some restrictions on the vector fields Z_i or Y_i, and these additional conditions are such that the above-described set of partial differential equations is transformed into a set of algebraic equations. This will allow us to give some more explicit characterizations than for the present case of general FDE.

5. Feedback and diffeomorphism equivalence with matching conditions.

5.1. Definitions: FDEM1 and FDEM2.

Let us define two properties that are more restrictive than FDE. See §5.2 for comments and interpretations.

DEFINITION 5.1 (FDEM1). *A family $\mathcal{S} = (f_o, f_1, \dots, f_m)$ is* locally FDEM1 (feedback and diffeomorphism equivalent with matching conditions of the first kind) *at $(\bar{p}, \bar{x})$ (resp., globally FDEM1) if and only if it is locally FDE at $(\bar{p}, \bar{x})$ (resp., globally FDE), and φ has the property that there exists a smooth map v_1:*

$$(p, \dot{p}, x) \longmapsto v_1(p, \dot{p}, x) \in \mathbb{R}^m$$

defined for $(p, \dot{p}, x)$ in $\mathbb{R}^l \times \mathbb{R}^l \times M^n$ such that $(p, x) \in U$ (resp., for any $(p, \dot{p}, x)$ in $\mathbb{R}^l \times \mathbb{R}^l \times M^n$) satisfying

$$(80) \qquad \frac{\partial \varphi}{\partial x}(p, x) g(p, x) v_1(p, \dot{p}, x) + \frac{\partial \varphi}{\partial p}(p, x)\dot{p} = 0$$

at any point where both sides are defined.

DEFINITION 5.2 (FDEM2). *A family $\mathcal{S} = (f_o, f_1, \ldots, f_m)$ is* locally FDEM2 (feedback and diffeomorphism equivalent with matching conditions of the second kind) *at $(\bar{p}, \bar{x})$ (resp., globally FDEM2) if and only if it is locally FDE at $(\bar{p}, \bar{x})$ (resp., globally FDE), and φ has the property that there exists a smooth map v_2:*

$$(p, q, \dot{p}, x) \longmapsto v_2(p, q, \dot{p}, x) \in I\!R^m$$

defined for $(p, q, \dot{p}, x)$ in $I\!R^l \times I\!R^l \times I\!R^l \times M^n$ such that $(p, x) \in U$ and $(q, x) \in U$ (resp., any $(p, q, \dot{p}, x)$ in $I\!R^l \times I\!R^l \times I\!R^l \times M^n$) satisfying

$$\frac{\partial \varphi}{\partial x}(p, x) g(q, x) v_2(p, q, \dot{p}, x) + \frac{\partial \varphi}{\partial p}(p, x)\dot{p} = 0 \tag{81}$$

at any point where both sides are defined.

Remark 2. We have defined five properties for a parameterized family of vector fields: a parameterized family of systems can be constant, pure feedback equivalent (FE), feedback and diffeomorphism equivalent (FDE), or feedback and diffeomorphism equivalent with one of the matching conditions (FDEM1 and FDEM2), and also it can have none of these properties (a consequence of [11] is that most of the parameterized families of systems have none of these properties). It is rather obvious that

$$\text{Constant} \;\Rightarrow\; \text{FE} \;\Rightarrow\; \text{FDEM2} \;\Rightarrow\; \text{FDEM1} \;\Rightarrow\; \text{FDE}, \tag{82}$$

and actually all these implications are strict, although we will see that for some particular classes of systems FDEM1 and FDEM2 are equivalent.

5.2. Some interpretations of the "matching conditions" in FDEM1 and FDEM2. Note that FE does not refer to any φ, but it can also be understood as FDE with φ independent of p. FDEM1 or FDEM2 allow p-dependence of φ, but they restrict it.

These "matching conditions" are generalizations of those considered in [10], [5], [4] in the following sense: the "strict matching assumption" made in [10] is equivalent to FE plus one of the systems in the family being feedback linearizable and the "extended matching assumption" in [5], [4] is equivalent to FDEM1 or FDEM2 plus one of the systems in the family being feedback linearizable (from Theorem 5.11, FDEM1 and FDEM2 are equivalent for feedback linearizable systems). In these cases, system (50) can be taken a linear system.

The main motivation for these "matching conditions" comes from adaptive nonlinear control, i.e., for instance, the problem of stabilizing a system $\mathcal{S}_p$ (p constant) without the knowledge of p, i.e., with a controller that does not depend on p. Let us very briefly outline the use of these conditions for adaptive nonlinear control. For some precisions, the reader is referred to [10], [5], [4] and to [8], [9], [6]. Assume that the family $(\mathcal{S}_p)$ is globally FDE and that a feedback law $v = v_{ST}(y)$ is known that globally asymptotically stabilizes the origin in (50); then, for each p, the control law $u_p(x) = \alpha(p, x) + \beta(p, x) v_{ST}(\varphi(p, x))$ globally asymptotically stabilizes $\varphi^{-1}(p, 0)$ in system $\mathcal{S}_p$. In addition, if $U(y)$ is a Lyapunov function for $\dot{y} = \tilde{f}_0(y) + \tilde{g}(y) u_{ST}(y)$, then $V_p(x) = U(\varphi(p, x))$ decreases along the solutions of $\mathcal{S}_p$ in closed loop with $u = u_p(x)$. This is not a solution to the problem since the control law $u_p(x)$ depends on p, but we are in the framework of the most current available solutions for nonlinear adaptive stabilization, if we also make the assumption that f_0 and g depend *linearly* on the parameter p.

Under these assumptions and if not only FDE, but FE, is met, $V_p(x)$ does not depend on p. In that case (see [10], [9], [6], [8]), one may design a controller of the form $u = u_{\hat{p}}(x)$ with $\dot{\hat{p}}$ a suitable function of x and $\hat{p}$ such that $W = V_{\hat{p}}(x) + \|\hat{p} - p\|^2$ decreases in the closed-loop system (whose state is $(\hat{p}, x)$) obtained by controlling $\mathcal{S}_p$ with this controller. This dynamic controller does not depend on p and therefore provides a solution to the problem. If FE is not met, then $V_p(x)$ does depend on p in general and the time derivative of $V_{\hat{p}}(x)$ depends explicitly on $\dot{\hat{p}}$, making the previous design method inefficient. However, FDEM1, if satisfied, provides a way (see [5], [4], [8]) of cancelling this dependence by adding $v_1(\hat{p}, \dot{\hat{p}}, x)$ to the original control; this would mean setting $u = u_{\hat{p}}(x) + v_1(\hat{p}, \dot{\hat{p}}, x)$ and applying the previous design method for $\dot{\hat{p}}$; this unfortunately provides $\dot{\hat{p}}$ as a function of x, $\hat{p}$, and v_1, i.e., of x, $\hat{p}$, and $\dot{\hat{p}}$, which results in general in some unavoidable singularities when trying to define $\dot{\hat{p}}$ as a function of $(\hat{p}, x)$. A way to counteract this fact, if FDEM2 is met, is (see [8], [9], [6]) to use a "bigger" controller, with state $(\hat{p}, \hat{q})$: $u = u_{\hat{p}}(x) + v_2(\hat{p}, \hat{q}, \dot{\hat{p}}, x)$ where the v_2-term is still designed to cancel the dependence on $\dot{\hat{p}}$ of the time derivative of $V_{\hat{p}}(x)$; $\dot{\hat{p}}$ may then be defined as a function of x and $\hat{p}$, only, and $\dot{\hat{q}}$ as a function of x, $\hat{p}$, and v_2, i.e., of x, $\hat{p}$, and $\hat{q}$ (recall that $\dot{\hat{p}}$ is a function of x and $\hat{p}$).

Let us give another interpretation of FDEM1 and FDEM2 in terms of disturbance rejection or model matching. Suppose that FDE is satisfied and, therefore, that certain φ, α, and β are defined; see (50).

The paragraph after the definition of FDE (50) allows one to understand FDE as the possibility to render the input–output behavior $v \longrightarrow y$ with $y = \varphi(p, x)$ independent of p for any *constant* p by performing a suitable change of input (feedback transformation depending on x and p): $u = \alpha(p, y) + \beta(p, y)v$. Now suppose that p is time varying; i.e., consider the following time-varying system with state x and output y ($y \in M^n$):

$$(83) \qquad \begin{cases} \dot{x} = f_0(p(t), x) \;+\; g(p(t), x)\, u, \\ y = \varphi(p(t), x). \end{cases}$$

FDEM1 is a condition on φ, necessary and sufficient for the possibility, for any possible time-dependence $p(t)$, to design a change of control

$$(84) \qquad u = v_1(p(t), \dot{p}(t), x) + \alpha(p(t), \varphi(p(t), x)) + \beta(p(t), \varphi(p(t), x))\, v$$

for (83) to "match" the time-invariant model (50). In fact, this model-matching problem also amounts to the problem of rejecting the measured disturbance w in

$$(85) \qquad \begin{cases} \dot{x} = f_0(p, x) + g(p, x)\, u, \\ \dot{p} = w, \\ y = \varphi(p, x) \end{cases}$$

where (x, p) is the state and y is the output. By "rejecting the *measured* disturbance w", we mean building a control $u = \gamma(x, p, w, v)$, nonsingular with respect to v, such that the behavior of the output is affected by v and not by w. FDEM2 is more restrictive: the existence of v_2 can be interpreted as the possibility to reject, for any value of q (maybe time-varying), the disturbance w in the following system with inputs u_1 and u_2, where only the input u_2 is allowed, as a feedback, to depend on w:

$$(86) \qquad \begin{cases} \dot{x} = f_0(p, x) + g(p, x)\, u_1 + g(q, x)\, u_2, \\ \dot{p} = w, \\ y = \varphi(p, x). \end{cases}$$

Finally, a simple interpretation of FDEM1 or FDEM2 is (see Proposition 6.4) that the family (1) is FE if and only if the following family is FDEM1 or FDEM2:

$$\dot{x} = f_0(p,x) \;+\; g(p,x)\,z, \qquad \dot{z} = u.$$

5.3. Necessary and sufficient conditions for FDEM1. The following theorem gives some necessary and sufficient conditions for a family of systems to be FDEM1. The comments given before Theorem 4.2 hold for it as well; in particular, the conditions labeled 3 are simpler than the conditions labeled 2 (2 ⇒ 3 is obvious), and the practical content of these theorems is 1 ⟺ 3. The abstract conditions given here are translated into more explicit ones in §5.5, and some practically tractable conditions are given in §§5.6 and 5.7.

THEOREM 5.3. *Let $\mathcal{S} = (F_0, F_1, \dots, F_m)$ be a parameterized family of systems.*

Local FDEM1. The following three propositions are equivalent:

1. *$\mathcal{S}$ is locally FDEM1 at $(\bar{p},\bar{x})$.*

2. *There exist some C^∞ vector fields $Z_1,\dots,Z_l$, defined on a certain neighborhood of $(\bar{p},\bar{x})$, such that, on their domain of definition,*

$$Z_i \in \mathcal{G}, \qquad i = 1,\dots,l, \tag{87}$$

$$\left[\frac{\partial}{\partial p_i} + Z_i,\ F_k\right] \in \mathcal{G}, \qquad i = 1,\dots,l\ ,\ \ k = 0,1,\dots,m, \tag{88}$$

$$\left[\frac{\partial}{\partial p_i} + Z_i,\ \frac{\partial}{\partial p_j} + Z_j\right] = 0, \qquad i,j = 1,\dots,l. \tag{89}$$

3. *There exist some C^∞ vector fields $Y_1,\dots,Y_l$, defined on a certain neighborhood of $(\bar{p},\bar{x})$, such that, on their domain of definition,*

$$Y_i \ \in\ \mathcal{G}, \qquad i = 1,\dots,l, \tag{90}$$

$$\left[\frac{\partial}{\partial p_i} + Y_i\ ,\ F_k\right] \ \in\ \mathcal{G}, \qquad i = 1,\dots,l\ ,\ \ k = 0,1,\dots,m, \tag{91}$$

$$\left[Y_i,\ \frac{\partial}{\partial p_j}\right] \ \in\ \mathcal{G}, \qquad i,j = 1,\dots,l\ ,\ \ i \neq j. \tag{92}$$

Moreover, if these conditions are satisfied, φ may then be computed according to (56) *on a neighborhood of $(\bar{p},\bar{x})$.*

Global FDEM1. If one adds, in 2 *and* 3, *the condition that the vector fields $\frac{\partial}{\partial p_i} + Z_i$ and $\frac{\partial}{\partial p_i} + Y_i$ be defined all over $I\!R^l \times M^n$ and complete, then the result holds globally, i.e.,* 1 *can be replaced by "$\mathcal{S}$ is globally FDEM1" and φ may then be computed according to* (56) *all over $I\!R^l \times M^n$, choosing any $\bar{p}$ in $I\!R^l$.*

Proof of Theorem 5.3. 1 ⇒ 2. As in the proof of Theorem 4.2, we define the vector fields $Z_1,\dots,Z_l$ according to (63). The only property to be proved is that they belong (where they are defined) to $\mathcal{G}$. This is true because, from (65), (80) may be rewritten, with $\dot{p} = (\dot{p}_1,\dots,\dot{p}_l)$ and $v_1(p,\dot{p},x) = (v_1^1(p,\dot{p},x),\dots,v_1^m(p,\dot{p},x))$ as follows:

$$\sum_{i=1}^{l} Z_i(p,x)\dot{p}_i = \sum_{k=1}^{m} v_1^k(p,\dot{p},x)\, F_k(p,x). \tag{93}$$

2 ⇒ 3. The Y_i's defined by $Y_i = Z_i$ work. (89) implies (92) because $[\frac{\partial}{\partial p_i} + Z_i, \frac{\partial}{\partial p_j}]$ $= -[\frac{\partial}{\partial p_i} + Z_i, Z_j]$ is in $\mathcal{G}$ from (88) because Z_j is in $\mathcal{G}$.

$\underline{3 \Rightarrow 1.}$ (Note that a more intuitive proof is given in Remark 3 for the case when the rank of some distributions are constant.) As in the proof of Theorem 4.2, we define φ from the Y_i's according to (56) or, equivalently, Φ according to (67). The only additional property to prove, compared to Theorem 4.2, is the existence of the smooth v_1 satisfying (80).

Let us define the Z_i's by (63) and note that, from (65), (80) is equivalent to (93). Let $\mathcal{H}$ be the module defined from $\mathcal{G}$ by

$$\mathcal{H} = \mathcal{G} + \operatorname{Span}\left\{\frac{\partial}{\partial p_1}, \dots, \frac{\partial}{\partial p_l}\right\}. \tag{94}$$

Since $\mathcal{H}$ is obviously finitely generated and (91)–(92) imply $[\frac{\partial}{\partial p_i} + Y_i, \mathcal{H}] \subset \mathcal{H}$, the lemma given in the appendix implies that, for any $t \in I\!R$ and at any point where ϕ_i^t is defined,

$$X \in \mathcal{H} \Rightarrow \phi_{i*}^t X \in \mathcal{H}, \quad i = 1, \dots, l. \tag{95}$$

Therefore (72) and (90) imply that

$$\Phi_*^{-1} \frac{\partial}{\partial p_i} \in \mathcal{H}, \quad i = 1, \dots, l; \tag{96}$$

or, considering (63) and (65),

$$Z_i \in \mathcal{G}, \quad i = 1, \dots, l. \tag{97}$$

This is equivalent to the existence of smooth functions a_i^k such that

$$Z_i = \sum_{k=1}^{m} a_i^k F_k. \tag{98}$$

Note that, from (65), (80) is equivalent to (93). One v_1 meeting (80) is then given by

$$v_1 = (v_1^1(p, \dot{p}, x), \dots, v_1^m(p, \dot{p}, x)),$$
$$v_1^k(p, \dot{p}, x) = \textstyle\sum_{i=1}^l a_i^k(p, x)\, \dot{p}_i. \qquad \square$$

Remark 3. In the case where the ranks of both the distribution spanned by $\mathcal{G}$ and the distribution spanned by the module $\mathcal{L}$ defined by (131) are constant, it is possible to give a simpler proof of 3 ⇒ 1 based only on some involutive distributions and which is independent of the proof of Theorem 4.2. If we define the module $\mathcal{K}$ by

$$\mathcal{K} = \{X \in \mathcal{H} / [X, F_k] \in \mathcal{H},\ k = 0 \dots m \text{ and } \left[\frac{\partial}{\partial p_i}, X\right] \in \mathcal{H},\ i = 1 \dots l\}, \tag{99}$$

$\mathcal{H}$ being defined by (94), it is not difficult to show that the following three facts are true. First, since it may be rewritten as

$$\mathcal{K} = \{X \in \mathcal{H} / [X, \mathcal{H}] \subset \mathcal{H} \text{ and } [X, F_0] \in \mathcal{H}\}, \tag{100}$$

$\mathcal{K}$ is stable under Lie bracket. Second, (90) and (91) exactly mean that $\frac{\partial}{\partial p_i} + Y_i$ is in $\mathcal{K}$. Third, the set of vector fields that are in $\mathcal{K}$ and have a zero p-component, i.e., the kernel of the restriction of Π_1 to $\mathcal{K}$ (see (5)–(6)), is exactly $\mathcal{L}$ defined by (131).

The two last facts plus the fact that we assume here that $\mathcal{L}$ has constant rank imply that $\mathcal{K}$ has constant rank $(l + \text{Rank}\mathcal{L})$ and, therefore, from the first fact, $\mathcal{K}$ spans an integrable distribution. Now the construction of φ according to (56) implies that two points on a level submanifold of φ might be joined by a finite concatenation of integral curves of the vector fields $\frac{\partial}{\partial p_i} + Y_i$, which, since these vector fields are in $\mathcal{K}$ and $\mathcal{K}$ is integrable, implies that the level submanifolds of φ are included in the integral submanifolds of $\mathcal{K}$. Since the tangent space to the level submanifolds of φ is spanned by the vector fields $\Phi_*^{-1}\frac{\partial}{\partial p_i}$, this proves that $\Phi_*^{-1}\frac{\partial}{\partial p_i}$ belongs to $\mathcal{K}$, which proves both (because it implies (62)) that φ transforms the family $\mathcal{S}$ into a family $\widetilde{\mathcal{S}}$, which is FE, and (because it implies (96) or (97)) that φ satisfies (80).

5.4. Necessary and sufficient conditions for FDEM2. Theorem 5.6 gives necessary and sufficient conditions for a family of systems to be locally or globally FDEM2. Unfortunately, these conditions are not as good as those given for FDE or FDEM1 in Theorems 4.2 and 5.3; i.e., we are not able to give a set of conditions similar to those labeled 3 in those theorems. The conditions we give are similar to those labeled 2 in those theorems; i.e., they contain some commutation relations that make them hard to check in practice. Actually, we give a counterexample showing that a condition like these labeled 3 in Theorems 4.2 and 5.3 would not be sufficient. This is the motivation for Theorem 5.7 which gives two different sufficient conditions for FDEM2 that do not involve any commutation relation; these conditions are not necessary. The conditions given in the present section are explained in §5.5, and §§5.7 and 5.8 present some cases where it is possible to give a simple characterization.

The conditions we are going to state require the definition of the following submodule of $\mathcal{G}$.

DEFINITION 5.4. *For U an open subset of $\mathbb{R}^l \times M^n$ of the form*

$$U = C \times V \tag{101}$$

where V is an open subset of M^n and $C \subset \mathbb{R}^l$ is a product of intervals, we define the submodule $\mathcal{G}^{\cap,U}$ of $\mathcal{G}$ the following way: a vector field Y is in $\mathcal{G}^{\cap,U}$ if and only if it is in $\mathcal{G}$ and, for any (p,x) in U and $t_1,\ldots,t_m$ such that $\tau_1^{t_1} \circ \tau_2^{t_2} \circ \cdots \circ \tau_l^{t_l}(p,x)$ is in U, we have

$$\tau_1^{t_1}{}_*\tau_2^{t_2}{}_* \cdots{}_* \tau_l^{t_l}{}_*Y(p,x) = \sum_{k=1}^{m} b^k(t_1,\ldots,t_l,p,x)F_k(p,x) \tag{102}$$

where $\tau_i^{t_i}$ is the flow at time t_i of the vector field $\frac{\partial}{\partial p_i}$ (see (68)) and $b^1 \ldots b^m$ are C^∞ functions from the part of $\mathbb{R}^l \times C \times M^n$ made of triplets (t,p,x) such that $p+t$ is in C to $\mathbb{R}$.

We will write $\mathcal{G}^{\cap,\text{glob}}$ instead of $\mathcal{G}^{\cap,\mathbb{R}^l \times M^n}$.

Remark 4. Using parameterized families of vector fields instead of parameter-preserving vector fields, we may equivalently say that a vector field Y is in $\mathcal{G}^{\cap,U}$ if and only if it is in $\mathcal{G}$ and the associated parameterized family of vector fields y is such that there exists some C^∞ functions $\widehat{b}^1 \ldots \widehat{b}^m$ of (p,q,x) such that, for any p, q, and x such that both (p,x) and (q,x) are in U, we have

$$y(q,x) = \widehat{b}^1(p,q,x)f_1(p,x) + \cdots + \widehat{b}^m(p,q,x)f_m(p,x). \tag{103}$$

The functions $\widehat{b}^1,\ldots,\widehat{b}^m$ may easily be deduced from $b^1,\ldots,b^m$ in (102) since q stands for $p-(t_1,\ldots,t_m)$.

If the module $\mathcal{G}^{\cap,U}$ is finitely generated, then a simpler characterization can be given. This is useful to actually compute $\mathcal{G}^{\cap,U}$:

PROPOSITION 5.5. • *If the largest submodule of $\mathcal{G}$ that is invariant by the vector fields $\frac{\partial}{\partial p_1}, \dots, \frac{\partial}{\partial p_m}$ on U is finitely generated, then it is equal to $\mathcal{G}^{\cap,U}$.*

• *If the module $\mathcal{G}$ spans a* distribution *of constant rank on $\mathbb{R}^l \times M^n$, which we denote by $\mathcal{G}$ as well, and if the* distribution *Δ defined on U by*

$$\Delta(p,x) = \bigcap_{q \in V} \mathcal{G}(q,x) \tag{104}$$

has constant rank on U, then the distribution spanned by $\mathcal{G}^{\cap,U}$ is exactly Δ.

Note that the intersection in (104) makes sense since the tangent space at the point (q,x) to $\mathbb{R}^l \times M^n$ is $T_q\mathbb{R}^l \times T_xM^n$, which may be identified to $\mathbb{R}^l \times T_xM^n$, so that the different $\mathcal{G}(q,x)$ may be considered as subspaces of the same vector space $\mathbb{R}^l \times T_xM^n$.

Proof of Proposition 5.5. The first point is a straightforward consequence of the lemma given in the appendix, with $\mathcal{G}$ as $\mathcal{D}$, $\frac{\partial}{\partial p_1}$, as X, and $F = 0$. Δ given by (104) is invariant by the vector fields $\frac{\partial}{\partial p_1}, \dots, \frac{\partial}{\partial p_m}$ and contains any subdistribution of $\mathcal{G}$ invariant by these vector fields; since it has constant rank, it uniquely defines a finitely generated module that, from the first point, is exactly $\mathcal{G}^{\cap,U}$ □

Let us now state our necessary and sufficient conditions.

THEOREM 5.6. *Let $\mathcal{S} = (F_0, F_1, \dots, F_m)$ be a family of systems.*

<u>*Local FDEM2.*</u> *The two following propositions are equivalent:*

1. *$\mathcal{S}$ is locally FDEM2 at $(\bar{p},\bar{x})$.*
2. *There exist some $\mathcal{C}^\infty$ vector fields $Z_1, \dots, Z_l$, defined on a certain neighborhood U of $(\bar{p},\bar{x})$, such that, on their domain of definition,*

$$Z_i \in \mathcal{G}^{\cap,U}, \qquad i = 1,\dots,l, \tag{105}$$

$$\left[\frac{\partial}{\partial p_i} + Z_i \ , \ F_k\right] \in \mathcal{G}, \qquad i = 1,\dots,l, \ k = 0,1,\dots,m, \tag{106}$$

$$\left[\ \frac{\partial}{\partial p_i} + Z_i,\ \frac{\partial}{\partial p_j} + Z_j\right] = 0, \qquad i,j = 1,\dots,l. \tag{107}$$

Moreover, if these conditions are satisfied, φ may then be computed according to (56), *with Y_i replaced by Z_i, on a neighborhood of $(\bar{p},\bar{x})$.*

<u>*Global FDEM2.*</u> *If one adds, in* 2, *the condition that the vector fields $\frac{\partial}{\partial p_i} + Z_i$ be defined all over $\mathbb{R}^l \times M^n$ and complete, and replaces* (105) *with*

$$Z_i \in \mathcal{G}^{\cap,\text{glob}}, \quad i = 1,\dots,l, \tag{108}$$

then the result holds globally, i.e., 1 *can be replaced by "$\mathcal{S}$ is globally FDEM2" and φ may then be computed according to* (56), *with Y_i replaced by Z_i, all over $\mathbb{R}^l \times M^n$, choosing any $\bar{p}$ in $\mathbb{R}^l$.*

Proof of Theorem 5.6. 1 ⇒ 2. The only thing that is not implied by Theorem 4.2 is that the Y_i's in that theorem belong to $\mathcal{G}^\cap$ under the present assumptions. This is true from Definition 5.4 because, from (63) and (65), we may rewrite (81) as

$$\sum_{i=1}^{l} y_i(p,x)\dot{p}_i = \sum_{k=1}^{m} g(q,x)v_{2,k}(p,\dot{p},q,x) \tag{109}$$

where $v_2(p,\dot p,q,x) = (v_{2,1}(p,\dot p,q,x),\dots,v_{2,m}(p,\dot p,q,x))$, $\dot p = (\dot p_1,\dots,\dot p_l)$, and y_i is the parameterized family of vector fields associated with the parameter-preserving vector field Y_i.

<u>$2 \Rightarrow 1$</u>. We can use the proof of $3 \Rightarrow 1$ in Theorem 4.2, replacing Y_i by Z_i. Relation (107) actually simplifies this proof considerably since it implies that the flows $\phi_1^{t_1},\dots,\phi_l^{t_l}$ commute. Therefore (67) and (57) imply that

$$\Phi \circ \phi_i^t \circ \Phi^{-1}\ (p_1,\dots,p_l,x) = (p_1,\dots,p_i+t,\dots,p_l,x),$$

i.e., $\Phi \circ \phi_i^t \circ \Phi^{-1}$ is the flow at time t of $\frac{\partial}{\partial p_i}$, and therefore

$$\Phi_* \left(\frac{\partial.}{\partial p_i} + Z_i\right) = \frac{\partial}{\partial p_i}. \tag{110}$$

(66) is then a consequence of (110) and (106). Note that when, as it is the case here, the vector fields from which we construct φ commute, then defining φ by (56) (substituting Z_i to Y_i) means taking φ constant along the integral submanifolds of $\{\frac{\partial}{\partial p_1} + Z_1,\dots,\frac{\partial}{\partial p_l} + Z_l\}$.

We now need to prove the existence of v_2 satisfying (81). This is a consequence of Z being in $\mathcal{G}^{\cap}$. From Definition 5.4 and (103), if z_i is the parameterized family of vector fields on M^n associated to Z_i, there exists some functions b_i^k such that

$$z_i(p,x) = \sum_{k=1}^{m} b_i^k(p,q,x)\, f_k(q,x), \tag{111}$$

so (81) is satisfied defining v_2 by

$$\begin{aligned} v_2 &= (v_2^1(p,q,\dot p,x),\dots,v_2^m(p,q,\dot p,x)),\\ v_2^k(p,q,\dot p,x) &= \textstyle\sum_{i=1}^l b_i^k(p,q,x)\,\dot p_i. \qquad \Box \end{aligned}$$

As noticed above (see also the discussion in §5.5), these theorems are not as convenient as those concerning FDE and FDEM1 since they do not give a condition free of commutation relations. The next theorem gives two different sufficient conditions for FDEM2, involving no commutation relation. The commutation relations are actually replaced by (114) or a certain module being a Lie algebra. The theorem is only proved in the case where $\mathcal{G}^{\cap,U}$ is finitely generated.

THEOREM 5.7. *Let $\mathcal{S} = (F_0,F_1,\dots,F_m)$ be a parameterized family of systems. Conditions* 1 *and* 2 *below are both sufficient for $\mathcal{S}$ to be locally FDEM*2 *at $(\bar p,\bar x)$ (resp., globally FDEM*2*).*

1. *For a certain neighborhood U of $(\bar p,\bar x)$, the module $\mathcal{G}^{\cap,U}$ is finitely generated and there exist some C^∞ vector fields $Y_1,\dots,Y_l$, defined on U, such that, on U,*

$$Y_i \in \mathcal{G}^{\cap,U}, \qquad i=1,\dots,l, \tag{112}$$

$$\left[\frac{\partial}{\partial p_i} + Y_i\ ,\ F_k\right] \in \mathcal{G}, \qquad i=1,\dots,l\ ,\ \ k=0,1,\dots,m, \tag{113}$$

$$\left[\frac{\partial}{\partial p_i} + Y_i,\ \mathcal{G}^{\cap,U}\right] \subset \mathcal{G}^{\cap,U}, \qquad i=1,\dots,l. \tag{114}$$

(resp., the same with $U = \mathbb{R}^l \times M^n$, plus the vector fields $\frac{\partial}{\partial p_1} + Y_1,\dots,\frac{\partial}{\partial p_l} + Y_l$ are complete).

2. *For a certain neighborhood U of $(\bar{p}, \bar{x})$, the module*

$$\mathcal{J}^U = \{ X \in \mathcal{G}^{\cap,U} ,\ [X, \mathcal{G}^{\cap,U}] \in \mathcal{G} \} \tag{115}$$

is finitely generated and stable under Lie bracket, and there exist some C^∞ vector fields $Y_1, \ldots, Y_l$, defined on U, satisfying (112) *and* (113) *on U (resp., the same with $U = I\!R^l \times M^n$, plus the vector fields $\frac{\partial}{\partial p_1} + Y_1, \ldots, \frac{\partial}{\partial p_l} + Y_l$ are complete).*
Moreover, φ may still then be computed according to (56).

Proof of Theorem 5.7. Condition 1 is sufficient. This is very similar to the 3 ⇒ 1 of the proof of Theorem 5.3. Replace $\mathcal{G}$ with $\mathcal{G}^{\cap,U}$ and $\mathcal{H}$ with

$$\mathcal{H}^{\cap,U} \;=\; \mathcal{G}^{\cap,U} \;+\; \text{Span}\left\{ \frac{\partial}{\partial p_1}, \ldots, \frac{\partial}{\partial p_l} \right\}. \tag{116}$$

The fact that $\mathcal{G}^{\cap,U}$, and therefore $\mathcal{H}^{\cap,U}$, are finitely generated plus condition (114) still enable one to derive (95) and therefore (96) using (72) and (112). (97) becomes $Z_i \in \mathcal{G}^{\cap,U}$, and v_2 may therefore be constructed as in the proof of Theorem 5.6; see (111) and (5.4).

Condition 2 is sufficient. This is also similar to the part 3 ⇒ 1 of the proof of Theorem 5.3. After (94), consider

$$\mathcal{M} = \{ X \in \mathcal{H}^{\cap,U} ,\ [X, \mathcal{G}^{\cap,U}] \in \mathcal{G} \}. \tag{117}$$

The vector fields $\frac{\partial}{\partial p_i} + Y_i$ are in $\mathcal{M}$ from (112) and (113). Since $\mathcal{J}^U$ is stable under the Lie bracket, $\mathcal{M}$ is stable under the Lie bracket too (use the fact that $[\frac{\partial}{\partial p_i}, \mathcal{G}^{\cap,U}] \subset \mathcal{G}^{\cap,U}$). From the lemma given in the appendix, and since $\mathcal{M}$ is finitely generated ($\mathcal{M} = \mathcal{J}^U \oplus \text{Span}\{\frac{\partial}{\partial p_1} \cdots \frac{\partial}{\partial p_l}\}$), this implies (95) and (96) with $\mathcal{H}$ replaced with $\mathcal{H}^{\cap,U}$. This implies $Z_i \in \mathcal{G}^{\cap,U}$, which allows the same construction of v_2 as in the proof of Theorem 5.6; see (111) and (5.4). □

Example. The present example in $I\!R^8$ proves that our Theorem 5.6 concerning FDEM2 is false in general if we remove the commutation conditions (107) and that the sufficient conditions given in Theorem 5.7 are not necessary.

Dimension 8 is almost minimal: from Theorem 5.13, we need $\mathcal{G}^\cap$ to have dimension at least 4 and $\mathcal{G}$ to have dimension at least 6. We do not know if it is possible to find a counterexample where the dimension is exactly 6; in the present case it is 7, so in order for the system to be nontrivial, the state must have dimension at least 8.

Let us consider the family of systems in $I\!R^8$ ($M^n = I\!R^8$) depending on two parameters ($l = 2$), with seven inputs ($m = 7$), defined by

$$\left.\begin{aligned}
&\dot{x}_1 = -x_5 u_3 - (p_1 + x_7)(p_2 + x_8) u_5 - (p_2 + x_8) u_6 - (p_1 + x_7) u_7,\\
&\dot{x}_2 = u_7 + x_6 u_2,\\
&\dot{x}_3 = u_6 - x_5 u_1,\\
&\dot{x}_4 = u_5 - x_7 u_1,\\
&\dot{x}_5 = u_4 \ , \quad \dot{x}_6 \;=\; u_3 \ , \quad \dot{x}_7 \;=\; u_2 \ , \quad \dot{x}_8 \;=\; u_1;
\end{aligned}\right\} \tag{118}$$

i.e.,

$$\left.\begin{aligned}
&F_0 = 0, \quad F_1 \;=\; -x_5 \frac{\partial}{\partial x_3} - x_7 \frac{\partial}{\partial x_4} + \frac{\partial}{\partial x_8},\\
&F_2 = x_6 \frac{\partial}{\partial x_2} + \frac{\partial}{\partial x_7}, \quad F_3 = -x_5 \frac{\partial}{\partial x_1} + \frac{\partial}{\partial x_6},\\
&F_4 = \frac{\partial}{\partial x_5}, \quad F_5 = -(p_1 + x_7)(p_2 + x_8) \frac{\partial}{\partial x_1} + \frac{\partial}{\partial x_4},\\
&F_6 = -(p_2 + x_8) \frac{\partial}{\partial x_1} + \frac{\partial}{\partial x_3}, \quad F_7 = -(p_1 + x_7) \frac{\partial}{\partial x_1} + \frac{\partial}{\partial x_2}.
\end{aligned}\right\} \tag{119}$$

The equation of $\mathcal{G}$ is $\mathrm{d}p_1 = \mathrm{d}p_2 = \omega = 0$ where ω is the differential form given by

$$\begin{aligned} \omega = {} & \mathrm{d}x_1 + (p_1+x_7)\,\mathrm{d}x_2 + (p_2+x_8)\,\mathrm{d}x_3 + (p_1+x_7)\,(p_2+x_8)\,\mathrm{d}x_4 + x_5\mathrm{d}x_6 \\ & -x_6\,(p_1+x_7)\,\mathrm{d}x_7 + x_5\,(p_2+x_8)\,\mathrm{d}x_8 + x_7\,(p_1+x_7)\,(p_2+x_8)\,\mathrm{d}x_8. \end{aligned} \tag{120}$$

We have, for any U

$$\mathcal{G}^{\cap,U} = \mathcal{G}^{\cap,\mathrm{glob}} = \mathcal{G}^{\cap} = \mathrm{Span}\{\,F_1\,,\,F_2\,,\,F_3\,,\,F_4\,\} \tag{121}$$

because the right-hand side is invariant under $\frac{\partial}{\partial p_1}$ and $\frac{\partial}{\partial p_2}$ and no linear combination of F_5, F_6, and F_7 can have both its Lie brackets with $\frac{\partial}{\partial p_1}$ and with $\frac{\partial}{\partial p_2}$ in $\mathcal{G}$.

We are looking for

$$\begin{aligned} Y_1 &= a_1^1\,F_1 + a_1^2\,F_2 + a_1^3\,F_3 + a_1^4\,F_4, \\ Y_2 &= a_2^1\,F_1 + a_2^2\,F_2 + a_2^3\,F_3 + a_2^4\,F_4 \end{aligned} \tag{122}$$

such that $[\frac{\partial}{\partial p_i} + Y_i, F_k] \in \mathcal{G}$. This is equivalent to

$$\left\langle \omega\,,\, \left[\frac{\partial}{\partial p_i} + Y_i, F_k\right]\right\rangle = 0, \quad i=1,2\ ,\ k=1,\dots,7. \tag{123}$$

Here, we have

$$\left.\begin{array}{lcl} \langle\,\omega\,,\,[\frac{\partial}{\partial p_1} +.Y_1, F_1]\,\rangle & = & -\,(p_1+x_7)\,(p_2+x_8)\,a_1^2 - (p_2+x_8)\,a_1^4, \\ \langle\,\omega\,,\,[\frac{\partial}{\partial p_1} + Y_1, F_2]\,\rangle & = & (p_1+x_7)\,(p_2+x_8)\,a_1^1\ + (p_1+x_7)\,a_1^3, \\ \langle\,\omega\,,\,[\frac{\partial}{\partial p_1} + Y_1, F_3]\,\rangle & = & -\,(p_1+x_7)\,a_1^2 - a_1^4, \\ \langle\,\omega\,,\,[\frac{\partial}{\partial p_1} + Y_1, F_4]\,\rangle & = & (p_2+x_8)\,a_1^1\ + a_1^3, \\ \langle\,\omega\,,\,[\frac{\partial}{\partial p_1} + Y_1, F_5]\,\rangle & = & -\,(p_2+x_8) - (p_1+x_7)\,a_1^1\ - (p_2+x_8)\,a_1^2, \\ \langle\,\omega\,,\,[\frac{\partial}{\partial p_1} + Y_1, F_6]\,\rangle & = & -a_1^1, \\ \langle\,\omega\,,\,[\frac{\partial}{\partial p_1} + Y_1, F_7]\,\rangle & = & -1 - a_1^2 \end{array}\right\} \tag{124}$$

and

$$\left.\begin{array}{lcl} \langle\,\omega\,,\,[\frac{\partial}{\partial p_2} + Y_2, F_1]\,\rangle & = & -\,(p_1+x_7)\,(p_2+x_8)\,a_2^2\ - (p_2+x_8)\,a_2^4, \\ \langle\,\omega\,,\,[\frac{\partial}{\partial p_2} + Y_2, F_2]\,\rangle & = & (p_1+x_7)\,(p_2+x_8)\,a_2^1\ + (p_1+x_7)\,a_2^3, \\ \langle\,\omega\,,\,[\frac{\partial}{\partial p_2} + Y_2, F_3]\,\rangle & = & -\,(p_1+x_7)\,a_2^2\ - a_2^4, \\ \langle\,\omega\,,\,[\frac{\partial}{\partial p_2} + Y_2, F_4]\,\rangle & = & (p_2+x_8)\,a_2^1\ + a_2^3, \\ \langle\,\omega\,,\,[\frac{\partial}{\partial p_2} + Y_2, F_5]\,\rangle & = & -\,(p_1+x_7) - (p_1+x_7)\,a_2^1\ - (p_2+x_8)\,a_2^2, \\ \langle\,\omega\,,\,[\frac{\partial}{\partial p_2} + Y_2, F_6]\,\rangle & = & -1 - a_2^1 \\ \langle\,\omega\,,\,[\frac{\partial}{\partial p_2} + Y_2, F_7]\,\rangle & = & -a_2^2. \end{array}\right\} \tag{125}$$

It can be seen easily that (123)–(125) has only one solution for the a_k^j's:

$$\begin{array}{llcl} a_1^1 = 0, & a_2^1 & = & -1, \\ a_1^2 = -1, & a_2^2 & = & 0, \\ a_1^3 = 0, & a_2^3 & = & p_2+x_8, \\ a_1^4 = p_1+x_7, & a_2^4 & = & p_1+x_7 \end{array}$$

yielding, from (122), a unique solution of (123):

$$Y_1 = -x_6\frac{\partial}{\partial x_2} + (p_1+x_7)\,\frac{\partial}{\partial x_5} - \frac{\partial}{\partial x_7}, \tag{126}$$

$$Y_2 = -x_5\,(p_2+x_8)\,\frac{\partial}{\partial x_1} + x_5\frac{\partial}{\partial x_3} + x_7\frac{\partial}{\partial x_4} + (p_2+x_8)\,\frac{\partial}{\partial x_6} - \frac{\partial}{\partial x_8}. \tag{127}$$

These are the *only* solutions, and their Lie bracket is not zero:

$$\left[\frac{\partial}{\partial p_1}+Y_1, \frac{\partial}{\partial p_2}+Y_2\right] = -(p_1+x_7)(p_2+x_8)\frac{\partial}{\partial x_1}+(p_2+x_8)\frac{\partial}{\partial x_2} \tag{128}$$
$$+(p_1+x_7)\frac{\partial}{\partial x_3}-\frac{\partial}{\partial x_4}.$$

The vector fields given by (126)–(127) satisfy the conditions required for the Z_i in Theorem 5.6—those for global FDEM2 since it is easy to check that they are complete—except the commutation relation (107). Since they are the only solutions, there exists no other vector field satisfying both these same relations and (107). The family therefore is not (either locally or globally) FDEM2 because the conditions given by Theorem 5.6 are necessary. This proves that condition (107) cannot be omitted from Theorem 5.6 (see first paragraph of §5.4).

This example also allows us to see that the sufficient conditions given in Theorem 5.7 are not necessary by considering the family depending on one parameter only that is obtained by fixing $p_2 = 0$. It is (globally) FDEM2 because Y_1 given by (126) satisfies the conditions of Theorem 5.6 (there is no longer any commutation relation to check because there is only one Y_i). However, we may check that neither condition 1 nor condition 2 of Theorem 5.7 is satisfied.

Condition 1. Y_1 given by (126) is the unique vector field satisfying (112) and (113), and it does not satisfy the additional requirement (114) in condition 1 of Theorem 5.7 since, for example, F_3 is in $\mathcal{G}^\cap$ and $[\frac{\partial}{\partial p_1}+Y_1, F_3] = -(p_1+x_7)\frac{\partial}{\partial x_1}+\frac{\partial}{\partial x_2}$ is not in $\mathcal{G}^\cap$ from (119) and (121).

Condition 2. Since there is no longer any dependence on p_2, $\mathcal{G}^\cap$ is larger than it was in (121): it is now spanned by F_1, F_2, F_3, F_4, $F_5 - x_8F_7$, and F_6. The vector fields Y_1 given by (126) and Y_1' given by

$$Y_1' = F_5 - x_8F_7 + (p_1+x_7)F_6$$

are in $\mathcal{J}^U$ (which does not depend on U) since they are in $\mathcal{G}^\cap$ and their Lie brackets with F_1, F_2, F_3, F_4, $F_5 - x_8F_7$, and F_6 are in $\mathcal{G}$. However, since $[Y_1, Y_1'] = F_6$, $[F_6, F_1] = -\frac{\partial}{\partial x_1}$, and $\frac{\partial}{\partial x_1}$ is not in $\mathcal{G}$, $[Y_1, Y_1']$ is not in $\mathcal{J}^U$, so $\mathcal{J}^U$ is not stable under the Lie bracket.

5.5. How to use the necessary and sufficient conditions. Let us outline, as we did in §4.3 for FDE, the way to check in practice whether FDEM1 or FDEM2 is satisfied or not.

Let us first consider the case of FDEM1. Given a family of systems defined by certain F_0, $F_1, \dots, F_m$, Theorem 5.3 tells us that the systems of the family are FDEM1 if and only if one may find some vector fields $Y_1, \dots, Y_l$ satisfying (90), (91), and (92). If we try to go through the same process as we did in §4.3, we may of course define the Y_i's according to (76) because (90) implies (54). Since (91) is similar to (55), it may be translated into (79) if $\mathcal{G}$ has constant rank. The new feature is that (90) implies that the terms $\langle \eta_j, Y_i\rangle$ are all identically zero, so the first term in (79) vanishes. Since this was the only term involving the derivatives, (79) is *no longer a set of PDEs but simply a set of linear algebraic equations* in the b_i^j's. (90) and (92) can also be translated into linear algebraic equations in the b_i^j's. It is then fairly easy to give conditions under which this system of linear equations has some solutions, at least if its rank is constant. Theorem 5.9 in §5.6 gives this condition, using a dual characterization of existence of solutions for a linear system.

The case of FDEM2 is not as simple. (106) can still be translated, if both $\mathcal{G}$ and $\mathcal{G}^{\cap}$ have constant rank, into some linear algebraic equations in the components of Z_i, but we have to add some nonlinear, coupled differential conditions in these same components to account for the commutation relations (107). This does not allow such explicit conditions as in the case of FDEM1. Sections 5.8 and 5.7 review some cases in which explicit conditions may still be given because these nonlinear conditions are consequences of the linear ones.

5.6. An explicit characterization of local FDEM1. Here, we translate the conditions on each Y_i, which lies in $\mathcal{G}$, into some conditions on its coefficients as a linear combination of $F_1, \ldots, F_m$. As noticed in §5.5, these conditions do not involve derivatives of these coefficients.

PROPOSITION 5.8. *The family $\mathcal{S}$ is locally FDEM1 if and only if there exists, around $(\bar{p}, \bar{x})$, some C^∞ functions a_i^s $(i = 1 \ldots l,\ s = 1 \ldots m)$ such that*

$$\left[\frac{\partial}{\partial p_i}, F_k\right] - \sum_{s=1}^{m} a_i^s \,[F_k, F_s] \in \mathcal{G}, \qquad \begin{array}{l} k = 0, 1, \ldots, m, \\ i = 1, \ldots, l, \end{array} \tag{129}$$

$$\sum_{s=1}^{m} a_i^s \left[\frac{\partial}{\partial p_j}, F_s\right] \in \mathcal{G}, \qquad \begin{array}{l} i, j = 1, \ldots, l \\ i \neq j. \end{array} \tag{130}$$

It is globally FDEM1 if and only if there exists some C^∞ functions a_i^s $(i = 1 \ldots l,$ $s = 1 \ldots m)$ defined on $\mathbb{R}^l \times M^n$, satisfying (129) and (130) on $\mathbb{R}^l \times M^n$, and such that the vector fields $\frac{\partial}{\partial p_i} + \sum_{s=1}^{m} a_i^s F_s$ are complete.

Proof. Use Theorem 5.3, and translate (90) into the fact that Y_i may be written as $\sum_{s=1}^{m} a_i^s F_s$. (91) reads (129) and (92) reads (130). The terms that would involve some Lie derivatives of the functions a_i^s may be dropped because they are in $\mathcal{G}$. □

If both the distribution spanned by the module $\mathcal{G}$ and the distribution spanned by the module $\mathcal{L}$ of parameter-preserving vector fields on $\mathbb{R}^l \times M^n$ defined by

$$\mathcal{L} \triangleq \left\{ X \in \mathcal{G} \,/\, [X, F_k] \in \mathcal{G},\ k = 0, 1, \ldots, m, \text{ and } \left[\frac{\partial}{\partial p_i}, X\right] \in \mathcal{G},\ i = 1, \ldots, l \right\} \tag{131}$$

have constant rank, we may translate (129) and (130) into some scalar equations and give some conditions for existence of the solutions a_i^s.

THEOREM 5.9. *Let $\mathcal{S} = (F_o, F_1, \ldots, F_m)$ be a family with $\mathcal{G}$ and $\mathcal{L}$ of constant rank around $(\bar{p}, \bar{x})$. $\mathcal{S}$ is locally FDEM1 at $(\bar{p}, \bar{x})$ if and only if the following property holds on a certain neighborhood of $(\bar{p}, \bar{x})$. Any set of $m + 1 + l$ differential forms of degree 1 on $\mathbb{R}^l \times M^n$ $\Omega_o, \Omega_1, \ldots, \Omega_m, \Omega_{m+1}, \ldots, \Omega_{m+l}$ that all vanish on $\mathcal{H}$ (i.e., vanish on $\mathcal{G}$ and have no* dp*-component):*

$$\langle \Omega_k, \mathcal{H} \rangle = 0, \qquad k = 0, 1, \ldots, m + l, \tag{132}$$

and satisfy, for any $s = 1, \ldots, m$,

$$\sum_{k=0}^{m} \langle \Omega_k, [F_k, F_s] \rangle + \sum_{j=1}^{l} \left\langle \Omega_{m+j}, \left[\frac{\partial}{\partial p_j}, F_s\right] \right\rangle = 0, \tag{133}$$

also satisfy, for any $i = 1, \ldots, l$,

$$\sum_{k=0}^{m} \left\langle \Omega_k, \left[\frac{\partial}{\partial p_i}, F_k\right] \right\rangle = 0. \tag{134}$$

In the case $l = 1$ (scalar parameter p), Ω_{m+1} can a priori be chosen zero.

Remark 5. Instead of using vector fields and differential forms on $I\!R^l \times M^n$, this could be formulated with families of vector fields and differential forms on M^n. Suppose that the parameterized families of distribution $\mathcal{g}$ and

$$\mathcal{L} = \left\{ f \in \mathcal{g} \,/\, [f, f_k] \in \mathcal{g},\ k = 0, 1, \ldots, m, \text{ and } \frac{\partial f}{\partial p_i} \in \mathcal{g},\ i = 1, \ldots, l \right\}$$

are of constant rank around $(\bar{p}, \bar{x})$. $\mathcal{S}$ is locally FDEM1 at $(\bar{p}, \bar{x})$ if and only if, on a certain neighborhood of $(\bar{p}, \bar{x})$, any $m + 1 + l$ parameterized families of 1-forms ω_o, $\omega_1, \ldots, \omega_m$, $\omega_{m+1}, \ldots, \omega_{m+l}$ that all vanish on $\mathcal{g}$ and satisfy, for $s = 1, \ldots, m$,

$$\sum_{k=0}^{m} \langle \omega_k\, ,\, [f_k, f_s] \rangle + \sum_{j=1}^{l} \left\langle \omega_{m+j}\, ,\, \frac{\partial f_s}{\partial p_j} \right\rangle = 0$$

also satisfy, for $i = 1, \ldots, l$, $\sum_{k=0}^{m} \langle \omega_k\, ,\, \frac{\partial f_k}{\partial p_i} \rangle = 0$.

Proof of Theorem 5.9. Necessity is obvious from Proposition 5.8. Let us prove sufficiency. (135) and (136) are equivalent to the vector fields $[\frac{\partial}{\partial p_i}, F_k] - \sum_{s=1}^{m} a_i^s [F_k, F_s]$ and $\sum_{s=1}^{m} a_i^s [\frac{\partial}{\partial p_j}, F_s]$ belonging to $\mathcal{H}$, because these vector fields cannot have a nonzero p-component, so they are in $\mathcal{G}$ if and only if they are in $\mathcal{H}$. $\mathcal{H}^o$, the annihilator of $\mathcal{H}$ ($\mathcal{H}^o$ is the set of 1-forms that vanish on all the F_k's and have no dp_i component) has constant rank, say $n - m'$ for a certain integer m'. Let $\overline{\Omega}_{m'+1}, \ldots, \overline{\Omega}_n$ be locally a basis of $\mathcal{H}^o$. (129) and (130) are equivalent to

$$\text{(135)}\quad \sum_{s=1}^{m} \langle\, \overline{\Omega}_q,\ [\, F_k\, ,\, F_s\,]\, \rangle\, a_i^s = \left\langle\, \overline{\Omega}_q\, ,\ \left[\, \frac{\partial}{\partial p_i}\, ,\, F_k \right] \right\rangle, \qquad \begin{array}{l} q = m' + 1\, , \ldots,\, n, \\ k = 0,\, 1\, , \ldots,\, m, \\ i = 1, \ldots, l, \end{array}$$

$$\text{(136)}\qquad \sum_{s=1}^{m} \left\langle\, \overline{\Omega}_q\, ,\ \left[\, \frac{\partial}{\partial p_j}\, ,\, F_s \right] \right\rangle a_i^s = 0, \qquad \begin{array}{l} q = m' + 1\, , \ldots,\, n, \\ i, j = 1, \ldots, l,\ j \neq i. \end{array}$$

Let us fix i. (135)–(136) is a linear system in $a_i^1, \ldots, a_i^m$. The condition of the theorem states (the Ω_k's are linear combinations of $\overline{\Omega}_{m'+1}, \ldots, \overline{\Omega}_n$) that whenever a linear combination of the left-hand sides of this system is identically zero (i.e., independently of $a_i^1, \ldots, a_i^m$), the same linear combination of the right-hand sides is also zero. This is a classical characterization of the existence of solutions for linear systems.

In the case $l = 1$, there is no equation in (136), so that, in the linear combinations, there will be no form corresponding to this equation. □

5.7. Some cases in which FDEM1 and FDEM2 are equivalent. We now consider the two particular cases when the control distributions are known either to be integrable or to be independent of the parameter. $\mathcal{L}$ is still defined by (131).

THEOREM 5.10. *Suppose that the module $\mathcal{G}$ is invariant by $\frac{\partial}{\partial p_1}, \ldots, \frac{\partial}{\partial p_l}$ (i.e., the control module does not depend on the parameter). Let $\mathcal{K}$ be defined by*

$$\text{(137)}\qquad \mathcal{K} \ \triangleq\ \{ X \in \mathcal{G} \,/\, [X, \mathcal{G}] \subset \mathcal{G} \}.$$

FDEM1 and FDEM2 are equivalent and are both locally equivalent to

$$\text{(138)}\qquad \left[\frac{\partial}{\partial p_i}\, ,\, F_o \right], \in \mathcal{G}\ +\ [\, F_o\, ,\, \mathcal{K}], \quad i = 1, \ldots, l.$$

THEOREM 5.11. *Suppose that the module $\mathcal{G}$ is stable under the Lie bracket. Then FDEM1 and FDEM2 are equivalent, and they are both equivalent locally to*

$$\left[\frac{\partial}{\partial p_i}, \mathcal{G}\right] \subset \mathcal{G}, \quad i = 1, \ldots, l; \tag{139}$$

$$\left[\frac{\partial}{\partial p_i}, F_o\right] \in \mathcal{G} + [F_o, \mathcal{G}], \quad i = 1, \ldots, l. \tag{140}$$

Remark 6. This theorem gives, as a particular case, the result stated in [5], [4], saying that, if all the systems $\mathcal{S}_p$ are fully feedback linearizable and the dependence on p is linear, then a sufficient condition for the linearizing coordinates to satisfy a condition similar to the one φ satisfies in FDEM1 is that $\mathcal{g}$ be independent of p (this is (139)) and that

$$\frac{\partial f_o}{\partial p_i} \;\in\; \mathcal{g} + [f_o, \mathcal{g}], \quad i = 1, \ldots, l, \tag{141}$$

which is (140) (linearity in p implies that $\frac{\partial f_o}{\partial p_i}$ is a vector field independent of p).

Proof of Theorem 5.10. If $\mathcal{G}$ is invariant by $\frac{\partial}{\partial p_1}, \ldots, \frac{\partial}{\partial p_l}$, then, from the lemma given in the appendix, there exists an invertible matrix $\beta(p, x)$ such that $g(\bar{p}, x) = g(p, x)\beta(p, x)$. Then FDEM1 implies FDEM2 since (80) implies (81) with

$$v_2(p, q, x, \dot{p}) = \beta(q, x)\,\beta(p, x)^{-1}\,v_1(p, x, \dot{p}).$$

Since FDEM2 always implies FDEM1 (see Remark 2), this proves the equivalence between FDEM1 and FDEM2 in this case (FDEM1⇒FDEM2 can be seen as a consequence of Theorems 5.3 and 5.6 and the fact that, since $\mathcal{G}$ is invariant by $\frac{\partial}{\partial p_1}, \ldots, \frac{\partial}{\partial p_l}$, we have $\mathcal{G}^{\cap} = \mathcal{G}$). Let us use Theorem 5.3 to prove that they are equivalent to (138). Clearly (90) and (91) imply that Y_i is in $\mathcal{K}$ (because $[\frac{\partial}{\partial p_i}, F_k] \in \mathcal{G}$, $k = 1, \ldots, m$) and then imply (138) (using (91) for $k = 0$). Conversely, the fact that $\mathcal{G}$ is invariant by $\frac{\partial}{\partial p_1}, \ldots, \frac{\partial}{\partial p_l}$ implies that (92) and (91) for $k \neq 0$ are satisfied by any Y_i in $\mathcal{K}$. (138) exactly means the existence of a Y_i in $\mathcal{K}$ satisfying (91) for $k = 0$. □

Proof of Theorem 5.11. If $\mathcal{G}$ is a Lie algebra, FDEM1 implies (139)–(140) since (90), (91) and $[\mathcal{G}, \mathcal{G}] \subset \mathcal{G}$ obviously imply (139), i.e., invariance of $\mathcal{G}$ by $\frac{\partial}{\partial p_1}, \ldots, \frac{\partial}{\partial p_l}$ and, from Theorem 5.10, FDEM1 plus invariance of $\mathcal{G}$ imply (138), which is equivalent, since $\mathcal{K} = \mathcal{G}$, to (140).

If $\mathcal{G}$ is a Lie algebra, (139)–(140) imply FDEM2 from Theorem 5.10 because (139) means that $\mathcal{G}$ is invariant by $\frac{\partial}{\partial p_1}, \ldots, \frac{\partial}{\partial p_l}$ and, since $\mathcal{K} = \mathcal{G}$, (138) is equivalent to (140).

FDEM2 implies FDEM1 in general (see (82)). □

5.8. An explicit characterization of local FDEM2 in some particular cases. To simplify our discussion, we suppose here that $\mathcal{G}$ spans a distribution of constant rank; $\mathcal{G}^{\cap, U}$ does not depend on the open set U, so that we can write simply $\mathcal{G}^{\cap}$; and the distribution spanned by this module $\mathcal{G}^{\cap}$ also has constant rank.

We will not be able to give an explicit necessary and sufficient condition for FDEM2 in general. This paragraph only gives a characterization in the case when the parameter is one dimensional ($l = 1$) and in the case when either the rank of the distribution $\mathcal{G}^{\cap}$ or the difference between this rank and the rank of $\mathcal{G}$ is small.

If l is 1, Theorem 5.6 contains no commutation relation, and we may therefore write a condition for the set of linear algebraic equations defining the coefficients of Z_1 to have a solution.

THEOREM 5.12. *Let $\mathcal{S} = (F_o, F_1, \ldots, F_m)$ be a family of systems depending on one parameter only (i.e., $l = 1$). Suppose that $\mathcal{G}$, $\mathcal{G}^\cap$, and $\mathcal{G}^\cap \cap \mathcal{L}$ ($\mathcal{L}$ is defined in* (131)) *span some constant rank distributions around $(\bar{p}, \bar{x})$. In particular, locally, some vector fields $G_1^\cap, \ldots, G_r^\cap$ span $\mathcal{G}^\cap$:*

$$\mathcal{G}^\cap = \text{Span}\,\{ G_1^\cap, \ldots, G_r^\cap \}. \tag{142}$$

The family $\mathcal{S}$ is locally FDEM2 at $(\bar{p}, \bar{x})$ if and only if the following property holds on a certain neighborhood of $(\bar{p}, \bar{x})$: any $m+1$ 1-forms on $I\!R^l \times M^n$ $\Omega_o, \Omega_1, \ldots, \Omega_m$ that vanish on $\mathcal{H}$ (i.e., vanish on $\mathcal{G}$ and have no dp*-component) and satisfy*

$$\sum_{k=0}^{m} \langle\, \Omega_k\,,\, [F_k, G_s^\cap]\,\rangle = 0, \qquad s = 1, \ldots, r, \tag{143}$$

also satisfy

$$\sum_{k=0}^{m} \left\langle\, \Omega_k\,,\, \left[\frac{\partial}{\partial p_i}, F_k\right] \right\rangle = 0, \quad i = 1. \tag{144}$$

Proof of Theorem 5.12. This condition is equivalent to the existence of some $\mathcal{C}^\infty$ functions a^s such that

$$\left[\frac{\partial}{\partial p}\,,\, F_k\right] - \sum_{s=1}^{m} a^s [\, F_k\,,\, G_s^\cap\,] \in \mathcal{G}, \qquad k = 0, 1, \ldots, m \tag{145}$$

(see the proof of Theorem 5.9), which is in turn equivalent to the conditions in Theorem 5.6, since (107) is automatically satisfied when $l = 1$. □

When $l > 1$, no such theorem holds in general, since the commutation relations required in Theorem 5.6 cannot be removed; see counterexample (118)–(119) in §5.4. However, in some situations, as illustrated by the following theorem, these commutation conditions may be implied by the others.

THEOREM 5.13. *Let $\mathcal{S} = (F_o, F_1, \ldots, F_m)$ be a family of systems. Suppose that $\mathcal{G}$, $\mathcal{G}^\cap$, $\mathcal{L}$, and $\mathcal{G}^\cap \cap \mathcal{L}$ ($\mathcal{L}$ is defined in* (131)) *span some constant rank distributions around $(\bar{p}, \bar{x})$. Suppose also that*

- *either $\mathcal{G}^\cap$ is integrable,*
- *or* Rank $\mathcal{G}^\cap \leq 3$,
- *or* Rank $\mathcal{G}$ − Rank $\mathcal{G}^\cap \leq 1$.

Then one can remove the commutation relations in Theorem 5.6. *As a consequence, in these same cases, Theorem* 5.12 *is valid even for $l > 1$.*

Proof. The last line is a consequence of the first part because the conditions in Theorem 5.12 are equivalent to the condition (105)–(106) (without the commutation relation (107)) in Theorem 5.6. Let us prove the first part of the theorem. The conditions of Theorem 5.6 obviously remain necessary if we *remove* (107). It remains to prove that they remain sufficient in the cases quoted in the theorem.

By Theorem 5.7, it is enough to prove that the module $\mathcal{J}^U$ defined by (115) is stable under the Lie bracket. Let us denote by $\mathcal{J}^U$ as well the distribution spanned by the module $\mathcal{J}^U$. Since (105) and (106) imply that $\frac{\partial}{\partial p_i} + Z_i \in \mathcal{J}^U$ and $\mathcal{L} \cap \mathcal{G}^\cap$

contains exactly all the vector fields in $\mathcal{J}$ with a zero p-component, we have $\mathcal{J}^U = (\mathcal{L} \cap \mathcal{G}^\cap) \oplus \mathrm{Span}\{\frac{\partial}{\partial p_1} + Z_1, \ldots, \frac{\partial}{\partial p_l} + Z_l\}$ and therefore $\mathcal{J}^U$ does not depend on U—we shall write simply $\mathcal{J}$—and has constant rank $l + \mathrm{Rank}\,\mathcal{L} \cap \mathcal{G}^\cap$. All we have to check is that the distribution $\mathcal{J}$ is involutive. This is the case in all but one of the cases considered in the theorem:

- If $\mathcal{G}^\cap$ itself is involutive, then $\mathcal{J}$ is equal to $\mathcal{G}^\cap$.
- If $\mathcal{G}^\cap$ has rank 1, we are in the previous case.
- If $\mathcal{G}^\cap$ has rank 2, then the rank of $\mathcal{J}$ is either 0 or 1 or 2. In the two first cases, it is obviously involutive and in the last one, it is involutive as well since then $\mathcal{J} = \mathcal{G}^\cap$, so that $[\mathcal{G}^\cap, \mathcal{G}^\cap] \subset \mathcal{G}$, which implies $[\mathcal{G}^\cap, \mathcal{G}^\cap] \subset \mathcal{G}^\cap$ since $[\mathcal{G}^\cap, \mathcal{G}^\cap]$ is invariant under the $\frac{\partial}{\partial p_i}$'s and $\mathcal{G}^\cap$ is the largest subdistribution of $\mathcal{G}$ invariant by the $\frac{\partial}{\partial p_i}$'s.
- If $\mathcal{G}^\cap$ has rank 3, then the rank of $\mathcal{J}$ is either 0 or 1 or 2 or 3. We may conclude as above if it is 0, 1, or 3 (if it is 3, $\mathcal{J} = \mathcal{G}^\cap$). It cannot be 2 because if G_1, G_2, G_3 are three vector fields locally spanning $\mathcal{G}^\cap$, then $\lambda_1 G_1 + \lambda_2 G_2 + \lambda_3 G_3$ is in $\mathcal{J}$ if and only if $(\lambda_1, \lambda_2, \lambda_3)$ is in the kernel of a certain number of skew symmetric matrices; the rank of the corresponding system in $\lambda_1, \lambda_2, \lambda_3$ cannot be 1 because the ranks of all these matrices are even.

In the last case ($\mathrm{Rank}\,\mathcal{G} - \mathrm{Rank}\,\mathcal{G}^\cap \leq 1$), $\mathcal{J}$ is not involutive in general. If $\mathrm{Rank}\,\mathcal{G} - \mathrm{Rank}\,\mathcal{G}^\cap = 0$, then $\mathcal{G} = \mathcal{G}^\cap$, so $\mathcal{G}$ is invariant under the vector fields $\frac{\partial}{\partial p_i}$ and one may conclude using Theorem 5.10. If $\mathrm{Rank}\,\mathcal{G} - \mathrm{Rank}\,\mathcal{G}^\cap = 1$, then it is possible to find, locally, some vector fields $G_1, \ldots, G_m$ such that

$$\begin{aligned} \mathcal{G} &= \mathrm{Span}\,\{G_1, \ldots, G_m\}, \\ \mathcal{G}^\cap &= \mathrm{Span}\,\{G_1, \ldots, G_{m-1}\}, \\ \mathcal{J} &= \mathrm{Span}\,\{G_1, \ldots, G_r\} \end{aligned} \tag{146}$$

where $r \leq m - 1$. Consider the module

$$\mathcal{J}_1 = \{\, X \in \mathcal{G}^\cap \,/\, [X, \mathcal{G}] \subset \mathcal{G} \,\}. \tag{147}$$

It is obvious from the definitions that $\mathcal{J}_1 \subset \mathcal{J}$. We distinguish two cases depending on whether this inclusion is strict or not:

- If $\mathcal{J}_1 = \mathcal{J}$, then $\mathcal{J}$ is involutive: for $j \leq r$ and $k \leq r$, let us prove that $[G_j, G_k]$ is in $\mathcal{J}$. First of all it is in $\mathcal{G}$ because G_j is in $\mathcal{J}$ and G_k is in $\mathcal{G}$ (see definition of $\mathcal{J}$). Now, $[\frac{\partial}{\partial p_i}, [G_j, G_k]]$ is in $\mathcal{G}$ because, from Jacobi identity, $[\frac{\partial}{\partial p_i}, [G_j, G_k]] = [G_j, [\frac{\partial}{\partial p_i}, G_k]] + [G_k, [\frac{\partial}{\partial p_i}, G_j]]$ where $[\frac{\partial}{\partial p_i}, G_k]$ and $[\frac{\partial}{\partial p_i}, G_j]$ are in $\mathcal{G}^\cap$ (G_k and G_j are in $\mathcal{G}^\cap$, which is invariant under $\frac{\partial}{\partial p_i}$). On the other hand, since $[G_j, G_k]$ is in $\mathcal{G}$, there exist some functions $\lambda_{kl}^1, \ldots, \lambda_{kl}^m$ such that

$$[G_j, G_k] = \sum_{s=1}^{m} \lambda_{kl}^s G_s, \tag{148}$$

and therefore

$$\left[\frac{\partial}{\partial p_i}, [G_j, G_k]\right] = \sum_{s=1}^{m} \frac{\partial \lambda_{kl}^s}{\partial p_i} G_s + \sum_{s=1}^{m} \lambda_{kl}^s \left[\frac{\partial}{\partial p_i}, G_s\right] \in \mathcal{G}. \tag{149}$$

This implies that the coefficient λ_{kl}^m is identically zero because $[\frac{\partial}{\partial p_i}, G_m] \notin \mathcal{G}$ (otherwise $\mathcal{G} = \mathcal{G}^\cap$) and all the other terms are in $\mathcal{G}$. This proves that $[G_j, G_k] \in \mathcal{G}^\cap$. We have now to check that $[[G_j, G_k], \mathcal{G}^\cap] \subset \mathcal{G}$: let Y be in $\mathcal{G}^\cap$, we have (Jacobi identity)

$$[\,[G_j, G_k]\,, Y\,] = [G_j, [G_k, Y]] \; - \; [G_k, [G_j, Y]],$$

which is in $\mathcal{G}$ because $[G_k, Y]$ and $[G_j, Y]$ are in $\mathcal{G}$ and G_j and G_k are in $\mathcal{J}_1 = \mathcal{J}$.

- If $\mathcal{J} \neq \mathcal{J}_1$ we shall use the following fact: condition 2 in Theorem 5.7 is also sufficient for FDEM2 if we replace the module $\mathcal{J}^U$ with $\mathcal{M}_1$:

$$\mathcal{M}_1 = \{ X \in \mathcal{H}^\cap / [X, \mathcal{H}] \subset \mathcal{H} \}. \tag{150}$$

The proof is simpler than the proof of 2 in Theorem 5.7, replacing $\mathcal{M}$ with $\mathcal{M}_1$. Here, $\mathcal{M}_1$ is finitely generated because $\mathcal{M}_1 = \mathcal{L} \oplus \text{Span}\{\frac{\partial}{\partial p_1} + Y_1, \ldots, \frac{\partial}{\partial p_l} + Y_l\}$, where the vector fields Y_i are those satisfying (105) and (106), and it is stable under the Lie bracket because the module

$$\mathcal{M}_2 = \{ X \in \mathcal{H} / [X, \mathcal{H}] \subset \mathcal{H} \} \tag{151}$$

is naturally stable under the Lie bracket; to conclude the proof, we establish that $\mathcal{M}_1 = \mathcal{M}_2$. Since the inclusion $\mathcal{J} \subset \mathcal{J}_1$ is strict, there is a nonzero vector field Y_o in $\mathcal{J}$ that is not in $\mathcal{J}_1$, i.e.,

$$Y_o \in \mathcal{J} \ ; \ [Y_o, G_m] \notin \mathcal{G}. \tag{152}$$

Then, let X be in $\mathcal{M}_2$; since it is in $\mathcal{H}$, it may be written

$$X = \sum_1^l \mu_i \frac{\partial}{\partial p_i} + \sum_1^m \nu_k G_k, \tag{153}$$

and we have

$$[X, Y_o] = \left[\sum_{i=1}^l \mu_i \frac{\partial}{\partial p_i}, Y_o \right] + \sum_{k=1}^m \nu_k [G_k, Y_o] - \sum_{k=1}^m (L_{Y_o} \nu_k) G_k$$

where the left-hand side is in $\mathcal{H}$ because X is in $\mathcal{M}_2$ and Y_o is in $\mathcal{J} \subset \mathcal{H}$, and the only term in the right-hand side that is not naturally in $\mathcal{H}$ is $\nu_m[G_m, Y_o]$; from (152), this implies than the coefficient ν_m is zero and, from (153), that X is in $\mathcal{M}_1$. We have proved that $\mathcal{M}_2 \subset \mathcal{M}_1$; $\mathcal{M}_1 \subset \mathcal{M}_2$ is obvious from (150) and (151). □

5.9. Explicit calculation of the transformations on an example. Here, we continue to work out example (118)–(119) and calculate explicitly the transformations. As seen previously, the family of systems described in (118)–(119) is not FDEM2. It is, however, globally FDEM1 from Theorem 5.3 because the Y_i's computed in (126)–(127) are in $\mathcal{G}^\cap$ and therefore satisfy (92). Let us compute the corresponding α and β. The flows ϕ_1 and ϕ_2 of vector fields $\frac{\partial}{\partial p_1} + Y_1$ and $\frac{\partial}{\partial p_2} + Y_2$ are

$$\phi_1^t(p,x) = (p_1 + t,\ p_2,\ x_1,\ x_2 - t x_6,\ x_3,\ x_4,\ x_5 + (p_1 + x_7)t,\ x_6,\ x_7 - t,\ x_8), \tag{154}$$

$$\begin{aligned} \phi_2^t(p,x) = (&p_1,\ p_2 + t,\ x_1 - t x_5 (p_2 + x_8),\ x_2,\ x_3 + t x_5,\ x_4 + t x_7, \\ &x_5,\ x_6 + t(p_2 + x_8),\ x_7,\ x_8 - t\ ), \end{aligned} \tag{155}$$

so computing φ according to (56), choosing $\bar{p} = 0$, gives

$$\begin{aligned} \varphi(p,x) = (\,&x_1 + p_2 x_5 (p_2 + x_8)\,,\ x_2 + p_1 [x_6 - p_2(p_2 + x_8)]\,,\ x_3 - p_2 x_5\,, \\ &x_4 + p_2 x_7\,,\ x_5 - p_1(p_1 + x_7)\,,\ x_6 - p_2(p_2 + x_8)\,,\ x_7 + p_1\,,\ x_8 + p_2). \end{aligned} \tag{156}$$

This φ transforms the family $\mathcal{S}$ into $\widetilde{\mathcal{S}}$ described by

$$\begin{aligned}
\dot\xi_1 &= p_2(\xi_5+p_1\xi_7)u_1-(\xi_5+p_1\xi_7)u_3+p_2\xi_8u_4 \\
&\quad -\xi_7\xi_8u_5-\xi_8u_6-\xi_7u_7, \\
\dot\xi_2 &= u_7-p_1p_2u_1+(\xi_6+p_2\xi_8)u_2-p_1u_3, \\
\dot\xi_3 &= u_6-(\xi_5+p_1\xi_7)u_1-p_2u_4, \\
\dot\xi_4 &= u_5-(\xi_7-p_1)u_1,
\end{aligned}
\qquad
\begin{aligned}
\dot\xi_5 &= u_4-p_1u_2, \\
\dot\xi_6 &= u_3-p_2u_1, \\
\dot\xi_7 &= u_2, \\
\dot\xi_8 &= u_1
\end{aligned}$$

(with $\xi=\varphi(p,x)$). $\widetilde{\mathcal{S}}$ is FE, the feedback defined by the following α and β, turning it into a constant family:

$$\alpha(p,\xi)=0;\quad \beta(p,\xi)=\begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ p_2 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & p_1 & 0 & 1 & 0 & 0 & 0 \\ p_1 & p_2 & 0 & 0 & 1 & 0 & 0 \\ p_1\xi_7 & p_1p_2 & 0 & p_2 & 0 & 1 & 0 \\ 0 & -p_2\xi_8 & -p_1 & 0 & 0 & 0 & 1 \end{pmatrix}. \tag{157}$$

A simple computation allows one to compute explicitly $\frac{\partial\varphi}{\partial x}^{-1}\frac{\partial\varphi}{\partial p_1}$ and $\frac{\partial\varphi}{\partial x}^{-1}\frac{\partial\varphi}{\partial p_2}$ and to express them as a linear combination of $F_1\ldots F_7$, i.e., φ, defined by (156), that satisfies (80) with $v_1(p,\dot p,x)$ given by

$$v_1(p,\dot p,x)=\begin{pmatrix} -\dot p_2 \\ -\dot p_1 \\ (p_2+x_8)\dot p_2 \\ (p_1+x_7)(\dot p_1+\dot p_2) \\ \dot p_1 \\ -(p_1+x_7)\dot p_1 \\ -(p_2+x_8)\dot p_1 \end{pmatrix}; \tag{158}$$

hence, FDEM1 is met.

6. The case of nonaffine systems.

The previous sections are devoted only to systems that are affine in the control u. We now consider more general systems. Being pure feedback equivalent (FE) or feedback and diffeomorphism equivalence (FDE) is quite meaningful for a parameterized family of nonaffine systems. We give some generalizations of the main results on FE and FDE.

Let us consider the family Σ of systems parameterized by $p\in\mathbb{R}^l$ where the system corresponding to a certain value of p is Σ_p described by

$$\dot x = f(p,x,u). \tag{159}$$

Equivalence (FDE or FE) will have exactly the same meaning as for affine systems, except we cannot restrict the feedback transformation to be affine in the control, and the diffeomorphism acts on the "multi-vector field f" instead of acting separately on the vector fields $f_0,\ldots,f_m$ in the affine-in-the-control case.

DEFINITION 6.1. • *A family Σ of systems is* constant *on an open subset U of $\mathbb{R}^l\times M^n$ if, for all x, p^1, and p^2 such that (p^1,x) and (p^2,x) are in U and any u in $\mathbb{R}^m$ we have*

$$f(p^1,x,u)=f(p^2,x,u). \tag{160}$$

- *A family* Σ *is* FE (pure feedback equivalent) *on an open subset* U *of* $\mathbb{R}^l \times M^n$ *if there exists a* C^∞ *map*

$$\gamma : \; U \times \mathbb{R}^m \longrightarrow \mathbb{R}^m,$$

such that for any (p, x) *the map* $u \mapsto \gamma(p, x, u)$ *is a diffeomorphism of* $\mathbb{R}^m$ *and the family* $\widetilde{\Sigma}$ *defined by*

$$\widetilde{f}(p, x, v) = f(p, x, \gamma(p, x, v)) \tag{161}$$

is constant on U. Σ *is locally FE at* $(\bar{p}, \bar{x})$ *if it is FE on a certain open neighborhood* U *of* $(\bar{p}, \bar{x})$ *in* $\mathbb{R}^l \times M^n$. *It is globally FE if it is FE on* $\mathbb{R}^l \times M^n$.

- *A family* Σ *is locally* FDE (feedback and diffeomorphism equivalent) *at* $(\bar{p}, \bar{x})$ *if and only if there exists a neighborhood* U *of* $(\bar{p}, \bar{x})$ *and a map*

$$\psi : \; U \times \mathbb{R}^m \longrightarrow \mathbb{R}^m$$

such that for any (p, x) *the map* $x \mapsto \psi(p, x, u)$ *is a diffeomorphism from* $U \times \mathbb{R}^l$ *to its image and the family* $\widetilde{\Sigma}$, *defined in* $\psi(U \times \mathbb{R}^l)$ *by*

$$\widetilde{f} = \psi_* f, \tag{162}$$

is locally FE at $(\bar{p}, \varphi(\bar{p}, \bar{x}))$. *It is globally FDE if and only if* ψ *is defined all over* $\mathbb{R}^l \times M^n$ *and the family* $\widetilde{\Sigma}$ *defined on* $\mathbb{R}^l \times M^n$ *by* (162) *is globally FE.*

To a given (non-necessarily-affine) parameterized family of systems Σ on M^n with m controls, we may associate a parameterized family of affine systems, just adding an integrator to each control, so that the new controls appear obviously linearly. Precisely, to Σ, we associate $\mathcal{S} = \mathcal{A}(\Sigma)$, where the system $\mathcal{S}_p$ is defined by

$$\left\{ \begin{array}{rcl} \dot{x} & = & f(p, x, z), \\ \dot{z} & = & v. \end{array} \right. \tag{163}$$

Equivalently, the corresponding parameter-preserving family of systems on $\mathbb{R}^l \times M^n \times \mathbb{R}^m$ is defined by

$$\begin{array}{l} F_0(p, X) = (0, f(p, X), 0), \\ F_1 = \frac{\partial}{\partial z_1}, \ldots, F_m = \frac{\partial}{\partial z_m}. \end{array} \tag{164}$$

We have the following obvious properties as a consequence of the fact that $\mathcal{A}(\Sigma)$ is a particular family of systems (control vector fields independent of the parameter; see also Theorem 5.10).

PROPOSITION 6.2. *The family* $\mathcal{S} = \mathcal{A}(\Sigma)$ (*for any family* Σ) *is constant* (*locally/globally*) *if and only if it is FE* (*locally/globally*). *The family* Σ *is constant* (*locally/globally*) *if and only if* $\mathcal{S} = \mathcal{A}(\Sigma)$ *is constant or FE* (*locally/globally*). *The family* $\mathcal{S} = \mathcal{A}(\Sigma)$ (*for any family* Σ) *is FDEM*1 (*locally/globally*) *if and only if it is FDEM*2 (*locally/globally*).

It has already been noticed, in [3] for example, that the classification of general control systems is equivalent to the classification of affine-in-the-control systems with a control distribution of constant rank. This is the same for *families* of systems. More precisely, we also have the following two properties, the first of which is the "parametric" counterpart of the remark made in [3] and the second of which relates the property FE for a nonaffine family Σ to the notion of feedback equivalence with matching that we have developed for affine-in-the-control systems. The (elementary) proof of these propositions is given further.

PROPOSITION 6.3. *The family Σ is FDE (locally/globally) if and only if the family $\mathcal{S} = \mathcal{A}(\Sigma)$ is FDE (locally/globally).*

PROPOSITION 6.4. *The family Σ is FE (locally/globally) if and only if the family $\mathcal{S} = \mathcal{A}(\Sigma)$ is FDEM1 or FDEM2 (locally/globally).*

For the sake of ease of notation, we define F such that, for any u, $(p,x) \mapsto F(p,x,u)$ is the parameter-preserving vector field on $\mathbb{R}^l \times M^n$ (as intended in §1) defined by $F(p,x,u) = (0, f(p,x,u))$. Since for a fixed (p,x) $F(p,x,u)$ stays in the same tangent space, the "derivative" $\frac{\partial F}{\partial u_k}$ can be defined in the same way as the derivative with respect to parameters of a parameterized family of vector fields (see §2.4); $\frac{\partial F}{\partial u_k}$ is, for any fixed u, a parameter-preserving vector field on $\mathbb{R}^l \times M^n$.

The following two theorems give some characterizations of FE and FDE for a family of nonaffine systems. They are easy consequences of Theorems 4.2 and 5.3 and the above propositions. The proof of Theorem 6.6 is omitted (use Theorem 4.2 with $Y_i = (X_i, \sum_{j=1}^k a_i^j \frac{\partial}{\partial z_j})$).

THEOREM 6.5. *The family of systems Σ is locally FE at $(\bar{p},\bar{x})$ if the following relation is true for any u in $\mathbb{R}^m$ and for (p,x) in a neighborhood of $(\bar{p},\bar{x})$ (resp., if it is true for any u in $\mathbb{R}^m$ and any (p,x) in $\mathbb{R}^l \times M^n$):*

$$\left[\frac{\partial}{\partial p_i}, F\right] \in \operatorname{Span}\left\{\frac{\partial F}{\partial u_1}, \dots, \frac{\partial F}{\partial u_m}\right\}, \quad i = 1, \dots, l, \tag{165}$$

where Span$\{\frac{\partial F}{\partial u_1}, \dots, \frac{\partial F}{\partial u_m}\}$ *denotes the module over smooth functions of (p,x,u) generated by $\frac{\partial F}{\partial u_1}, \dots, \frac{\partial F}{\partial u_m}$. It is globally FE if relation* (165) *is true everywhere and if it is possible to chose the coefficients $a_i^1, \dots, a_i^m$ in the module such that the vector fields $\frac{\partial}{\partial p_i} + a_i^1 \frac{\partial}{\partial z_1} + a_i^m \frac{\partial}{\partial z_m}$ are complete.*

Proof of Theorem 6.5. From Proposition 6.4, all we have to prove is that the given condition is necessary and sufficient for $\mathcal{S} = \mathcal{A}(\Sigma)$ to be locally/globally FDEM1. For the family of systems $\mathcal{S}$, the module $\mathcal{G}$ is that generated by the vector fields $\frac{\partial}{\partial z_k}$. From Theorem 5.3, a condition for $\mathcal{S}$ to be FDEM1 is that there exists some functions a_i^j such that

$$\left[\frac{\partial}{\partial p_i} + a_i^1 \frac{\partial}{\partial z_1} + \cdots + a_i^m \frac{\partial}{\partial z_m}, F_k\right] \in \operatorname{Span}\left\{\frac{\partial}{\partial z_1}, \dots, \frac{\partial}{\partial z_m}\right\}, \tag{166}$$

where the F_k's are those defined in (164). The relation for $k = 1, \dots, m$ are automatically satisfied. The relation for $k = 0$ is equivalent to the left-hand side having a zero x-component; it is simple to prove that this is exactly equivalent to (165) (the a_i^j's are the coefficients in the module Span$\{\frac{\partial F}{\partial u_1}, \dots, \frac{\partial F}{\partial u_m}\}$). The completeness condition for global FE is exactly that for global FDEM1 in Theorem 5.3. □

THEOREM 6.6. *The family of systems Σ is locally FDE at $(\bar{p},\bar{x})$ if there exists some parameter-preserving vector fields $X_1, \dots, X_l$ on a neighborhood of $(\bar{p},\bar{x})$ in $\mathbb{R}^l \times M^n$, such that the following relation is true for any u in $\mathbb{R}^m$ and for (p,x) in a neighborhood of $(\bar{p},\bar{x})$:*

$$\left[\frac{\partial}{\partial p_i} + X_i, F\right] \in \operatorname{Span}\left\{\frac{\partial F}{\partial u_1}, \dots, \frac{\partial F}{\partial u_m}\right\}, \quad i = 1, \dots, l. \tag{167}$$

It is globally FDE if the parameter-preserving vector fields $X_1, \dots, X_l$ are defined all over $\mathbb{R}^l \times M^n$, (167) *is true for any u in $\mathbb{R}^m$ and any (p,x) in $\mathbb{R}^l \times M^n$, and it is* *possible to choose the coefficients* $a_i^1, \ldots, a_i^m$ *in the module such that the vector fields* $\frac{\partial}{\partial p_i} + X_i + a_i^1 \frac{\partial}{\partial z_1} + a_i^m \frac{\partial}{\partial z_m}$ *are complete.*

Remarks. 1. Note that X_i does not depend on u: relation (167) has to be satisfied for any u with the same X_i.

2. Of course, Theorem 4.2 is a consequence of Theorem 6.6 in the particular case where the systems are affine in the control, i.e.,

$$f(p, x, u) \quad = \quad f_o(p, x) + u_1 f_1(p, x) + \cdots + u_m f_m(p, x)$$

since in this case $\frac{\partial F}{\partial u_k}(p, x, u)$ is $F_i(p, x)$, and the fact that (167) is true for any u implies it is true separately for all the F_k's, which are independent of u. Therefore (55) is satisfied. The completeness conditions do not involve the $\frac{\partial}{\partial z_k}$ parts of the vector fields because it is always possible to choose the coefficients linear in z in the affine case (see the lemma given in the appendix).

Proof of Proposition 6.4. If Σ is (locally/globally) FE and γ is the corresponding family of feedback transformations, let us define φ by

$$\varphi(p, X) = \varphi(p, x, z) = \left(x \, , \, \gamma^{-1}(p, x, z) \right) , \tag{168}$$

where γ^{-1} means inverting γ with respect to z only:

$$\gamma^{-1}(\, p \, , \, x \, , \, \gamma(p, x, z) \,) = \gamma(\, p \, , \, x \, , \, \gamma^{-1}(p, x, z) \,) \equiv z. \tag{169}$$

The family of diffeomorphisms φ transforms the family of systems $\mathcal{S}$, described by (163), into

$$\left\{ \begin{array}{lcl} \dot{\xi} & = & f(p, \xi, \gamma(p, \xi, \zeta)), \\ \dot{\zeta} & = & \sum_{k=1}^{m} \frac{\partial \gamma^{-1}}{\partial z_k}(p, x, z) \, v_k \end{array} \right. \tag{170}$$

($\xi = x$, $\zeta = \gamma^{-1}(p, x, z)$), which is FE (locally/globally) from Theorem 3.4 since by assumption $f(p, \xi, \gamma(p, \xi, \zeta))$ does not depend on p, and the module $\widetilde{\mathcal{G}}$ is spanned by $\frac{\partial}{\partial z_1} \cdots \frac{\partial}{\partial z_m}$ because the matrix of the $\frac{\partial \gamma^{-1}}{\partial z_k}(p, x, z)$'s is invertible (for $z \mapsto \gamma^{-1}(p, x, z)$ is a diffeomorphism). In addition, φ satisfies the additional property (80) required for FDEM1, simply taking

$$v_1(p, x, z, \dot{p}) = \beta(p, x, z) \frac{\partial \gamma^{-1}}{\partial p}(p, x, z) \dot{p}$$

where $\beta(p, x, z)$ is the inverse of the matrix whose kth column is $\frac{\partial \gamma^{-1}}{\partial z_k}(p, x, z)$.

Conversely, if $\mathcal{S}$ is (locally/globally) FDEM1, let φ be a family of diffeomorphisms, transforming $\mathcal{S}$ into the FE family $\widetilde{\mathcal{S}}$. Since the affine-in-the-control family $\widetilde{\mathcal{S}}$ is FE, we have, from Theorem 3.4 and (164),

$$\varphi_* \frac{\partial}{\partial z_k} \in \operatorname{Span} \left\{ \frac{\partial}{\partial z_1}, \ldots, \frac{\partial}{\partial z_m} \right\}, \quad k = 1, \ldots, m \ , \tag{171}$$

which implies that the x component of φ does not depend on z. In addition, φ satisfies the property (80), which implies

$$\frac{\partial \varphi}{\partial z_k} \quad \in \quad \operatorname{Span} \left\{ \frac{\partial}{\partial z_1}, \ldots, \frac{\partial}{\partial z_m} \right\}, \quad k = 1, \ldots, m, \tag{172}$$

i.e., that the x component of φ does not depend on p either. Altogether, we have

$$\varphi(p,X)=\varphi(p,x,z)=(\,\varphi_1(x)\,,\,\varphi_2(p,x,z)\,)\,. \tag{173}$$

Actually, we may suppose that $\varphi_1(x)=x$:

$$\varphi(p,x,z)=(\,x\,,\,\varphi_2(p,x,z)\,) \tag{174}$$

because transforming it by the diffeomorphism $(x,z)\mapsto(\varphi_1(x),z)$, independent of p, does not change the fact that $\widetilde{\mathcal{S}}$ is FE. Let us then define γ by $\gamma(p,x,v)=\varphi_2^{-1}(p,x,v)$, where φ_2^{-1} means inverting φ_2 with respect to v (or z) only. The family of feedback transformations γ transforms the system $\widetilde{\sigma}$ that is constant because $\widetilde{\mathcal{S}}=\mathcal{A}(\widetilde{\sigma})$ and $\widetilde{\mathcal{S}}$ is FE. □

Proof of Proposition 6.3. If Σ is (locally/globally) FDE, let ψ be the corresponding family of diffeomorphisms transforming Σ into an FE family and γ be the family of feedback transformations transforming this FE family into a constant family. Let us define φ by

$$\varphi(p,X)=\varphi(p,x,z)=\left(\,\psi(p,x)\,,\,\gamma^{-1}(\psi(p,x),z)\,\right), \tag{175}$$

where γ^{-1} means inverting with respect to z for any (p,x). This does transform $\mathcal{S}$ into a family that is FE (if $(\xi,\zeta)=\varphi(p,x,z)$, $\dot{\xi}$ does not depend on p, as a function of p,ξ,ζ, and $\mathcal{G}$ is spanned by $\frac{\partial}{\partial z_1}\cdots\frac{\partial}{\partial z_m}$).

Conversely, if $\mathcal{S}$ is (locally/globally) FDE and φ is the corresponding family of feedback transformations, let us define φ_1 and φ_2 as the x and z components of φ. Since the affine-in-the-control family $\widetilde{\Sigma}$ is FE, we have, from Theorem 3.4 and (164),

$$\varphi_*\frac{\partial}{\partial z_k}\ \in\ \operatorname{Span}\left\{\frac{\partial}{\partial z_1},\ldots,\frac{\partial}{\partial z_m}\right\},\qquad k=1,\ldots,m, \tag{176}$$

which implies that φ_1 does not depend on z:

$$\varphi(p,X)=\varphi(p,x,z)=(\,\varphi_1(p,x)\,,\,\varphi_2(p,x,z)\,)\,. \tag{177}$$

It is clear that φ being a diffeomorphism implies that $x\mapsto\varphi_1(p,x)$ is a diffeomorphism for any p and $z\mapsto\varphi_2(p,x,z)$ is a diffeomorphism for any (p,x). Let us define ψ and γ by

$$\begin{array}{rcl}\psi(p,x)&=&\varphi_1(p,x),\\ \gamma(p,x,z)&=&\varphi_2^{-1}(p,x,z).\end{array} \tag{178}$$

This transforms Σ into a family $\widetilde{\Sigma}$ that is constant because φ transforms $\mathcal{S}$ into $\widetilde{\mathcal{S}}$ which is FE, and whose "$\dot{x}$" part is therefore independent of p. □

7. Conclusion. We have given some general conditions for the systems of a smoothly parameterized family to be equivalent via state feedback and diffeomorphism transformations, and the interest of the method used here is that it is constructive: under these conditions, we actually compute the transformations.

Among others, the interest of the problem considered here is to allow, if it can be solved, to control the systems of the family "in the same way": since all the systems are equivalent to a single one, it is possible to design a controller for this system and bring it back to the original systems of the family through the transformations.

The case where it is possible to choose a family of diffeomorphisms meeting some so-called "matching assumptions" has been studied in details. As mentioned above, the motivation for this study comes from adaptive nonlinear control where these conditions play an important role; they are a generalization of those introduced in the case of feedback linearizable systems; see [5], [4]. It turns out that our conditions are more explicit for feedback and diffeomorphism satisfying these matching conditions. The geometric conditions given in [5], [4] are a particular case of these.

Of course, feedback and diffeomorphism equivalence plays a very important role of its own in control theory. The point of view of parameterized families is rather unusual, and it facilitates the rather natural derivation of some global results. This has been developed in [7].

Appendix: A technical lemma. LEMMA. *Let $\mathcal{D}$ be a finitely generated module of $\mathcal{C}^\infty$ vector fields on a $\mathcal{C}^\infty$ manifold $\mathcal{N}$ and X and F be two $\mathcal{C}^\infty$ vector fields on $\mathcal{N}$. Let $G_1, \dots, G_s$ be any system of generators of $\mathcal{D}$:*

$$\mathcal{D} = \operatorname{Span}\{G_1, \dots, G_s\}. \tag{179}$$

Also let $\phi_X : J \times U \to \mathcal{N}$ be any partial flow of X, i.e., a $\mathcal{C}^\infty$ mapping ϕ_X, J an open interval of $I\!R$ containing 0, and U an open subset of $\mathcal{N}$, such that (we will often write $\phi_X^t(x)$ instead of $\phi_X(t,x)$):

$$\phi_X(0,x) = x \quad \textit{for all } x \in U,$$
$$\frac{\partial \phi_X}{\partial t} = X \circ \phi_X \quad \textit{on } J \times U.$$

Let $\mathcal{D}_U$ be the module of vector fields on U generated by the restrictions to U of the vector fields belonging to $\mathcal{D}$. Let X^U and F^U be the restrictions of X and F to U. If we have

$$\left\{ \begin{array}{l} [X^U, \mathcal{D}_U] \subset \mathcal{D}_U, \\ [X^U, F^U] \in \mathcal{D}_U, \end{array} \right. \tag{180}$$

then there exist $\mathcal{C}^\infty$ functions c_k^j, $1 \le k \le s$, $1 \le j \le s$, defined on the open subset $\mathcal{O} = \bigcup_{t \in J} (\{t\} \times \phi_X^t(U))$ of $I\!R \times \mathcal{N}$, such that, for any (t, χ) in $\mathcal{O}$,

$$\left\{ \begin{array}{l} \phi^t_{X*} G_k(\chi) = \sum_{i=1}^{s} c_k^j(\chi, t) F_j(\chi,) \quad k = 1, \dots, s, \\ \phi^t_{X*} F(\chi) = F(\chi) + \sum_{i=1}^{s} c_0^j(\chi, t) G_j(\chi). \end{array} \right. \tag{181}$$

Moreover, the $s \times s$ matrix

$$[c_k^j(\chi, t)]_{1 \le k \le s,\, 1 \le j \le s} \tag{182}$$

is invertible for any (χ, t) in $\mathcal{O}$.

Note that the converse of this lemma is obvious ((181) implies (180)) and that it is known (180) implies that for any t $\phi^t_{X*}F$ and the $\phi^t_{X*}G_k$'s are in $\mathcal{D}$. The precision here is the continuous dependence of the coefficients on t and also the fact that the matrix $[c_k^j(\chi, t)]$ remains invertible.

Proof. (180) implies that there exists some $\mathcal{C}^\infty$ functions b^j_k, $1 \le j \le s$, $0 \le k \le s$, defined on U, such that

$$\text{(183)} \qquad \left\{ \begin{array}{rcl} [X^U , G^U_k] & = & - \sum_{j=1}^s b^j_k(\chi) G^U_j(\chi), \\ {[X^U , F^U]} & = & - \sum_{i=1}^s b^j_0(\chi) G^U_j \end{array} \right.$$

where G^U_j stands for the restriction of G_j to U. Denote by F_t and $G_{k,t}$, $1 \le k \le s$, the vector fields ${\phi^t_X}_* F$ and ${\phi^t_X}_* F$ respectively, defined on the open subset $\phi(t, U)$ of $\mathcal{N}$. The mappings $(t, x) \mapsto F_t(\chi)$ and $(t, x) \mapsto G_{k,t}(\chi)$ are $\mathcal{C}^\infty$ from $\mathcal{O}$ to $T\mathcal{N}$, and by the definition of the Lie bracket,

$$\text{(184)} \qquad \frac{\partial F_t}{\partial t} = \phi^t_{X*}[X^U, F^U]; \qquad \frac{\partial G_{k,t}}{\partial t} = \phi^t_{X*}[X^U, G^U_k].$$

Hence, from (183), F_t and the $G_{k,t}$'s satisfy the linear differential system

$$\text{(185)} \qquad \left\{ \begin{array}{rcl} \dfrac{\partial F_t}{\partial t} & = & \displaystyle\sum_{i=1}^s (b^i_0 \circ \phi^{-t}_X)\, G_{i,t}, \\ \dfrac{\partial G_{k,t}}{\partial t} & = & \displaystyle\sum_{i=1}^s (b^i_k \circ \phi^{-t}_X)\, G_{i,t} \end{array} \right.$$

and the initial condition $F_0 = F$, $G_{k,0} = G_k$. The system (185) has a *unique* solution satisfying these initial conditions. Hence, to find parameterized vector fields F_t and $G_{k,t}$, all we have to do is find a solution of (185) satisfying the initial conditions. To do this, let us look for a solution of the form $F_t = F + \sum_{i=1}^s c^i_{0,t} G_i$, $G_{k,t} = \sum_{i=1}^s c^i_{k,t} G_i$, where the c^i_k are $\mathcal{C}^\infty$ functions defined on $\mathcal{O}$ to be computed such that

$$\text{(186)} \qquad c^i_k(0, x) = \left\{ \begin{array}{ll} 1 \ , \ x \in U & \text{if } i = k,\ 1 \le k \le s, \\ 0 \ , \ x \in U & \text{if } i \ne k \ \text{ or } k = 0 \end{array} \right.$$

and $c^i_{k,t} \in \mathcal{C}^\infty(\phi(t, U))$ is the function $x \in \phi(t, U) \mapsto c^i_k(t, x)$. This "ansatz" satisfies the initial conditions and satisfies (185) if

$$\text{(187)} \qquad \left\{ \begin{array}{l} \displaystyle\sum_{i=1}^s \frac{\partial c^i_{0,t}}{\partial t} G_i = \sum_{i=1}^s \sum_{j=1}^s (b^i_0 \circ \phi^{-t}_X)\, c^j_{i,t}\, G_j, \\ \displaystyle\sum_{i=1}^s \frac{\partial c^i_{k,t}}{\partial t} G_i = \sum_{i=1}^s \sum_{j=1}^s (b^i_k \circ \phi^{-t}_X)\, c^j_{i,t}\, G_j. \end{array} \right.$$

(187) will be satisfied if the functions c^i_j satisfy the linear differential system

$$\text{(188)} \qquad \left\{ \begin{array}{c} \dfrac{\partial c^i_0}{\partial t} = \displaystyle\sum_{j=1}^s \widehat{b}^j_0\, c^i_j, \\ \dfrac{\partial}{\partial t} \begin{pmatrix} c^1_1 & \cdots & c^1_s \\ \vdots & & \vdots \\ c^s_1 & \cdots & c^s_s \end{pmatrix} = \begin{pmatrix} \widehat{b}^1_1 & \cdots & \widehat{b}^1_s \\ \vdots & & \vdots \\ \widehat{b}^s_1 & \cdots & \widehat{b}^s_s \end{pmatrix} \begin{pmatrix} c^1_1 & \cdots & c^1_s \\ \vdots & & \vdots \\ c^s_1 & \cdots & c^s_s \end{pmatrix} \end{array} \right.$$

with initial conditions (186), where

$$\widehat{b}^i_k \in \mathcal{C}^\infty(\mathcal{O}); \quad \widehat{b}^i_k(t, \chi) = b^i_k(\phi^{-t}_X(\chi)).$$

This linear system has a unique solution, defined for any t, and the matrix in the right-hand side of (188), i.e. the matrix $[c_k^j(\chi,t)]_{1\leq k\leq s,\, 1\leq j\leq s}$, is always invertible. We have constructed the functions c_k^j satisfying the required conditions. □

REFERENCES

[1] B. Bonnard, *Feedback equivalence for nonlinear systems and the time-optimal control problem*, SIAM J. Control Optim., 29 (1991), pp. 1300–1321.

[2] R. B. Gardner and W. F. Shadwick, *Feedback equivalence of control systems*, Systems Control Lett., 8 (1987), pp. 463–465.

[3] B. Jakubczyk, *Equivalence and invariants of nonlinear control systems*, in Nonlinear Controllability and Optimal Control, H. J. Sussmann, ed., Marcel Dekker, New York, 1990.

[4] I. Kanellakopoulos, P. V. Kokotovic, and R. Marino, *Robustness of adaptive nonlinear control under an extended matching condition*, in Proc. of the IFAC Symposium on nonlinear control design, A. Isidori, ed., Capri, Italy, June 1989, IFAC Symposium Series, Pergamon Press, Oxford, UK, 1990.

[5] ———, *An extended direct scheme for robust adaptive nonlinear control*, Automatica, 27 (1991), pp. 247–255.

[6] J.-B. Pomet, *Sur la commande adaptative des systèmes non-linéaires*, thèse de doctorat, Ecole des Mines de Paris, Paris, France, Sept. 1989.

[7] J.-B. Pomet and I. A. K. Kupka, *Global aspects of feedback equivalence for a parametrized family of systems*, in Analysis of Controlled Dynamical Systems, B. Bonnard, B. Bride, J.-P. Gauthier, and I. A. K. Kupka, eds., Progress in Systems and Control Theory 8, Birkhäuser, Boston, 1990, pp. 337–346.

[8] J.-B. Pomet and L. Praly, *Adaptive control of feedback equivalent systems*, in Analysis and Optimization of Systems, J.-L. Lions and A. Bensoussan, eds., Lecture Notes in Control and Information Science 144, Springer-Verlag, New York, 1990, pp. 808–817.

[9] L. Praly, G. Bastin, J.-B. Pomet, and Z.-P. Jiang, *Adaptive stabilization of nonlinear systems*, in Foundations of Adaptive Control, P. V. Kokotovic, ed., Lecture Notes in Control and Information Science 160, Springer-Verlag, New York, 1991, pp. 345–433.

[10] D. G. Taylor, P. V. Kokotovic, R. Marino, and I. Kanellakopoulos, *Adaptive regulation of nonlinear systems with unmodelled dynamics*, IEEE Trans. Automat. Control, 34 (1989), pp. 405–412.

[11] K. Tchoń, *The only stable normal forms of affine systems under feedback are linear*, Systems Control Lett., 8 (1987), pp. 359–365.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 33, No. 4, pp. 1208–1223, July 1995

© 1995 Society for Industrial and Applied Mathematics
013

NECESSARY CONDITIONS FOR BILEVEL DYNAMIC OPTIMIZATION PROBLEMS*

JANE J. YE[†]

Abstract. In this paper we study the bilevel dynamic optimization problem, which is a hierarchy of two optimization problems where the constraint region of the upper-level problem is determined implicitly by the solution to the lower-level problem and where the upper-level decision variable is a vector while the lower-level decision variable is an admissible control function. To obtain optimality conditions, we reformulate the bilevel dynamic optimization problem as a single-level optimal control problem that involves the value function of the lower-level problem. A sensitivity analysis of the lower-level problem with respect to the perturbation in the upper-level decision variable is given, and the first-order necessary optimality conditions are derived by using nonsmooth analysis.

Key words. necessary conditions, bilevel dynamic optimization problems, sensitivity analysis, nonsmooth analysis, value function

AMS subject classifications. 90D65, 49K40

1. Introduction.

Let us consider a two-level hierarchical system where the higher level (hereafter the "leader") and the lower level (hereafter the "follower") must find vectors $z \in Z$ and control functions $u(\cdot)$, respectively, to minimize their individual objective functions $J_1(z, u)$ and $J_2(z, u)$. The leader is assumed first to select his decision vector $z \in Z$ and the follower next to select his decision control function $u(\cdot) \in \mathcal{U}$, where Z is a nonempty subset of R^n and $\mathcal{U}$ is the set of admissible controls. Under these assumptions on the order of play, the game will proceed as follows. Given any decision vector $z \in Z$ chosen by the leader, the follower will select his decision control function $u_z(\cdot) \in \mathcal{U}$ (depending on the decision vector z chosen by the leader) to minimize his objective $J_2(z, u_z)$. Assume that the game is cooperative, i.e., if the follower's problem has several optimal controls for a given parameter z, then the follower allows the leader to choose which of them is actually used. Thus the leader chooses his optimal decision vector $z \in Z$ to minimize the leader's objective $J_1(z, u_z)$. In other words, given any decision vector $z \in Z$ chosen by the leader, the follower faces the ordinary (*single-level*) optimal control problem involving a parameter z:

$$\begin{aligned} P_2(z) \qquad \min J_2(z,u) &= \int_{t_0}^{t_1} G(t, x(t), z, u(t))dt + g(x(t_1)), \\ \text{s.t. } \dot{x}(t) &= \phi(t, x(t), z, u(t)) \qquad \text{a.e.}, \\ x(t_0) &= x_0, \quad x(t_1) \in C_1, \\ u(t) &\in U(t) \qquad \text{a.e.}, \end{aligned}$$

while the leader faces the *bilevel dynamic optimization problem*:

$$P_1 \qquad \min J_1(z, u_z) = \int_{t_0}^{t_1} F(t, x_z(t), z, u_z(t))dt + f(x_z(t_1))$$

over $z \in Z$ and all optimal pairs (x_z, u_z) of $P_2(z)$.

* Received by the editors May 28, 1993; accepted for publication (in revised form) February 8, 1994. This research was supported by Natural Sciences and Engineering Research Council of Canada grant WFA 0123160.

[†] Department of Mathematics and Statistics, University of Victoria, Victoria, British Columbia, Canada V8W 3P4 (janeye@sol.uvic.ca).

The bilevel dynamic optimization problem has many applications in economics and management science. For instance, the leader may be the government that sets up the taxation policy z and the follower may be a company that seeks the optimal policy $u_z(t)$ in reaction to the government's taxation policy.

The bilevel static problem where both leader's and follower's decisions are vectors instead of control functions was first introduced by von Stackelberg [10] for an economic model. The bilevel dynamic problem where both leader's and follower's decisions are control functions was first considered by Chen and Cruz in [2]. The bilevel dynamic optimization problem studied in this paper is a special case of the bilevel dynamic problem as in Zhang [13]. Several names for bilevel (static or dynamic) optimization problems have been used in the literature, such as Stackelberg game, principal-agent problem, bilevel programming problem, and two-level hierarchical optimization problem. Most of the bilevel (static or dynamic) problems are attacked by reducing the lower-level problem through first-order necessary conditions (cf. Bard and Falk [1] and Zhang [13], [14] for the bilevel static problem and Zhang [13] for the bilevel dynamic problem). The reduction is equivalent if and only if the lower-level problem satisfies certain convexity assumptions since in this case the first-order necessary condition is also sufficient. Apart from the strong convexity assumption, the resulting optimality conditions of the above approach involve second-order (generalized in nonsmooth case [13]) derivatives and a larger system since the reduced problem minimizes over the set of original decision variables as well as the set of multipliers of the lower-level problem.

The purpose of this paper is to provide first-order necessary conditions for problem P_1 under very general assumptions (in particular, without convexity assumptions on the lower-level problem).

Define the *value function of the lower-level optimal control problem* as an extended-valued function $V : Z \to \bar{R}$ defined by

$$V(z) := \inf \left\{ \begin{array}{ll} \int_{t_0}^{t_1} G(t,x(t),z,u(t))dt + g(x(t_1)) : & \dot{x}(t) = \phi(t,x(t),z,u(t)) \text{ a.e.} \\ & u(t) \in U(t) \qquad \text{a.e.} \\ & x(t_0) = x_0, \quad x(t_1) \in C_1 \end{array} \right\},$$

where $\bar{R} := R \cup \{-\infty\} \cup \{+\infty\}$ is the extended real line and $\inf \emptyset = +\infty$ by convention. Our approach is to reformulate P_1 as in the following single-level optimal control problem:

$$\begin{aligned} \widetilde{P}_1 \qquad & \min J_1(z,u) = \int_{t_0}^{t_1} F(t,x(t),z(t),u(t))dt + f(x(t_1)), \\ & \text{s.t. } \dot{x}(t) = \phi(t,x(t),z(t),u(t)) \qquad \text{a.e.}, \\ & \dot{z}(t) = 0, \\ & x(t_0) = x_0, \quad x(t_1) \in C_1, \\ & u(t) \in U(t) \qquad \text{a.e.}, \\ & \int_{t_0}^{t_1} G(t,x(t),z(t),u(t))dt + g(x(t_1)) \le V(z(t_1)). \end{aligned}$$

The above problem is obviously equivalent to the original bilevel dynamic optimization problem P_1 and is a standard optimal control problem except that the endpoint constraints involve the value function V of the lower-level optimal control problem. In general V is not an explicit function of the problem data and is nonsmooth even in the case where all problem data are smooth functions. Recent developments in nonsmooth analysis allow us to study the generalized derivatives of the value function V and relate them to the multiplier sets for the lower-level optimal control problem, hence deriving a necessary condition for optimality. This approach was first used by Ye and Zhu [12] to derive first-order necessary conditions for the static bilevel optimization problem. The following basic assumptions are in force throughout this paper:

(A1) $Z \subset \mathbb{R}^n$ and C_1 are closed.

(A2) $U(t) : [t_0, t_1] \to \mathbb{R}^m$ is a nonempty compact-valued set-valued map. The graph of $U(t)$ (i.e., the set $\{(s,r) : s \in [t_0,t_1], r \in U(s)\}$), denoted by Gr$U$, is $\mathcal{L}\times\mathcal{B}$ measurable, where $\mathcal{L}\times\mathcal{B}$ denotes the σ-algebra of subsets of $[t_0,t_1]\times\mathbb{R}^m$ generated by product sets $M \times N$ where M is a Lebesgue measurable subset of $[t_0,t_1]$ and N is a Borel subset of $\mathbb{R}^m$.

(A3) There exists an integrable function k defined on $[t_0,t_1]$ such that for each $(t,u) \in \text{Gr}U$, the functions $\phi(t,\cdot,\cdot,u), F(t,\cdot,\cdot,u), G(t,\cdot,\cdot,u)$ are locally Lipschitz of rank $k(t)$. For each $(x,z) \in \mathbb{R}^d \times \mathbb{R}^n$, the functions $\phi(\cdot,x,z,\cdot) : [t_0,t_1] \times \mathbb{R}^m \to \mathbb{R}^d$, $F(\cdot,x,z,\cdot) : [t_0,t_1] \times \mathbb{R}^m \to \mathbb{R}$, $G(\cdot,x,z,\cdot) : [t_0,t_1] \times \mathbb{R}^m \to \mathbb{R}$ are $\mathcal{L} \times \mathcal{B}$ measurable.

(A4) The functions $f, g : \mathbb{R}^d \to \mathbb{R}$ are locally Lipschitz continuous.

(A5) For any $z \in Z$, $P_2(z)$ has an admissible pair (whose definition is given below).

A *control function* is a (Lebesgue) measurable selection $u(\cdot)$ for $U(\cdot)$, that is, a measurable function satisfying $u(t) \in U(t)$ a.e. $t \in [t_0,t_1]$. An *arc* is an absolutely continuous function. An *admissible pair* for $P_2(z)$ is a pair of functions $(x(\cdot),u(\cdot))$ on $[t_0,t_1]$ of which $u(\cdot)$ is a control function and $x(\cdot) : [t_0,t_1] \to \mathbb{R}^d$ is an arc that satisfies the differential equation $\dot{x}(t) = \phi(t,x(t),z,u(t))$ a.e., together with the initial condition $x(t_0) = x_0$ and the endpoint constraint $x(t_1) \in C_1$. The first and the second components of an admissible pair are called an *admissible trajectory* and an *admissible control*, respectively. A *solution* to problem $P_2(z)$ is an admissible pair that minimizes the value of the cost functional $J_2(z,u)$ over all admissible pairs. An *admissible strategy* for P_1 includes a vector $z \in Z$ and an optimal control u_z for $P_2(z)$. The strategy (z,u_z) is *optimal* for the bilevel dynamic optimization problem P_1 if (z,u_z) minimizes the value of the cost functional $J_1(z,u_z)$ among all admissible strategies for P_1.

A plan of the paper is as follows. In §2, we give background material on nonsmooth analysis that will be referred to in the following sections. In §3, we study generalized differentiability of the value function $V(z)$. The necessary condition for optimality is given in §4. In §5, we consider an extension to the bilevel dynamic optimization problem defined in §1 to allow opportunity costs; a fishery regulation problem is used to demonstrate applications of the necessary condition for optimality derived.

2. Nonsmooth analysis background.

In this section we shall give a concise review of the material on nonsmooth analysis that will be required.

Let C be a nonempty closed set in $\mathbb{R}^n$. A vector $\zeta \in \mathbb{R}^n$ is a *proximal normal* to C at point $\bar{x} \in C$ if for $t > 0$ sufficiently small, the unique point of C nearest to $\bar{x} + t\zeta$ (in the Euclidean norm) is $\bar{x}$. It is a *limiting proximal normal* if there exist points $x_k \in C, x_k \to \bar{x}$, and proximal normals ζ_k to C at x_k, such that $\zeta_k \to \zeta$. Let the *limiting proximal normal cone to C at $\bar{x}$* be the set

$$\hat{N}_C(\bar{x}) := \{\zeta : \zeta \text{ is a limiting proximal normal to } C \text{ at } \bar{x}\}$$

and the *Clarke normal cone to C at $\bar{x}$* to be the set

$$N_C(\bar{x}) := \text{clco}\hat{N}_C(\bar{x}).$$

Now consider a lower semicontinuous function $\phi : I\!\!R^n \to I\!\!R \cup \{+\infty\}$ and a point $\bar{x} \in I\!\!R^n$ where ϕ is finite. A vector $\zeta \in I\!\!R^n$ is called a *proximal subgradient* of $\phi(\cdot)$ at $\bar{x}$ provided that there exist $M > 0, \delta > 0$ such that

$$\langle \zeta, x' - \bar{x} \rangle \leq \phi(x') - \phi(\bar{x}) + M\|x' - \bar{x}\|^2, \qquad x' \in \bar{x} + \delta B,$$

where $\langle a, b \rangle$ denotes the inner product of vectors a and b. The set of all proximal subgradients of $\phi(\cdot)$ at $\bar{x}$ is denoted $\partial^\pi \phi(\bar{x})$. The *limiting subgradient* of ϕ at $\bar{x}$ is the set

$$\hat{\partial}\phi(\bar{x}) := \left\{ \lim_{k \to \infty} \zeta_k : \zeta_k \in \partial^\pi \phi(x_k), x_k \to \bar{x}, \phi(x_k) \to \phi(\bar{x}) \right\}.$$

The *singular limiting subgradient* of ϕ at $\bar{x}$ is the set

$$\hat{\partial}^\infty \phi(\bar{x}) := \left\{ \lim_{k \to \infty} t_k \zeta_k : \zeta_k \in \partial^\pi \phi(x_k), x_k \to \bar{x}, \phi(x_k) \to \phi(\bar{x}), t_k \downarrow 0 \right\}.$$

The limiting subgradient is a smaller object than the *Clarke generalized gradient.* In fact, if ϕ is Lipschitz continuous near x, we have $\partial\phi(x) = \mathrm{co}\hat{\partial}\phi(x)$, where $\partial\phi$ and coA denote the Clarke generalized gradient of ϕ and the convex hull of the set A, respectively. For the definition and the precise relation between the limiting subgradient and the Clarke generalized gradient, the reader is referred to Clarke [5] and Rockafellar [9].

The following proposition summarizes the prerequisites regarding limiting subgradients and limiting proximal normal cones.

PROPOSITION 2.1. (a) *If C is a nonempty closed convex set, the limiting proximal normal cone to C coincides with the normal cone in the sense of convex analysis, i.e., one has $\zeta \in \hat{N}_C(\bar{x})$ if and only if*

$$\langle \zeta, x - \bar{x} \rangle \leq 0 \quad \forall x \in C.$$

(b) *The function $\phi(\cdot)$ is Lipschitz near x if and only if $\hat{\partial}^\infty \phi(x) = \{0\}$.*

(c) *If $\hat{\partial}\phi(x) \neq \emptyset$, then*

$$\hat{\partial}(s\phi)(x) = s\hat{\partial}\phi(x) \qquad \forall s \geq 0.$$

(d) (*Clarke* [5, *Prop.* 1.5]) *Let ϕ and $\psi : I\!\!R^n \to I\!\!R \cup \{+\infty\}$ be lower semicontinuous functions finite at x, with $\hat{\partial}^\infty \phi(x) \cap (-\hat{\partial}^\infty \psi(x)) = \{0\}$. Then we have*

$$\hat{\partial}(\phi + \psi)(x) \subset \hat{\partial}\phi(x) + \hat{\partial}\psi(x).$$

(e) *Let $\Psi_C(x)$ be the indicator function of the set C. Then*

$$\hat{N}_C(x) = \hat{\partial}\Psi_C(x) = \hat{\partial}^\infty \Psi_C(x).$$

(f) *Let S_1 and S_2 be closed subets of $I\!\!R^n$ and let $\bar{x} \in S_1 \cap S_2$. If $\hat{N}_{S_1}(\bar{x}) \cap (-\hat{N}_{S_2}(\bar{x})) = \{0\}$, then we have*

$$\hat{N}_{S_1 \cap S_2}(\bar{x}) \subset \hat{N}_{S_1}(\bar{x}) + \hat{N}_{S_2}(\bar{x}).$$

(g) (*chain rule*) *Let $\phi(x) := f(F(x))$ where $F : I\!\!R^n \to I\!\!R$ is Lipschitz on some neighbourhood of $\bar{x}$, while $f : I\!\!R \to I\!\!R \cup \{+\infty\}$ is lower semicontinuous with $F(\bar{x})$ in $\mathrm{dom} f := \{y : f(y) \neq +\infty\}$. Then if*

$$0 \notin \hat{\partial}(\zeta F)(\bar{x}) \qquad \forall \text{ nonzero vectors } \zeta \in \hat{\partial}^\infty f(F(\bar{x})),$$

we have

$$\hat{\partial}\phi(\bar{x}) \subset \cup\{\hat{\partial}(\zeta F)(\bar{x}) : \zeta \in \hat{\partial} f(F(\bar{x}))\}.$$

3. Differentiability of the value function. To discuss generalized differentiability of the value function $V(z)$, we will refer to the following assumptions:

(A6) For some $\alpha \geq 0, \beta \geq 0$, the function $(\phi(t,\cdot,z,u), G(t,\cdot,z,u))$ satisfies the following growth condition: for all $z \in Z$, $(t,u) \in \mathrm{Gr}U$, one has

$$|(\phi(t,x,z,u), G(t,x,z,u))| \leq \alpha|x| + \beta.$$

(A6)$'$ The functions ϕ and G are continuously differentiable in x and z and lower semicontinuous in u. There exists an integrable function $k(t)$ such that

$$|\phi| + |\nabla_x \phi| + |G| + |\nabla_x G| \leq k(t).$$

(A7) For any $(t,x,z) \in [t_0,t_1] \times I\!R^d \times I\!R^n$, the set

$$\{(\phi(t,x,z,u), G(t,x,z,u)) : u \in U(t)\}$$

is convex.

(A7)$'$ For any $(t,x,z) \in [t_0,t_1] \times I\!R^d \times I\!R^n$, the set

$$\{(\phi(t,x,z,u), G(t,x,z,u) + \delta) : u \in U(t), \delta \geq 0\}$$

is convex.

The Hamiltonian for $P_2(z)$ is the function defined by

$$H_2(t,x,z,p_2;\lambda) := \sup\{p_2 \cdot \phi(t,x,z,u) - \lambda G(t,x,z,u) : u \in U(t)\}.$$

An index λ multiplier corresponding to an admissible trajectory x for $P_2(z)$ is an arc (p_2, q) such that

$$\begin{aligned} (-\dot{p_2}(t), -\dot{q}(t), \dot{x}(t)) &\in \partial_{(x,z,p_2)} H_2(t,x(t),z,p_2(t);\lambda) \qquad \text{a.e.} \\ -p_2(t_1) &\in \lambda \hat{\partial} g(x(t_1)) + \hat{N}_{C_1}(x(t_1)), \\ q(t_1) &= 0. \end{aligned}$$

The collection of all such arcs is the set $M^\lambda(x)$, the index λ multiplier set corresponding to x. Let Y be the set of all optimal trajectories x to problem $P_2(z)$. Let

$$M^\lambda(Y) := \bigcup_{x \in Y} M^\lambda(x).$$

For any index λ multiplier $(p_2,q) \in M^\lambda(x)$, we define $Q(p_2,q) = -q(t_0)$. The notation $QM^\lambda(x)$ designates the set of all possible values of $-q(t_0)$ obtained in this way, and $Q(M^\lambda(Y))$ denotes $\cup_{x \in Y} Q(M^\lambda(x))$. The following result relates the differential properities of V to the arcs q in the multiplier sets introduced above.

THEOREM 3.1. *In addition to assumptions* (A1)–(A5), *suppose either* (A6)–(A7) *or* (A6)$'$–(A7)$'$ *hold. If* $QM^0(Y) = \{0\}$, *then* V *is Lipschitz continuous near* z *and one has*

$$\hat{\partial} V(z) \subset QM^1(Y).$$

Theorem 3.1 under assumptions (A6)–(A7) can be obtained by reducing the original optimal control problem to an differential inclusion problem and applying the sensitivity result in Clarke and Loewen [6, Thm. 3.3]. Before proving Theorem 3.1 under assumptions (A1)–(A5) and (A6)$'$–(A7)$'$, we first give the following result.

LEMMA 3.2. *Let α_i be a sequence converging to α, and let (x_i, u_i) be an admissible pair for $P_2(\alpha_i)$. Then there exists a subsequence of $\{x_i\}$ converging uniformly to an arc x and a control u with (x, u) being an admissible pair for $P_2(\alpha)$ such that*

$$J_2(x,u) \leq \liminf J_2(x_i, u_i).$$

The proof can be reduced to an application of [4, Thm. 3.1.7] by studying the differential inclusion

$$(\dot{x}(t), \dot{y}(t), \dot{\alpha}(t)) \in \Gamma(t, x(t), y(t), \alpha(t)) \qquad \text{a.e.},$$

where $y \in I\!R$ and the convex multifunction Γ is defined via

$$\Gamma(t,x,y,\alpha) := \{[\phi(t,x,\alpha,u), r, 0] : G(t,x,\alpha,u) \leq r \leq k(t)+1, u \in U(t)\}.$$

The essential fact in the reduction is Filippov's lemma: (x, y, α) satisfies the above differential inclusion iff there is a control u for x such that (x, u) is an admissible pair for $P_2(\alpha)$ and y satisfies

$$G(t,x,\alpha,u) \leq \dot{y} \leq k(t)+1.$$

We now turn to the proof of the theorem. By (A5), $P_2(z)$ has an admissible pair. So $V(z)$ is finite. It follows from Lemma 3.2 that V is lower-semicontinuous.

Step 1. Let $\alpha \in Z$ be a point near z. Let $\zeta \in \partial^{\pi} V(\alpha)$, and let (x, u) be a solution of $P_2(\alpha)$ that exists by virtue of Lemma 3.2. Then by definition, for some $M > 0$ and for all α' near α, we have

$$\begin{aligned} V(\alpha') - \langle \zeta, \alpha' \rangle + M|\alpha' - \alpha|^2 &\geq V(\alpha) - \langle \zeta, \alpha \rangle \\ &= \int_{t_0}^{t_1} G(t, x(t), \alpha, u(t))dt + g(x(t_1)) - \langle \zeta, \alpha \rangle. \end{aligned}$$

Let (x', u') be an admissible pair for $P_2(\alpha')$. Then

$$\begin{aligned} &\int_{t_0}^{t_1} G(t, x'(t), \alpha', u'(t))dt + g(x(t_1)) - \langle \zeta, \alpha' \rangle + M|\alpha' - \alpha|^2 \\ &\qquad \geq \int_{t_0}^{t_1} G(t, x(t), \alpha, u(t))dt + g(x(t_1)) - \langle \zeta, \alpha \rangle. \end{aligned}$$

Hence (x, α, u) is a solution of the following optimal control problem:

$$\begin{aligned} \min \int_{t_0}^{t_1} & G(t, x'(t), \alpha'(t), u'(t))dt + g(x'(t_1)) - \langle \zeta, \alpha'(t_0) \rangle, \\ \text{s.t. } & \dot{x}'(t) = \phi(t, x'(t), \alpha'(t), u'(t)) \qquad \text{a.e.}, \\ & \alpha'(t) = 0, \\ & x'(t_0) = x_0, \quad x'(t_1) \in C_1, \\ & u'(t) \in U(t) \qquad \text{a.e.} \end{aligned}$$

In the proof of Theorem 5.2.1 of Clarke [4], if we replace the the Clarke generalized gradient ∂ by the limiting subgradient $\hat{\partial}$ in the transversality conditions, the argument goes through without modification (cf. Clarke [5]). It follows that there exist a scalar $\lambda \geq 0$ and arcs p_2, q such that

$$\begin{aligned}
&(1) \qquad -\dot{p_2}(t) = \nabla_x \phi(t, x(t), \alpha, u(t))^\top p_2(t) - \lambda \nabla_x G(t, x(t), \alpha, u(t)) \qquad \text{a.e.;}\\
&(2) \qquad -\dot{q}(t) = \nabla_\alpha \phi(t, x(t), \alpha, u(t))^\top p_2(t) - \lambda \nabla_\alpha G(t, x(t), \alpha, u(t)) \qquad \text{a.e.;}\\
&\qquad\quad \max_{u \in U(t)} \{p_2(t) \cdot \phi(t, x(t), \alpha, u) - \lambda G(t, x(t), \alpha, u)\}\\
&(3) \qquad\qquad = p_2(t) \cdot \phi(t, x(t), \alpha, u(t)) - \lambda G(t, x(t), \alpha, u(t)) \qquad \text{a.e.,}\\
&\qquad\quad -p_2(t_1) \in \lambda \hat{\partial} g(x(t_1)) + \hat{N}_{C_1}(x(t_1)),\\
&\qquad\quad q(t_0) = -\lambda \zeta, \qquad q(t_1) = 0,\\
&\qquad\quad \|p_2\|_\infty + \|q\|_\infty + \lambda > 0,
\end{aligned}$$

where ∂ denotes the Clarke generalized gradient, $\|\cdot\|_\infty$ denotes the supremum norm, and $^\top$ denotes the transpose.

By Clarke [4, Thm. 2.8.2], since ϕ and G are continuously differentiable in (x, z), $\partial_{(x,\alpha,p_2)} H_2(t, x, \alpha, p_2; \lambda)$ is the convex hull of all points of the form

$$[\nabla_x \phi(t, x, \alpha, u)^\top p_2 - \lambda \nabla_x G(t, x, \alpha, u), \nabla_\alpha \phi(t, x, \alpha, u)^\top p_2 - \lambda \nabla_\alpha G(t, x, \alpha, u), \phi(t, x, \alpha, u)],$$

where u in $U(t)$ is any point at which the maximum defining $H_2(t, x, \alpha, p_2; \lambda)$ is achieved. Hence (1), (2), and (3) imply that

$$(-\dot{p_2}(t), -\dot{q}(t), \dot{x}(t)) \in \partial_{(x,\alpha,p_2)} H_2(t, x(t), \alpha, p_2(t); \lambda) \qquad \text{a.e.}$$

Step 2. For any $\zeta \in \hat{\partial} V(z)$, by definition, $\zeta = \lim_{i \to \infty} \zeta_i$ where $\zeta_i \in \partial^\pi V(\alpha_i)$, $\alpha_i \to z$, and $V(\alpha_i) \to V(z)$. By Step 1, for each ζ_i, there exists an arc (p_2^i, q_i), a scalar λ_i, and an arc x_i that solves $P_2(\alpha_i)$ such that

$$\begin{aligned}
(-\dot{p_2}^i(t), -\dot{q}_i(t), \dot{x}_i(t)) &\in \partial_{(x,\alpha,p_2)} H_2(t, x_i(t), \alpha_i, p_2^i(t); \lambda_i) \qquad \text{a.e.,}\\
-p_2^i(t_1) &\in \lambda_i \hat{\partial} g(x_i(t_1)) + \hat{N}_{C_1}(x_i(t_1)),\\
q_i(t_0) &= -\lambda_i \zeta_i, \qquad q_i(t_1) = 0,\\
\|p_2^i\| + \|q_i\| + \lambda_i &> 0.
\end{aligned}$$

Since $M^0(Y) = \{0\}$, we must indeed have $\lambda_i = 1$ for i sufficiently large and $|p_2^i(0)|$ bounded (cf., Clarke and Loewen [6, p. 253]). Passing to a uniformly convergent subsequence of $\{(p_2^i, q_i, x_i)\}$ by Lemma 3.2 and Clarke [4, Thm. 3.1.7] leads to an optimal trajectory x for $P_2(z)$ and an arc (p_2, q) such that

$$\begin{aligned}
(-\dot{p_2}(t), -\dot{q}(t), \dot{x}(t)) &\in \partial_{(x,\alpha,p_2)} H_2(t, x(t), \alpha, p_2(t); \lambda) \qquad \text{a.e.,}\\
-p_2(t_1) &\in \hat{\partial} g(x(t_1)) + \hat{N}_{C_1}(x(t_1)),\\
q(t_0) &= -\zeta, \qquad q(t_1) = 0.
\end{aligned}$$

That is, $(p_2, q) \in QM^1(Y)$.

Similarly to Ye [11], one can show $\hat{\partial}^\infty V(z) \subset QM^0(Y)$ using results from Step 2. The Lipschitz continuity of V near z then follows by virtue of assumption $M^0(Y) = \{0\}$ and (b) of Proposition 2.1. The proof of Theorem 3.1 is now complete.

4. Necessary conditions for optimality. Define the *pseudo-Hamiltonian* for problem $(\widetilde{P}_1)$ as

$$H_1(t,x,z,p_1;\lambda,r) := p_1 \cdot \phi(t,x,z,u) - rG(t,x,z,u) - \lambda F(t,x,z,u),$$

for $t \in [t_0,t_1]$, $x, p_1 \in \mathbb{R}^d$, $z \in Z$, $\lambda, r \in \mathbb{R}$.

THEOREM 4.1. *Assume assumptions* (A1)–(A4) *hold. Let* $(z, u(t))$ *be an optimal strategy of the bilevel dynamic optimization problem* P_1 *and* $x(t)$ *the corresponding trajectory. Assume that the value function for the lower-level problem* V *is locally Lipschitz continuous. Then there exist* $\lambda \geq 0, r \geq 0$ *and arcs* p_1, η *such that:*

$$-(\dot{p_1}(t), \dot{\eta}(t)) \in \partial_{(x,z)} H_1(t,x(t),z,p_1(t),u(t);\lambda,r) \qquad a.e., \tag{4}$$

$$\max_{u\in U(t)} H_1(t,x(t),z,p_1(t),u;\lambda,r) = H_1(t,x(t),z,p_1(t),u(t);\lambda,r) \qquad a.e., \tag{5}$$

$$\eta(t_0) = 0,$$

$$-p_1(t_1) \in \lambda\hat{\partial} f(x(t_1)) + r\hat{\partial} g(x(t_1)) + \hat{N}_{C_1}(x(t_1)), \tag{6}$$

$$\eta(t_1) \in r\partial V(z), \tag{7}$$

$$\|p_1\|_\infty + \|\eta\|_\infty + \lambda + r > 0. \tag{8}$$

The following result, which is a limiting subgradient version of Corollary 1 of Theorem 2.4.7 in Clarke [4], will be useful in proving Theorem 4.1. We should prove it by using a chain rule.

LEMMA 4.2. *Let* $C = \{x : \psi(x) \leq 0\}$, *where* $\psi : \mathbb{R}^n \to \mathbb{R}$ *is Lipschitz continuous on some neighborhood of* $\bar{x} \in C$. *Suppose that* $0 \notin \hat{\partial}\psi(\bar{x})$. *Then*

$$\hat{N}_C(\bar{x}) \subset \bigcup_{r\geq 0} r\hat{\partial}\psi(\bar{x}). \tag{9}$$

Proof. If $\bar{x}$ is in the interior of C, then $\hat{N}_C(\bar{x}) = \{0\}$ and the above relation is trivially satisfied. Suppose $\bar{x}$ is in the boundary of C. By virtue of (a), (c), and (e) of Proposition 2.1, $0 \notin \hat{\partial}\psi(\bar{x})$ implies

$$0 \notin \hat{\partial} r\psi(\bar{x}) \qquad \forall \text{ nonzero scalars } r \in \mathbb{R}_+ = \hat{\partial}^\infty \Psi_{\mathbb{R}_-}(\psi(\bar{x})) = \hat{N}_{\mathbb{R}_-}(\psi(\bar{x})).$$

Since $\Psi_C(\bar{x}) = \Psi_{\mathbb{R}_-}(\psi(\bar{x}))$. by the chain rule ((g) of Proposition 2.1) we have

$$\hat{\partial}\Psi_C(\bar{x}) \subset \cup\{\hat{\partial}(r\psi)(\bar{x}) : r \in \hat{\partial}\Psi_{\mathbb{R}_-}(\psi(\bar{x}))\}, \tag{10}$$

which is the relation (9) thanks to Proposition 2.1(e). □

The proof of the following result is straightforward.

LEMMA 4.3. *Let* $F(x,y,z) : \mathbb{R}^d \times \mathbb{R}^m \times \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ *be a lower semicontinuous function and* $(\bar{x},\bar{y},\bar{z}) \in \operatorname{dom} F$. *Suppose* $F(x,y,z) = F_1(x) + F_2(y) + F_3(z)$. *Then*

$$\hat{\partial} F(\bar{x},\bar{y},\bar{z}) \subset \hat{\partial} F_1(\bar{x}) \times \hat{\partial} F_2(\bar{y}) \times \hat{\partial} F_3(\bar{z}).$$

Proof of Theorem 4.1. We pose the optimal control problem $\widetilde{P}_1$ equivalently as the problem

$$\widehat{P}_1 \qquad \min \int_{t_0}^{t_1} F(t,x(t),z(t),u(t))dt + f(x(t_1))$$

$$\begin{aligned} \text{s.t. } \dot{x}(t) &= \phi(t,x(t),z(t),u(t)) \quad \text{a.e.},\\ \dot{y}(t) &= G(t,x(t),z(t),u(t)) \quad \text{a.e.},\\ \dot{z}(t) &= 0,\\ u(t) &\in U(t) \quad \text{a.e.},\\ (x,y,z)(t_0) &\in \{x_0\}\times\{0\}\times I\!R,\\ (x,y,z)(t_1) &\in S := \{(x,y,z) : g(x)+y-V(z)\le 0, x\in C_1\}. \end{aligned}$$

The problem above is exactly in the form described in §5.2.1 of Clarke [4]. The pseudo-Hamiltonian is the function

$$H(t,x,y,z,p_1,p_2,\eta,u,\lambda) = p_1\cdot\phi(t,x,z,u) + p_2G(t,x,z,u) - \lambda F(t,x,z,u),$$

for $t\in[t_0,t_1], x, p_1\in I\!R^d, y, p_2, \eta, \lambda\in I\!R, z\in Z$. Applying Theorem 5.2.1 of Clarke [4] with the generalized gradient replaced by the limiting subgradient in the transversality conditions leads to the existence of a scalar $\lambda\ge 0$ and an arc (p_1,p_2,η) such that

$$-(\dot{p_1}(t),\dot{p_2}(t),\dot{\eta}(t)) \in \partial_{(x,y,z)}H(t,x(t),y(t),z(t),p_1(t),p_2(t),\eta(t),u(t),\lambda) \quad \text{a.e.,} \tag{11}$$

$$\max_{u\in U(t)} H(t,x(t),y(t),z(t),p_1(t),p_2(t),\eta(t),u,\lambda)$$

$$= H(t,x(t),y(t),z(t),p_1(t),p_2(t),\eta(t),u(t),\lambda) \quad \text{a.e.,} \tag{12}$$

$$(p_1(t_0),p_2(t_0),\eta(t_0)) \in \hat{N}_{\{x_0\}\times\{0\}\times I\!R}(x(t_0),y(t_0),z(t_0)), \tag{13}$$

$$-(p_1(t_1),p_2(t_1),\eta(t_1)) \in \lambda\hat{\partial}\hat{f}(x(t_1),y(t_1),z(t_1)) + \hat{N}_S(x(t_1),y(t_1),z(t_1)), \tag{14}$$

$$\|p_1\|_\infty + \|p_2\|_\infty + \|\eta\|_\infty + \lambda > 0, \tag{15}$$

where $\hat{f}(x,y,z) := f(x)$.

Let $\hat{F}(x,y,z) = g(x)+y-V(z)$. Then by Lemma 4.3, one has

$$\hat{\partial}\hat{F}(x,y,z) \subset \hat{\partial}g(x)\times\{1\}\times\hat{\partial}(-V(z)). \tag{16}$$

Therefore $0\notin\hat{\partial}\hat{F}(x,y,z)$.

Let $S_1 := \{(x,y,z) : g(x)+y-V(z)\le 0\}$ and $S_2 := C_1\times I\!R\times I\!R$. By Lemma 4.2 and inclusion (16), one has

$$\begin{aligned} \hat{N}_{S_1}(x,y,z) &\subset \bigcup_{r\ge 0} r\hat{\partial}\hat{F}(x,y,z)\\ &\subset \bigcup_{r\ge 0} r[\hat{\partial}g(x)\times\{1\}\times\hat{\partial}(-V)(z)]. \end{aligned}$$

Since $\Psi_{S_2}(x,y,z) = \Psi_{C_1}(x) + \Psi_{I\!R}(y) + \Psi_{I\!R}(z)$, by Lemma 4.3 and (e) of Proposition 2.1 one has

$$\hat{N}_{S_2}(x,y,z) \subset \hat{N}_{C_1}(x)\times\{0\}\times\{0\} \qquad \forall (x,y,z)\in C_1\times I\!R\times I\!R.$$

It follows that the second component of any triple in the set $-\hat{N}_{S_2}(x,y,z)$ is 0. The only vectors in $\hat{N}_{S_1}(x,y,z)$ that share this property are among those for which $r=0$ in the estimate above. Thus, $\hat{N}_{S_1}(x,y,z)\cap(-\hat{N}_{S_2}(x,y,z)) = \{0\}$ and Proposition 2.1 (f) gives

$$\begin{aligned} \hat{N}_S(x(t_1),y(t_1),z(t_1)) &\subset \hat{N}_{S_1}(x(t_1),y(t_1),z(t_1)) + \hat{N}_{S_2}(x(t_1),y(t_1),z(t_1))\\ &\subset \bigcup_{r\ge 0} r[\hat{\partial}g(x(t_1))\times\{1\}\times\hat{\partial}(-V)(z)]\\ &\quad + \hat{N}_{C_1}(x(t_1))\times\{0\}\times\{0\}. \end{aligned}$$

By Lemma 4.3, one has

$$\hat{\partial}\hat{f}(x(t_1), y(t_1), z(t_1)) \subset \hat{\partial}f(x(t_1)) \times \{0\} \times \{0\}.$$

Hence from (14), one has

$$\begin{aligned}
-(p_1(t_1), p_2(t_1), \eta(t_1)) \in\ & \lambda\hat{\partial}f(x(t_1)) \times \{0\} \times \{0\} \\
& + \bigcup_{r\geq 0} r[\hat{\partial}g(x(t_1)) \times \{1\} \times \hat{\partial}(-V)(z)] \\
& + \hat{N}_{C_1}(x(t_1)) \times \{0\} \times \{0\} \\
\subset\ & \lambda\hat{\partial}f(x(t_1)) \times \{0\} \times \{0\} \\
& + \bigcup_{r\geq 0} r[\hat{\partial}g(x(t_1)) \times \{1\} \times (-\partial V(z))] \\
& + \hat{N}_{C_1}(x(t_1)) \times \{0\} \times \{0\},
\end{aligned}$$

from which the transversality conditions (6) and (7) follow and one has $p_2(t_1) = -r$, where $r \geq 0$. Since H is independent of y, (11) implies that $\dot{p_2}(t) = 0$ and

$$(17) \quad -(\dot{p_1}(t), \dot{\eta}(t)) \in \partial_{(x,z)}H(t, x(t), y(t), z(t), p_1(t), p_2(t), \eta(t), u(t); \lambda) \qquad \text{a.e.}$$

Hence $p_2 \equiv -r$, where $r \geq 0$; and (4), (5), and (8) follow from (17), (12), and (15), respectively. From (13), one has $\eta(t_0) = 0$. The proof of the theorem is thus complete. □

Combining Theorem 4.1 and Theorem 3.1, one has the following necessary conditions for optimality for the general bilevel dynamic optimization problem.

THEOREM 4.4. *In addition to assumptions* (A1)–(A5), *suppose either assumptions* (A6)–(A7) *or* (A6)′–(A7)′ *hold. Let* (z, u) *be an optimal strategy of the bilevel dynamic optimization problem* P_1 *and* $x(t)$ *the corresponding trajectory. Suppose that* $QM^0(Y) = \{0\}$. *Then there exist scalars* $\lambda \geq 0, r \geq 0$, *integers* I, J, $\lambda_{ij} \geq 0$, $\sum_{i=1}^{I}\sum_{j=1}^{J}\lambda_{ij} = 1$, *optimal trajectories* $x_i(t)$ *of the lower-level problem* $P_2(z)$, *and arcs* $p_1, \eta, p_2^{ij}, q^{ij}$ *such that*

$$\begin{aligned}
(18) \quad & -(\dot{p_1}(t), \dot{\eta}(t)) \in \partial_{(x,z)}H_1(t, x(t), z, p_1(t), u(t); \lambda, r) \qquad \textit{a.e.}, \\
& \max_{u\in U(t)} H_1(t, x(t), z, p_1(t), u; \lambda, r) = H_1(t, x(t), z, p_1(t), u(t); \lambda, r) \qquad \textit{a.e.}, \\
& \eta(t_0) = 0, \\
& -p_1(t_1) \in \lambda\hat{\partial}f(x(t_1)) + r\hat{\partial}g(x(t_1)) + \hat{N}_{C_1}(x(t_1)), \\
& \eta(t_1) = r\sum_{ij}\lambda_{ij}q^{ij}(t_0); \\
(19) \quad & (-\dot{p}_2^{ij}(t), -\dot{q}^{ij}(t), \dot{x}_i(t)) \in \partial_{(x,z,p_2)}H_2(t, x_i(t), z, p_2^{ij}(t); 1) \qquad \textit{a.e.}, \\
& q^{ij}(t_1) = 0, \\
& -p_2^{ij}(t_1) \in \hat{\partial}g(x_i(t_1)) + \hat{N}_{C_1}(x_i(t_1)), \\
& \|p_1\|_\infty + \|\eta\|_\infty + \lambda + r > 0.
\end{aligned}$$

Remark 4.1. A sufficient condition for $QM^0(Y) = \{0\}$ to hold is $C_1 = I\!R^d$. Indeed, in this case, the index 0 multiplier set consists of all arcs (p_2, q) such that

$$(20) \quad (-\dot{p_2}(t), -\dot{q}(t), \dot{x}(t)) \in \partial_{(x,z,p_2)}H_2(t, x(t), z, p_2(t); 0) \qquad \text{a.e.}$$

$$p_2(t_1) = 0, \tag{21}$$

$$q(t_1) = 0. \tag{22}$$

Due to the Lipschitz continuity of ϕ in (x, z), by virtue of Theorem 2.8.2 of Clarke [4], (20) implies that

$$\|\dot{p_2}(t)\| \leq k(t)\|p_2(t)\|.$$

By Gronwall's Lemma, the above inequality implies that p_2 is either identically 0 or nonvanishing on $[t_0, t_1]$. Therefore (21) implies that $p_2 \equiv 0$. Hence $\dot{q}(t) = 0$ by virtue of (20). But q satisfies (22), therefore $q \equiv 0$. That is $QM^0(Y) = \{0\}$.

Another sufficient condition for $M^0(Y) = \{0\}$ to hold is that $\phi(t, x, z, u)$ be independent of z since in this case $q(t) \equiv 0$.

Remark 4.2. By Clarke [4, Thm. 2.8.2], $\partial_{(x,\alpha,p_2)}H_2(t, x, \alpha, p_2; 1)$ is the convex hull of all points of the form

$$[\nabla_x\phi(t,x,\alpha,u)^\top p_2 - \nabla_x G(t,x,\alpha,u), \nabla_\alpha\phi(t,x,\alpha,u)^\top p_2 - \nabla_\alpha G(t,x,\alpha,u), \phi(t,x,\alpha,u)],$$

where u in $U(t)$ is any point at which the maximum defining $H_2(t, x, \alpha, p_2; 1)$ is achieved. Therefore if in addition to assumptions (A1)–(A5) and (A6)′–(A7)′, we assume the set

$$\{(\nabla_x\phi(t,x,\alpha,u)^\top p_2 - \nabla_x G(t,x,\alpha,u), \nabla_\alpha\phi(t,x,\alpha,u)^\top p_2 - \nabla_\alpha G(t,x,\alpha,u) : u \in U(t)\}$$

is convex for any t, x, z, p_2, then the inclusion (19) becomes the following equations:

$$\begin{aligned}
-\dot{p}_2^{ij}(t) &= \nabla_x\phi(t, x_i(t), z, u_i(t))^\top p_2^{ij}(t) - \nabla_x G(t, x_i(t), z, u_i(t)) \qquad \text{a.e.,}\\
-\dot{q}^{ij}(t) &= \nabla_z\phi(t, x_i(t), z, u_i(t))^\top p_2^{ij}(t) - \nabla_z G(t, x_i(t), z, u_i(t)) \qquad \text{a.e.,}\\
&\max_{u\in U(t)} \{p_2^{ij}(t) \cdot \phi(t, x_i(t), z, u) - G(t, x_i(t), z, u)\}\\
&\quad = p_2^{ij}(t) \cdot \phi(t, x_i(t), z, u_i(t)) - G(t, x_i(t), z, u_i(t)) \qquad \text{a.e.,}\\
&\dot{x}_i(t) = \phi(t, x_i(t), z, u_i(t)) \qquad \text{a.e.,}
\end{aligned}$$

where $u_i(t)$ is an optimal control function associated with trajectory $x_i(t)$.

5. Extensions and an example.

There are many situations where an opportunity cost exists for the follower. That is, the follower will participate only if his optimal cost is less than or equal to the opportunity cost $L \geq 0$ that he may receive from somewhere else. In this case, the leader faces the following bilevel optimization problem:

$$\begin{aligned}
\overline{P}_1 \qquad \min J_1(z,u) &= \int_{t_0}^{t_1} F(t, x(t), z(t), u(t))dt + f(x(t_1)),\\
\text{s.t. } \dot{x}(t) &= \phi(t, x(t), z(t), u(t)) \qquad \text{a.e.,}\\
\dot{z}(t) &= 0,\\
x(t_0) &= x_0, \quad x(t_1) \in C_1,\\
u(t) &\in U(t) \qquad \text{a.e.,}\\
\int_{t_0}^{t_1} G(t, x(t), z(t), u(t))dt &+ g(x(t_1)) \leq V(z),\\
\int_{t_0}^{t_1} G(t, x(t), z(t), u(t))dt &+ g(x(t_1)) \leq L.
\end{aligned}$$

The technique described in the previous section can be applied to this more general problem in exactly the same way, and one obtains the following necessary conditions for optimality.

THEOREM 5.1. *Assume that in addition to* (A1)–(A5), *either assumptions* (A6)–(A7) *or* (A6)′–(A7)′ *hold. Let* (z, u) *be an optimal strategy of the bilevel dynamic optimization problem* $\overline{P}_1$ *and* $x(t)$ *the corresponding trajectory. Suppose that* $QM^0(Y) = \{0\}$. *Then there exist scalars* $\lambda \geq 0, r \geq 0, 0 \leq \hat{r} \leq r$, *integers* I, J, $\lambda_{ij} \geq 0$, $\sum_{i=1}^{I} \sum_{j=1}^{J} \lambda_{ij} = 1$, *optimal trajectories* $x_i(t)$ *of the lower-level problem* $P_2(z)$, *and arcs* $p_1, \eta, p_2^{ij}, q^{ij}$ *such that*

$$-(\dot{p}_1(t), \dot{\eta}(t)) \in \partial_{(x,z)} H_1(t, x(t), z, p_1(t), u(t); \lambda, r) \qquad a.e.,$$

$$\max_{u \in U(t)} H_1(t, x(t), z, p_1(t), u; \lambda, r) = H_1(t, x(t), z, p_1(t), u(t); \lambda, r) \qquad a.e.,$$

$$\eta(t_0) = 0,$$

$$-p_1(t_1) \in \lambda \hat{\partial} f(x(t_1)) + r \hat{\partial} g(x(t_1)) + \hat{N}_{C_1}(x(t_1)),$$

$$\eta(t_1) = \hat{r} \sum_{ij} \lambda_{ij} q^{ij}(t_0),$$

$$(-\dot{p}_2^{ij}(t), -\dot{q}^{ij}(t), \dot{x}_i(t)) \in \partial_{(x,z,p_2)} H_2(t, x_i(t), z, p_2^{ij}(t); 1) \qquad a.e.,$$

$$q^{ij}(t_1) = 0,$$

$$-p_2^{ij}(t_1) \in \hat{\partial} g(x_i(t_1)) + \hat{N}_{C_1}(x_i(t_1)),$$

$$\|p_1\|_\infty + \|\eta\|_\infty + \lambda + r > 0.$$

The following example is a simplified and finite horizon version of a fishery regulation problem first formulated and solved by Clarke and Munro using principal and agent analysis (see Clarke and Munro [7] and [8] for details).

Example. It has now been generally agreed that the fishery resources within the 200-mile zones are the property of the adjacent coastal states. For those coastal states opting to permit a distant water presence in their 200-mile zones, one of the problems they face is devising optimum terms and conditions of access to the Coastal State Exclusive Economic Zones to be imposed upon the distant water fleets.

Assume that the fish population follows the dynamic system

$$\dot{x}(t) = F(x(t)) - qE(t)x(t),$$

where $x(t)$ is the fish population at time t; $F(x)$ is the rate of natural growth; and $qE(t)x(t)$ is the rate of catch at time t, where $E(t)$ is the fishing effort at time t and q is a positive constant. We assume that $F(x)$ is a twice continuously differentiable function satisfying $F(x) > 0$ for $0 < x < \bar{x}, F(0) = F(\bar{x}) = 0$ and $F''(x) < 0$ for all $x > 0$, where $\bar{x}$ denotes the carrying capacity of the resource. It is also assumed that

$$0 \leq E(t) \leq E_{\max},$$

where $E_{\max}$ is an arbitary upper bound on $E(t)$. Suppose that the coastal state imposes the condition that at the terminal time T_1, the fish population cannot be less than $\widetilde{x} \geq 0$.

Suppose that the coastal states as a leader impose a unit tax n on catch $qE(t)x(t)$ and a unit tax m on effort $E(t)$. Then the distant water fleet would receive the profit in time period $[0, T_1]$

$$\int_0^{T_1} e^{-\delta t}[(p_0 - n)qx(t) - (c_0 + m)]E(t)dt$$

if he decided to use the fishing effort $E(\cdot)$ where p_0 and c_0 are the unit price on catch and unit cost on effort, respectively, $\delta > 0$ is the discount rate, and $x(\cdot)$ is the fish population corresponding to the fishing effort $E(\cdot)$. Hence for the given unit tax on catch and effort n and m, the distant water fleet as a follower faces the following optimal control problem:

$$\begin{aligned} P_2(n,m) \qquad & \max \int_0^{T_1} e^{-\delta t}[(p_0 - n)qx(t) - (c_0 + m)]E(t)dt, \\ & \text{s.t. } \dot{x}(t) = F(x(t)) - qE(t)x(t), \\ & x(0) = x_0, \quad x(T_1) \geq \widetilde{x}, \\ & E(t) \in [0, E_{\max}]. \end{aligned}$$

The optimal control problem $P_2(n,m)$ is linear. The necessary condition for (x, E) to solve $P_2(n,m)$ is the existence of an arc p_2 such that

$$-\dot{p_2}(t) = p_2(t)[F'(x(t)) - qE(t)] + e^{-\delta t}(p_0 - n)qE(t), \tag{23}$$

$$\begin{aligned} &\max_{E \in [0, E_{\max}]} \{p_2(t)[F(x(t)) - qEx(t)] + e^{-\delta t}[(p_0 - n)qx(t) - (c_0 + m)]E\} \\ &\quad = p_2(t)[F(x(t)) - qE(t)x(t)] + e^{-\delta t}[(p_0 - n)qx(t) - (c_0 + m)]E(t), \end{aligned} \tag{24}$$

$$p_2(T_1) \geq 0.$$

Since $E(t)$ has to maximize the Hamiltonian (see (24)), $E(t)$ must be either the singular control or else $E(t) = 0$ or $E_{\max}$. The singular control arises when the coefficient of E in the Hamiltonian is zero, implying that

$$p_2(\dot{t}) = e^{-\delta t}\left[(p_0 - n) - \frac{c_0 + m}{qx}\right] \tag{25}$$

$$\dot{p_2}(t) = e^{-\delta t}\left[-\delta\left[(p_0 - n) - \frac{c_0 + m}{qx}\right] + \frac{c_0 + m}{qx^2}\frac{dx}{dt}\right]. \tag{26}$$

From the adjoint equation (23), one has

$$\begin{aligned} \dot{p_2} &= -p_2[F'(x) - qE] - e^{-\delta t}[(p_0 - n)qE \\ &= -e^{-\delta t}\left\{\left[(p_0 - n) - \frac{c_0 + m}{qx}\right][F'(x) - qE] + (p_0 - n)qE\right\}, \end{aligned} \tag{27}$$

where (25) is used for p_2. When the two expressions for $\dot{p_2}(t)$, (26) and (27), are equated, the control variable E cancels out and the following equation emerges:

$$F'(x) + \frac{F(x)(c_0 + m)/qx^2}{p_0 - n - (c_0 + m)/qx} = \delta. \tag{28}$$

For fixed (n, m), this equation gives a unique solution x_* that is the optimal biomass and the optimal trajectory is the one that takes the most rapid path to the optimal biomass x_* (cf. Clark [3]).

Let $V(n,m)$ be the optimal value of the above problem. The distant water fleet will participate only when $V(n,m) \geq L$, the alternative remuneration from some other coastal state.

The coastal state as a leader now faces the following bilevel dynamic optimization problem:

$$
\begin{aligned}
P_1 \qquad & \max \int_0^{T_1} e^{-\delta t}(nqx(t)+m)E(t)dt, \\
& \text{s.t. } \dot{x}(t) = F(x(t)) - qE(t)x(t), \\
& x(0) = x_0, \quad x(T_1) \geq \widetilde{x}, \\
& E(t) \in [0, E_{\max}] \qquad \text{a.e.}, \\
V(n,m) \leq & \int_0^{T_1} e^{-\delta t}[(p_0-n)qx(t) - (c_0+m)]E(t)dt, \\
& V(n,m) \geq L.
\end{aligned}
$$

It is easy to show that all the conditions of Theorem 5.1 are satisfied. Notice that the lower-level problem $P(n,m)$ has a unique solution. By Theorem 5.1 and Remark 3.2, if (n,m,x,E) is an optimal solution to P_1, then there exist arcs $p_1, p_2, \eta_1, \eta_2, q_1, q_2$ and scalars $\lambda \geq 0$, $r \geq 0$, $0 \leq \hat{r} \leq r$ such that

$$-\dot{p_1} = p_1[F'(x) - qE) + e^{-\delta t}[r(p_0-n) + \lambda n]qE, \tag{29}$$

$$
\begin{aligned}
&\dot{\eta_1} = (r-1)e^{-\delta t}qxE, \\
&\dot{\eta_2} = (r-1)e^{-\delta t}E, \\
&\max_{E \in [0,E_{\max}]} \{p_1(t)[F(x(t)) - qEx(t)] + e^{-\delta t}[r[(p_0-n)qx(t) - (c_0+m)] \\
&\qquad\qquad\qquad +\lambda(nqx(t)+m)]E\} \\
&= p_1(t)[F(x(t)) - qE(t)x(t)] + e^{-\delta t}[r[(p_0-n)qx(t) - (c_0+m)] \\
&\qquad\qquad\qquad +\lambda(nqx(t)+m)]E(t),
\end{aligned}
\tag{30}
$$

$$(\eta_1, \eta_2)(0) = (0,0),$$

$$p_1(T_1) \geq 0, \tag{31}$$

$$(\eta_1, \eta_2)(T_1) = \hat{r}q(0),$$

$$-\dot{p_2} = p_2[F'(x) - qE] + e^{-\delta t}(p_0-n)qE, \tag{32}$$

$$
\begin{aligned}
&\dot{q_1} = e^{-\delta t}qxE, \\
&\dot{q_2} = e^{-\delta t}E,
\end{aligned}
$$

$$
\begin{aligned}
&\max_{E \in [0,E_{\max}]} \{p_2(t)[F(x(t)) - qEx(t)] + e^{-\delta t}[(p_0-n)qx(t) - (c_0+m)]E\} \\
&= p_2(t)[F(x(t) - qE(t)x(t)] + e^{-\delta t}[(p_0-n)qx(t) - (c_0+m)]E(t),
\end{aligned}
\tag{33}
$$

$$(q_1, q_2)(0) = (0,0),$$

$$p_2(T_1) \geq 0, \tag{34}$$

$$\|p_1\|_\infty + \|\eta\|_\infty + \lambda + r > 0.$$

Take $\lambda = 1$. As in the proof of (28), from (29) and (30) we can show that the steady state (n,m,x_*) for problem P_1 is a solution of the following equation:

$$F'(x_*) + \frac{F(x_*)(r(c_0+m) - m)/q{x_*}^2}{r(p_0-n) + n - (r(c_0+m)-m)/qx_*} = \delta, \tag{35}$$

and the optimal trajectory for P_1 is the one that takes the most rapid path to the optimal biomass (n, m, x_*). Since (n, m, x, E) is an optimal solution of P_1, (x, E) must be the optimal solution of the lower-level problem $P_2(n, m)$. Therefore x_* must be the optimal biomass associated with (n, m) defined by (28). Combining equations (28) and (35), one has

$$n = \rho p_0,$$
$$m = -\rho c_0,$$

where ρ is some constant to be determined. It is obvious that the optimal tax (n, m) must be such that $V(n, m) = L$. Let V_0 be the net global returns from the fishery, i.e.,

$$V_0 = \max \left\{ \int_0^{T_1} e^{-\delta t}(p_0 q x(t) - c_0) E(t) dt \right\}.$$

Then

$$\begin{aligned} (1-\rho)V_0 &= \max \left\{ \int_0^{T_1} e^{-\delta t}[(1-\rho)p_0 q x(t) - (1-\rho)c_0] E(t) dt \right\} \\ &= \max \left\{ \int_0^{T_1} e^{-\delta t}[(p_0 - n) q x(t) - (c_0 + m)] E(t) dt \right\} \\ &= V(n, m) = L, \qquad (36) \end{aligned}$$

from which it follows that $\rho = (V_0 - L)/V_0$. (36) also indicates that $E(t)$ will maximize the global net returns from the fishery. Hence the above necessary condition for optimality is indeed satisfied by $\lambda = 1$, $r = 1$, $\hat{r} = 0$, $n = \rho p_0$, $m = -\rho c_0$, and the corresponding fishing effort $E(t)$ since equations (29), (30), and (31) are necessary for $E(t)$ to maximize the net global returns from the fishery; (32), (33), and (34) are the necessary optimality conditions for the lower-level problem; and the rest of equations are easily seen to hold. The results agree with the work of Clarke and Munro [7].

Acknowledgments. The author would like to thank Philip Loewen for many helpful suggestions that led to improvements in the results of this paper and two anonymous referees for comments that led to the improvements in the style of the presentation in this version.

REFERENCES

[1] J. F. Bard and J. E. Falk, *An explicit solution to the multi-level programming problem*, Oper. Res., 9 (1982), pp.77–100.

[2] C. I. Chen and J. B. Cruz Jr., *Stackelberg solution for two-person games with baised information patterns*, IEEE Trans. Automat. Control, 6 (1972), pp. 791–798.

[3] C. W. Clark, *Mathematical Bioeconomics: The Optimal Management of Renewable Resources*, 2nd ed., John Wiley and Sons, New York, 1990.

[4] F. H. Clarke, *Optimization and Nonsmooth Analysis*, Wiley-Interscience, New York, 1983.

[5] ———, *Methods of Dynamic and Nonsmooth Optimization*, NSF-CBMS Regional Conf. Ser. in Appl. Math. 57, Society for Industrial and Applied Mathematics, Philadelphia, 1989.

[6] F. H. Clarke and P. D. Loewen, *The value function in optimal control: Sensitivity, controllability, and time-optimality*, SIAM. J. Control Optim., 24 (1986), pp. 243–263.

[7] F. H. Clarke and G. R. Munro, *Coastal states, distant water fishing nations and extended jurisdiction: A principal-agent analysis*, Natural Resource Modeling, 2 (1987), pp. 87–107.

[8] F. H. CLARKE AND G. R. MUNRO, *Coastal states and distant water fishing nations: Conflicting views of the future*, Natural Resource Modeling, to appear.

[9] R. T. ROCKAFELLAR, *Extensions of subgradient calculus with applications to optimization*, Nonlinear Analysis, Theory, Methods Appl., 9 (1985), pp. 665–698.

[10] H. VON STACKELBERG, *The Theory of the Market Economy*, Oxford University Press, Oxford, 1952.

[11] J. J. YE, *Perturbed infinite horizon optimal control problems*, J. Math. Anal. Its Appl., 182 (1994), pp.90–112.

[12] J. J. YE AND D. L. ZHU, *Optimality conditions for bilevel programming problems*, Optimization, to appear.

[13] R. ZHANG, *Problems of Hierarchical Optimization: Nonsmoothness and Analysis of Solutions*, Ph.D. Thesis, University of Washington, 1990.

[14] ———, *Problems of hierarchical optimization in finite dimensions*, SIAM J. Optim., 4 (1994), pp. 521–536.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 33, No. 4, pp. 1224–1246, July 1995
© 1995 Society for Industrial and Applied Mathematics
014

USING PERSISTENT EXCITATION WITH FIXED ENERGY TO STABILIZE ADAPTIVE CONTROLLERS AND OBTAIN HARD BOUNDS FOR THE PARAMETER ESTIMATION ERROR*

MILOJE S. RADENKOVIC† AND B. ERIK YDSTIE‡

Abstract. Two important instability problems in certainty equivalence adaptive control are solved by external excitation. The first instability is parameter drift along an unstable manifold when the excitation level is not high enough. The second instability is numerical and due to a division with zero in the adaptive law. Global methods based on excitation have been developed to solve this problem, but the energy of the excitation has been tuned on-line. The main contribution of the current paper is in showing that the estimator is stabilized when we apply excitation with *fixed and finite energy*. The level of excitation should be sufficiently high relative to the magnitudes of the external disturbances and the unmodeled dynamics. The approach can be generalized to more complex adaptive laws. This, together with the fact that we obtain hard bounds for the parameter estimation error, opens up for the possibility of designing robust controllers that are adaptive.

Key words. adaptive control, self-tuning regulator, stability, robustness, learning, excitation

AMS subject classifications. 93A, 93D, 93E

1. Introduction.

In this paper we present the global stability analysis of a direct adaptive control system for which a persistent excitation condition is satisfied. By *global* we mean that the results are valid for all initial conditions. In this way the analysis complements averaging methods which are initial-condition dependent. Averaging has been applied for the *local* analysis of adaptive systems with considerable success, and sufficient as well as necessary conditions for the stability of integral manifolds have been obtained by exploiting hyperbolicity [2], [12], [14]. The degree of locality of these results has not been determined. But it is an important problem worth investigating since averaging can give tight results. It comes as a surprise to the authors of this paper that averaging in adaptive control can have global validity when the singularities that may arise in the control design equations are avoided.

It was recognized early that the parameter estimates in certainty-equivalent adaptive control may drift, become unbounded, or cross through regions where the control law calculation is ill conditioned. The reason for this rather disappointing behaviour is that the adaptive controller does not destabilize in such a way that excitation is generated. In the ideal case the parameters cannot diverge [4]. However, self-stabilization of the estimator does not take place, and in the nonideal case small-amplitude, chaotic bursts as well as large transients and numerical instabilities may be observed [3], [15]. The drift problems can be solved by using artificial parameter bounding. In the case of leakage [6] and parameter projection [7] the parameters are constrained to belong to compact sets and additional measures are taken to ensure that the control law calculations are well behaved. Unfortunately, while boundedness of the input/output signals can be established, these methods do not ensure that the parameter estimator is stabilized and large bursts are possible. A number of methods for solving the singularity problem have been proposed. These include the adaptive excitation approach

* Received by the editors November 16, 1992; accepted for publication (in revised form) February 15, 1994. This research was supported in part by National Science Foundation contract NSF#CTS-8903160.

† Department of Electrical Engineering, University of Colorado at Denver, Denver, Colorado 80204.

‡ Department of Chemical Engineeering, Carnegie Mellon University, Doherty Hall, Pittsburgh, Pennsylvania 15213.

[1], parameter projection [7] approaches where the high-frequency gain is modified [5], and a recent method where the controller gains are switched to maintain a certain detectability property [8].

It is possible to solve the drift and singularity problems by ensuring that the estimator equations remain stable. The estimated parameters are then close to optimal and singularities and drift are avoided. The sufficient and necessary condition for such stabilization to take place is that the regression vector is persistently excited [14]. Since excitation is not automatically generated within the adaptive loop, the only possibility that remains is to generate the excitation externally. In this paper we do this by manipulating the reference signal using fixed and finite energy over a range of selected frequencies. Once excitation is generated this leads not only to stabilization of the adaptive loop but also to good conditions for the estimator. The parameter drift is arrested, and small parameter errors can be guaranteed. Reasonable system conditions can then be imposed to ensure that singularity problems in the solution of the Bezout identities do not arise, and methods can be developed to start and stop the estimator. Thus we do not only solve the problem of adaptive stabilization, we also solve the problem of identifying parameters in an uncertain environment. The latter problem has attracted considerable attention in the recent literature. See for example the recent IEEE T-AC special issue on identification. However, we believe that results developed in this paper are among the first that can be applied for the identification of an open-loop unstable plant.

We analyze a simple one step–step-ahead predictive controller with gradient estimator. Two types of data normalization are used in the analysis. In one instance we use the exponentially weighted normalization [11] to show robustness of adaptive control with respect to unmodeled dynamics and bounded noise. This approach requires knowledge of an upper bound for the largest time constant for the closed-loop system with an ideal controller. In order to relax this assumption we also analyse a second type of normalization sequence that does not require this type of a priori information. The idea here is to use the largest regressor up to time t to normalize the signals.

The results of the paper can be generalized to continuous time. Thus the projection used in [9] can be replaced with finite energy excitation of the reference.

2. Notation and terminology. For a function $x : T \longrightarrow R^+$, we define the following semi-norm

$$\|x(t)\|_\lambda = \left\{ \sum_{j=1}^{t} \lambda^{t-j} x(j)^2 \right\}^{1/2}, \qquad 0 < \lambda < 1.$$

Here T is the set of positive integers, while R^+ is the set of nonnegative real numbers. When $\|x(t)\|_\lambda$ is bounded for all $t \geq 0$, x is said to be in l_2^λ.

H_∞ will denote the space of transfer functions $T(z)$, which are analytic and bounded outside and on the unit circle in the z-plane. S^λ is the operator defined by

$$S^\lambda T(z) = T(\lambda^{1/2} z).$$

H_∞^λ is the space of transfer functions $T(z)$ such that $S^\lambda T(z) \in H_\infty$. In other words, $T(z) \in H_\infty^\lambda$, if $T(z)$ is analytic and bounded outside and on the circle $|z| = \lambda^{1/2}$ in the z-plane. For $T(z) \in H_\infty$, the H_∞ norm is defined by

$$\|T(z)\|_\infty = \max_{|z|=1} |T(z)|.$$

Likewise, the norm of the H_∞^λ space is defined by

$$\|T(z)\|_\infty^\lambda = \|S^\lambda T(z)\|_\infty = \max_{|z|=1} |T(\lambda^{1/2} z)|.$$

This norm is induced by the l_2^λ norm of the input and output signals of $T(z)$.

When performing majorizations in order to account for initial conditions, we use nonnegative functions

$$\xi_i(t) = c_i \beta_i^t, \qquad 0 \le c_i < \infty, \qquad 0 < \beta_i < 1.$$

When confusion cannot arise, we drop the subscripts.

3. Deterministic adaptive control and major assumptions. Let us consider the following discrete-time single-input single-output (SISO) system with unmodeled dynamics

$$(1) \quad A(q^{-1})y(t+1) = B(q^{-1})[1 + \Delta_1(q^{-1})]u(t) + A(q^{-1})\Delta_2(q^{-1})u(t) + w(t+1)$$

where $\{y(t)\}, \{u(t)\}$, and $\{w(t)\}$ are output, input, and disturbance sequences, respectively, while q^{-1} represents the unit delay operator. The polynomials $A(q^{-1})$ and $B(q^{-1})$ describe the nominal system model, which may be taken as being "centered" [16], and can be written as

$$A(q^{-1}) = 1 + a_1 q^{-1} + \cdots + a_{n_A} q^{-n_A},$$

$$B(q^{-1}) = b_0 + \cdots + b_{n_B} q^{-n_B}, \text{ with } b_0 \neq 0.$$

In equation (1), $\Delta_i(q^{-1})$ for $i = 1, 2$ denote multiplicative and additive system perturbations. The transfer functions $\Delta_i(q^{-1})$ for $i = 1, 2$ are causal and $\Delta_1(q^{-1})$ is stable.

The aim is to stabilize simultanously the input–output behaviour of the system (1) and estimate the parameters of the nominal model to a given precision. This is the problem of identification of open-loop unstable systems.

The nominal model is assumed to be stably invertible, the model mismatch small, the external perturbations bounded, and the reference signal excited with sufficient energy in a selected range of frequencies. In order to develop the theory we will apply direct adaptive control to minimize the criterion

$$J = [y(t) - y^*(t)]^2$$

where $y^*(t)$ is the given reference signal.

To define the adaptive law it is convenient to write system (1) in the form

$$(2) \qquad y(t+1) = \theta_0' \phi(t) + \gamma(t)$$

where

$$\theta_0' = (-a_1, -a_2, \ldots, -a_{n_A}, b_0, b_1, \ldots, b_{n_B})$$

is the vector of parameters that need to be estimated,

$$\phi(t)' = (y(t), y(t-1), \ldots, y(t - n_A + 1), u(t), u(t-1), \ldots, u(t - n_B))$$

is the regression vector, and finally, the modeling error is defined so that

$$\gamma(t) = \Delta_0(q^{-1})u(t) + w(t+1) \tag{3}$$

with

$$\Delta_0(q^{-1}) = B(q^{-1})\Delta_1(q^{-1}) + A(q^{-1})\Delta_2(q^{-1}).$$

From equation (2) it is obvious that when $\gamma(t) = 0$, the control law

$$\text{Solve for} \quad u(t): \quad \theta_0'\phi(t) - y^*(t+1) = 0 \tag{4}$$

is optimal. Internal stability is ensured by the stable invertibility of $B(q^{-1})$.

In certainty-equivalent adaptive control the parameter vector θ_0 is replaced by an estimate

$$\theta(t)' = (-\hat{a}_1(t), -\hat{a}_2(t), \ldots, -\hat{a}_{n_A}(t), \hat{b}_0(t), \hat{b}_1(t), \ldots, \hat{b}_{n_B}(t)).$$

The following algorithm is used for estimating the unknown parameter θ_0.

$$\theta(t+1) = \theta(t) + \frac{\mu}{r(t)}\phi(t)e(t) \tag{5}$$

where

$$e(t) = y(t) - y^*(t) \tag{6}$$

is the tracking error. The algorithm gain sequence may be given by

$$r(t) = r_0 + n_\phi(t)^2, \text{ with } 0 < r_0 < \infty$$

and

$$n_\phi(t)^2 = \lambda n_\phi(t-1)^2 + \|\phi(t)\|^2 \tag{7}$$

where λ is a tunable parameter chosen so that $0 < \lambda < 1$. Note that this gain sequence is similar to that proposed in [11]. In [16] the gain sequence is defined with $\lambda = 0$ so that

$$r(t) = r_0 + \|\phi(t)\|^2.$$

In the current paper we also analyse an algorithm with gain sequence chosen so that

$$r(t) = r_0 + \max_{1 \le \tau \le t} \|\phi(\tau)\|^2. \tag{8}$$

This normalization sequence does not involve the parameter λ as is the case with n_ϕ defined by equation (7). On the other hand, $r(t)$ defined by equation (8) satisfies the same property as the sequence given by (7), namely $r(t) \ge \lambda r(t-1)$. As a consequence of this we should expect that the two algorithms behave in a similar fashion as long as $\phi(t)$ remains uniformly bounded. The main results concerning normalization sequence (7) are given in Theorem 5.1. The results concerning normalization sequence (8) are given in Theorem 5.2. The analysis for the normalization sequence (7) with $\lambda = 0$ can be carried out using the approach developed in [16] for the discrete-time case and in [9] for the continuous-time case.

Equation (4) is ill defined when the estimate of b_0 is equal to zero and measures need to be taken to prevent large transients and instability of the algorithm. One approach, which is the one we follow here, is to depart from certainty equivalence on the event $A(t) = I_{\{|\hat{b}_0(t)|<\epsilon_1\}}$, where $I_{\{\cdot\}}$ is the indicator function and ϵ_1 is specified below, and implement the law

$$\text{(9)} \qquad \text{Solve for} \quad u(t): \quad \theta_c(t)'\phi(t) - y^*(t+1) = 0$$

with

$$\text{(10)} \qquad \theta_c(t)' = (-\hat{a}_1(t), -\hat{a}_2(t), \ldots, -\hat{a}_{n_A}(t), \epsilon(t) + \hat{b}_0(t), \hat{b}_1(t), \ldots, \hat{b}_{n_B}(t)).$$

Here

$$\epsilon(t) = \begin{cases} 0 & \text{if } |\hat{b}_0(t)| \geq \epsilon_1 > 0, \\ \epsilon_1 \text{sign}(\hat{b}_0(t)) & \text{if } |\hat{b}_0(t)| < \epsilon_1. \end{cases}$$

The sign function is defined so that $\text{sign}(x) = 1$ if $x \geq 0$ and $\text{sign}(x) = -1$ otherwise. The approach is similar to that used by Lozano–Leal, Collado, and Mondie [5] for the analysis of robustness of model-reference adaptive control. Other approaches include parameter projection [7] and controller switching [8]. These approaches give boundedness even when the signals are not excited, provided that projection is used to maintain finite parameters. But the modification then remains active and poor transients may result. We show that with excitation we get

$$\sum_{t=1}^{\infty} A(t) < \infty.$$

The modification applies a finite number of times, and the algorithm converges to a certainty-equivalence algorithm.

Assumption A1 (Concerning the reference signal and the disturbances): There exist constants k_w and k_{y^*} so that for all $t \geq 1$:

$$|w(t)| \leq k_w \qquad \text{and} \qquad |y^*(t)| \leq k_{y^*}.$$

In order to motivate the second assumption we develop closed-loop expressions for the adaptive system. From equations (2) and (9) it follows that

$$\text{(11)} \qquad e(t+1) = -z(t) + \gamma(t)$$

where

$$\text{(12)} \qquad z(t) = \tilde{\theta}_c'(t)\phi(t)$$

and the control parameter error satisfies

$$\tilde{\theta}_c(t) = \theta_c(t) - \theta_0.$$

From equations (9) and (12) we obtain

$$B(q^{-1})u(t) + q(1 - A(q^{-1}))y(t) = y^*(t+1) - z(t).$$

Combining this with equations (6) and (11) gives

$$B(q^{-1})u(t) = A(q^{-1})(-z(t) + y^*(t+1)) + (A(q^{-1}) - 1)\gamma(t). \tag{13}$$

Substituting this into equation (3) gives the closed loop

$$\gamma(t) = -\frac{\Delta_0(q^{-1})A(q^{-1})}{B(q^{-1}) - \Delta_0(q^{-1})(A(q^{-1}) - 1))}(z(t) - y^*(t+1)) \tag{14}$$

$$+\frac{B(q^{-1})}{B(q^{-1}) - \Delta_0(q^{-1})(A(q^{-1}) - 1))}w(t+1).$$

The transfer function from the reference signal to the model error $\gamma(t)$ plays a signficant role in the stability analysis. It is assumed to be stable and have small gain. In order to discuss this assumption and perform the analysis, define the following H_∞ norms:

$$C_{AB} = \left\| \frac{A(z)}{B(z)} \right\|_\infty^\lambda, \qquad C_A = \left\| \frac{A(z) - 1}{B(z)} \right\|_\infty^\lambda,$$

$$C_\gamma = \left\| \frac{\Delta_0(z)A(z)}{B(z) - \Delta_0(z)(A(z) - 1)} \right\|_\infty^\lambda, \qquad C_w = \left\| \frac{B(z)}{B(z) - \Delta_0(z)(A(z) - 1)} \right\|_\infty^\lambda.$$

Assumption A2 (Concerning the nominal system and unmodeled dynamics):

1. There exists a positive number $\lambda_0 < 1$ such that the zeros of $B(z^{-1})$ and the poles of the transfer functions

$$H_0(z^{-1}) = \frac{\Delta_0(z^{-1})A(z^{-1})}{B(z^{-1}) - \Delta_0(z^{-1})(A(z^{-1}) - 1)}$$

and

$$H_1(z^{-1}) = \frac{B(z^{-1})}{B(z^{-1}) - \Delta_0(z^{-1})(A(z^{-1}) - 1)}$$

are inside a circle with radius λ_0.

2. The transfer function $H_0(q^{-1})$ has small gain in the sense that the inequality

$$\rho_1(\epsilon_1) = 1 - \frac{\mu}{2} - (1-\mu)C_\gamma - \frac{\mu}{2}C_\gamma^2 - \epsilon_1 C_u(1 + C_\gamma) > 0$$

with

$$C_u = C_{AB} + C_A C_\gamma$$

is satisfied.

It is well known that the adaptive control algorithm described above may not be stable when there is no excitation. The problem is due to the presence of an unstable manifold along which the parameter estimates may drift to infinity. The problem can be avoided by applying parameter projection or leakage. However, these methods may give large transients and unrelenting bursting, unless the region into which the parameters are projected is small or the leakage center is defined close to the optimal parameter values. An alternative method, which we explore here, is to supply additional excitation. This approach has been shown to work well locally, and using averaging theory, necessary and sufficient conditions for the stability of integral manifolds have been established. The purpose of our analysis is to extend these results to be valid globally. In order to do this we introduce the following assumption about the level of excitation.

Define

$$\phi^*(t)' = \left(y^*(t),\ldots,y^*((t-n_A+1),\frac{A(q^{-1})}{B(q^{-1})}y^*(t+1),\ldots,\frac{A(q^{-1})}{B(q^{-1})}y^*(t-n_B+1)\right) \tag{15}$$

Assumption A3 (Persistent excitation): For all sufficiently large N

$$\sum_{t=1}^{N}\lambda^{N-t}\phi^*(t)\phi^*(t)' \geq \delta_1^* I$$

where, for some δ_2^*,

$$\frac{\delta_1^*}{2} - \left(n_1\Sigma_\gamma(\epsilon_1) + n_2\left[C_\gamma\Sigma_\gamma(\epsilon_1) + (C_\gamma k_{y^*} + (1+C_w)k_w)\frac{1}{(1-\lambda)^{1/2}}\right]\right) = \delta_2^* > 0 \tag{16}$$

with

$$\Sigma_\gamma(\epsilon_1)^2 = \max\left\{\frac{16}{\rho_1(\epsilon_1)^2}[(1-\mu+\mu C_\gamma+\epsilon_1 C_u)\Sigma_1+\epsilon_1(1+C_\gamma)\Sigma_2]^2;\ \frac{\mu\Sigma_1}{\rho_1(\epsilon_1)}(\Sigma_1+2\epsilon_1\Sigma_2)\right\} \tag{17}$$

with

$$\Sigma_1 = \frac{k_{y^*}C_\gamma + k_w C_w}{(1-\lambda)^{1/2}} \quad \text{and} \quad \Sigma_2 = \frac{k_{y^*}C_u + k_w C_A C_w}{(1-\lambda)^{1/2}}.$$

The constants n_1 and n_2 are defined so that

$$n_1 = \left(\sum_{i=1}^{n_A}\lambda^{-i} + C_{AB}^2\sum_{i=1}^{n_B}\lambda^{-i}\right)^{1/2} \quad \text{and} \quad n_2 = \left(\sum_{i=1}^{n_A}\lambda^{-i} + C_A^2\sum_{i=1}^{n_B}\lambda^{-i}\right)^{1/2}.$$

Comments.

1. Assumption A1 simply states that the reference and the disturbances should be uniformly bounded.
2. It is not difficult to see that Assumption A2 can be written as

$$(1-C_\gamma)\left(\left(1-\frac{\mu}{2}(1-C_\gamma)\right)-\epsilon_1 C_u\right) > 0.$$

This relation is satisfied if

$$C_\gamma < 1 \tag{18}$$

and

$$1-\frac{\mu}{2}+\frac{\mu}{2}C_\gamma - \epsilon_1 C_{AB} - \epsilon_1 C_A C_\gamma > 0. \tag{19}$$

Relationship (19) will be satisfied if ϵ_1 is selected so that

$$\epsilon_1 < \max\left\{(1-\frac{\mu}{2})\frac{1}{C_{AB}};\ \frac{\mu}{2C_A}\right\}, \tag{20}$$

and the admissible unmodeled dynamics are then specified by inequality (18).

3. Assumption A3 is more complicated and essentially means the following. The intensities of the unmodeled dynamics C_γ, the external disturbances, and the design parameter ϵ_1 should be small compared with the level of excitation. Moreover, $y^*(t)$ should have a spectral distribution function that is nonzero at $n_A + n_B + 1$ points (or more), and the transfer function of the nominal system model $B(z)/A(z)$ should be irreducible. This condition is not stronger than the similar conditions introduced in deterministic adaptive control and coincides with those obtained from the application of the averaging analysis.
4. Since we do not know the magnitude of the disturbances, the intensity of the unmodeled dynamics, and H_∞-norms C_{AB} and C_A, we cannot select ϵ_1 so that relationship (20) holds. The immediate consequence of this is that it is difficult to choose the right level of excitation, and we are still some way off from the target of having a completely adaptive control algorithm. In the following we assume that ϵ_1 and the level of excitation is chosen so that this relationship holds true.

Constants whose values are unimportant and do not depend on C_γ and C_w will be denoted by $C_i, i = 1, 2, \ldots$. Constants whose values are unimportant but depend on C_γ and C_w will be denoted by $\bar{C}_i, i = 1, 2, \ldots$.

4. Technical results. The following three results are useful for future reference. Lemma 4.1 simply states that all signals are bounded by the l_2^λ norm of the error signal $z(t)$ plus constants. Lemma 4.2 states that when the l_2^λ norm of $z(t)$ is small, the signal vector $\phi(t)$ is persistently exciting. Finally, Lemma 4.3 states that under similar conditions the parameter error vector is small.

LEMMA 4.1. *Let Assumptions* A1 *and* A2 *hold. Then*

1. $\|\gamma(t)\|_\lambda \le C_\gamma \|z(t)\|_\lambda + h_1(t)$ *where*

$$h_1(t) = \frac{C_\gamma k_{y^*} + C_w k_w}{(1-\lambda)^{1/2}} + \xi_1(t).$$

2. $\|u(t)\|_\lambda \le C_u \|z(t)\|_\lambda + h_2(t)$ *with*

$$h_2(t) = \frac{C_u k_{y^*} + C_A C_w k_w}{(1-\lambda)^{1/2}} + \xi_2(t).$$

3. $\|\phi(t)\|_\lambda \le C_{\phi_1} \|z(t)\|_\lambda + h_3(t)$ *where*

$$h_3(t) = \frac{C_{\phi_1} k_{y^*} + C_{\phi_2} k_w}{(1-\lambda)^{1/2}} + \xi_3(t),$$

$$C_{\phi_1} = C_1(\lambda^{-1/2}(1 + C_\gamma) + C_u) \; and \; C_{\phi_2} = C_1 C_w(\lambda^{-1} + C_A),$$

$$C_1 = \left(\sum_{i=0}^{\max\{n_A, n_B+1\}} \lambda^{-i} \right)^{1/2}.$$

4. *If there exists finite* t_0 *such that for all* $t \ge t_0$, $\|\theta_c(t)\| \le f_\theta < \infty$, *then*

$$r(t) \le \max\{C_\theta \|z(t-1)\|_\lambda^2, k_\theta + \xi_4(t)\}$$

with

$$C_\theta = 4\left(C_1\left(1+\frac{f_\theta C_1}{\epsilon_1}\right)(1+C_\gamma+C_u)\right)^2$$

and

$$k_\theta = 4\left(r_0^{1/2} + C_1\left(1+\frac{f_\theta C_1}{\epsilon_1}\right)\left(\frac{(1+C_\gamma+C_u)k_{y^*}+C_w(1+C_A)k_w}{(1-\lambda)^{1/2}}\right)\left(1+\frac{k_{y^*}C_1}{\epsilon_1}\right)\right)^2$$

Proof. Statement 1 of the lemma follows from equation (14). By using (13) we get

$$\|u(t)\|_\lambda \le C_{AB}(\|z(t)\|_\lambda + \|y^*(t+1)\|_\lambda) + C_A\|\gamma(t)\|_\lambda + \xi(t). \tag{21}$$

After substituting from statement 1 we have statement 2. From the definition of the regressor and equation (7) it follows that

$$\|\phi(t)\|_\lambda \le C_1(\|y(t)\|_\lambda + \|u(t)\|_\lambda) + \xi(t) \tag{22}$$

where C_1 was defined under statement 3. From equations (6) and (11) and statement 1 of the lemma we then get

$$\|y(t)\|_\lambda \le (1+C_\gamma)(\|z(t-1)\|_\lambda + \|y^*(t)\|_\lambda) + C_w\|w(t)\|_\lambda + \xi(t). \tag{23}$$

Statement 3 of the lemma follows by using the last three relations and the fact that $\|z(t-1)\|_\lambda^2 \le \lambda^{-1}\|z(t)\|_\lambda$.

From the control law (9) we get

$$\begin{aligned}\|u(t)\|_\lambda \le \frac{1}{\epsilon_1}\Bigg[& f_\theta\left(\sum_{i=1}^{n_B}\lambda^{-i}\right)^{1/2}\|u(t-1)\|_\lambda \\ & + f_\theta\left(\sum_{i=1}^{n_A}\lambda^{-i}\right)^{1/2}\|y(t)\|_\lambda + \|y^*(t+1)\|_\lambda\Bigg] + \xi(t).\end{aligned} \tag{24}$$

Substituting this into (22) we get

$$\|\phi(t)\|_\lambda \le C_1\left(1+\frac{f_\theta C_1}{\epsilon_1}\right)(\|y(t)\|_\lambda + \|u(t-1)\|_\lambda) + \frac{C_1}{\epsilon_1}\|y^*(t+1)\|_\lambda + \xi(t).$$

It is now obvious from equations (23), (24), and (21) that $\|\phi(t)\|_\lambda$ can be bounded in terms of $\|z(t-1)\|_\lambda$ and we get

$$\begin{aligned}\|\phi(t)\|_\lambda \le\ & C_1\left(1+\frac{f_\theta C_1}{\epsilon_1}\right)(1+C_\gamma+C_u)\left(\|z(t-1)\|_\lambda + \frac{k_{y^*}}{(1-\lambda)^{1/2}}\right) \\ & + C_1\left(1+\frac{f_\theta C_1}{\epsilon_1}\right)\frac{C_w(1+C_A)k_w}{(1-\lambda)^{1/2}} + \frac{C_1}{\epsilon_1}\frac{k_{y^*}}{(1-\lambda)^{1/2}} + \xi(t).\end{aligned}$$

The result follows by the application of the definition of $r(t)$. □

By using equations (6), (11), (12), and (13), the measurement vector can be written in the form

$$\phi(t) = \phi^*(t) + \phi_w(t) + \phi_z(t) + \phi_\gamma(t) \tag{25}$$

where

$$\phi^*(t)' = \left(y^*(t), \dots, y^*(t-n_A+1), \frac{A(q^{-1})}{B(q^{-1})}y^*(t+1), \dots, \frac{A(q^{-1})}{B(q^{-1})}y^*(t-n_B+1)\right),$$

$$\phi_w(t)' = \left(w(t), \dots, w(t-n_A+1), \frac{A(q^{-1})-1}{B(q^{-1})}w(t+1), \dots, \frac{A(q^{-1})-1}{B(q^{-1})}w(t-n_B+1)\right),$$

$$\phi_z(t)' = \left(-z(t-1), \dots, -z(t-n_A), -\frac{A(q^{-1})}{B(q^{-1})}z(t), \dots, -\frac{A(q^{-1})}{B(q^{-1})}z(t-n_B)\right),$$

$$\phi_\gamma(t)' = \left(\gamma(t-1), \dots, \gamma(t-n_A), \frac{A(q^{-1})-1}{B(q^{-1})}\gamma(t), \dots, \frac{A(q^{-1})-1}{B(q^{-1})}\gamma(t-n_B)\right),$$

while $\phi^*(t)$, of course, is assumed to be persistently excited.

We are now ready to establish the following lemma.

LEMMA 4.2. *Let Assumptions* A1–A3 *hold. Then, for N_p sufficiently large on every subsequence $\{N_p\}$, where*

$$\|z(N_p)\|_\lambda^2 \le \Sigma_\gamma(\epsilon_1)^2 + \xi_5(N_p),$$

the following holds:

$$\lambda_{\min}\left\{\sum_{t=1}^{N_p} \lambda^{N_p-t}\phi(t)\phi(t)'\right\} \ge \delta_2^* - \rho_0 > 0, \qquad 0 < \rho_0 << \delta_2^*.$$

Here $\lambda_{\min}(\cdot)$ denotes the minimal eigenvalue of the corresponding matrix.

Proof. From the decomposition in equation (25) we obtain

$$\sum_{t=1}^{N_p} \lambda^{N_p-t}\left(\eta'\phi_z(t)\right)^2 \le n_1^2\Sigma_\gamma(\epsilon_1)^2 + \xi(t) \tag{26}$$

where η is any vector satisfying $\|\eta\| = 1$. Similarly, using statement 1 of Lemma 4.1 we obtain

$$\sum_{t=1}^{N_p} \lambda^{N_p-t}\left(\eta'\phi_\gamma(t)\right)^2 \le n_2^2\left(C_\gamma\Sigma_\gamma(\epsilon_1) + \frac{C_\gamma k_{y^*} + C_w k_w}{(1-\lambda)^{1/2}}\right)^2 + \xi(t) \tag{27}$$

and

$$\sum_{t=1}^{N_p} \lambda^{N_p-t}\left(\eta'\phi_w(t)\right)^2 \le n_2^2\frac{k_w^2}{1-\lambda} + \xi_7(t). \tag{28}$$

Relationships (26)–(28) yield

$$\sum_{t=1}^{N_p} \lambda^{N_p-t}\left(\eta'\phi_1(t)\right)^2 \le \left(n_1\Sigma_\gamma(\epsilon_1) + n_2\left(C_\gamma\Sigma_\gamma(\epsilon_1) + \frac{C_\gamma k_{y^*} + (1+C_w)k_w}{(1-\lambda)^{1/2}}\right)\right)^2 + \xi(t) \tag{29}$$

where $\phi_1(t)$ is the component of the regressor that is not directly excited by the reference, i.e.,

$$\phi_1(t) = \phi(t) - \phi^*(t) = \phi_z(t) + \phi_\gamma(t) + \phi_w(t).$$

It then follows that

$$(\eta'\phi(t))^2 \geq \frac{1}{2}(\eta'\phi^*(t))^2 - (\eta'\phi_1(t))^2.$$

From these inequalities and Assumption A3 we then get

$$\lambda_{\min}\left\{\sum_{t=1}^{N_p}\lambda^{N_p-t}\phi(t)\phi(t)'\right\} \geq \delta_2^* - \xi(t).$$

Since $\xi(t)$ decays exponentially fast, it follows that for sufficiently large N_p we have $\xi(t) \leq \rho_0$ and the lemma is proved. □

We now have the following critical result.

LEMMA 4.3. *Let Assumptions* A1–A3 *hold. Then on the subsequence* $\{N_p\}$ *where*

$$\|z(N_p)\|_\lambda^2 \leq \Sigma_\gamma(\epsilon_1)^2 + \xi_6(N_p) \tag{30}$$

we have

$$\|\theta(N_p+1) - \theta_0\|^2 \leq d_0(\epsilon_1) + \xi_7(N_p)$$

where

$$d_0(\epsilon_1) = \frac{C_\phi^2}{(\delta_2^* - \rho_0)^2}$$

$$\times\left[\Sigma_\gamma(\epsilon_1)(1+\mu(1+C_\gamma)+\epsilon_1 C_u) + \frac{k_{y^*}(C_\gamma + C_u\epsilon_1) + C_w k_w(1+\epsilon_1 C_A)}{(1+\lambda)^{1/2}}\right]^2 \tag{31}$$

with

$$C_\phi = C_{\phi 1}\Sigma_\gamma(\epsilon_1) + \frac{C_{\phi 1}k_{y^*} + C_{\phi 2}k_w}{(1-\lambda)^{1/2}}. \tag{32}$$

Proof. From equation (11) and statement 1 of Lemma 4.1 we derive

$$\|e(t+1)\|_\lambda \leq (1+C_\gamma)\|z(t)\|_\lambda + \frac{C_\gamma k_{y^*} + C_w k_w}{(1-\lambda)^{1/2}} + \xi(t). \tag{33}$$

Using equation (30) we obtain

$$\|e(N_p)\|_\lambda^2 \leq \left((1+C_\gamma)\Sigma_\gamma(\epsilon_1) + \frac{C_\gamma k_{y^*} + C_w k_w}{(1-\lambda)^{1/2}}\right)^2 + \xi(t). \tag{34}$$

Similarly from statement 3 of Lemma 4.1 we conclude that

$$\|\phi(N_p)\|_\lambda \leq C_\phi + \xi(t) \tag{35}$$

where C_ϕ was defined in equation (32). Note also, from equations (10) and (12), it follows that

$$\tilde{\theta}(t+1)'\phi(t) = z(t) + \epsilon(t)u(t), \qquad \tilde{\theta}(t) = \theta(t) - \theta_0.$$

This together with the estimation equation (5) then yields

$$\tilde{\theta}(t+1)'p(t)^{-1} = \lambda'\tilde{\theta}(t)p(t-1)^{-1} + z(t)\phi(t)' + \epsilon(t)u(t)\phi(t)' + \frac{\mu}{r(t)}\phi(t)p(t)^{-1}e(t+1) \tag{36}$$

where

$$p(t)^{-1} = \lambda p(t-1)^{-1} + \phi(t)\phi(t)', \qquad p(0)^{-1} = p_0 I, \qquad \text{with } p_0 > 0. \tag{37}$$

It follows from (36) and (37) that

$$\tilde{\theta}(N+1)' = \lambda^N\tilde{\theta}(1)p_0p(N) + \left(\sum_{t=1}^{N}\lambda^{N-t}z(t)\phi(t)'\right)p(N)$$

$$+ \left(\sum_{t=1}^{N}\lambda^{N-t}\epsilon(t)u(t)\phi(t)'\right)p(N) + \mu\left(\sum_{t=1}^{N}\lambda^{N-t}\frac{\phi(t)p(t)^{-1}}{r(t)}e(t+1)\right)p(N).$$

Since

$$\left\|\sum_{t=1}^{N}\lambda^{N-t}\frac{\phi(t)p(t)^{-1}}{r(t)}e(t+1)\right\| \le \|e(N)\|_\lambda\|\phi(N)\|_\lambda,$$

$$\left\|\sum_{t=1}^{N}\lambda^{N-t}\epsilon(t)u(t)\phi(t)'\right\| \le \epsilon_1\|u(N)\|_\lambda\|\phi(N)\|_\lambda, \qquad \text{and}$$

$$\left\|\sum_{t=1}^{N}\lambda^{N-t}z(t)\phi(t)'\right\| \le \|z(N)\|_\lambda\|\phi(N)\|_\lambda,$$

we obtain

$$\|\tilde{\theta}(N+1)\| \le \|p(N)\|(p_0\lambda^N\|\tilde{\theta}(1)\| + (\|z(N)\|_\lambda + \epsilon_1\|u(N)\|_\lambda + \|e(N)\|_\lambda)\|\phi(N)\|_\lambda). \tag{38}$$

From Lemma 4.2 we have for p sufficiently large on the subsequence $\{N_p\}$

$$\|p(N_p)\| \le \frac{1}{\delta_2^* - \rho_0} \tag{39}$$

where δ_2^* was defined under Assumption A3. The lemma then follows by using inequalities (30), (34), (35), (38), and (39) together. □

Comments. $\Sigma_\gamma(\epsilon_1)$ is small when C_γ and k_w are small. It follows that during the intervals where $\|z(t)\|_\lambda$ is of the order of $\Sigma_\gamma(\epsilon_1)$, excitation provided by the reference signal neutralizes the effect of the unmodeled dynamics and the external disturbances and prevents the parameter drift that otherwise causes instability of the estimator and the eventual destabilization of the loop. An exact statement of this property is given in Lemma 4.2.

The remaining question to answer is: "What happens when $\|z(t)\|_\lambda$ is not small?"

5. Main results and mathematical formalization of self-stabilization. Below we give the stability result for parameter estimation and direct adaptive control with persistent excitation.

THEOREM 5.1. *Let Assumptions* A1–A3 *introduced in* §3 *hold, and assume that the estimator is implemented with gain sequence* (7) *with* $\lambda_0 \leq \lambda < 1$. *Then there exist nonnegative constants* $\delta_1^*, C_\gamma, k_w$, *and* ϵ_1 *so that*

1.

$$\limsup_{t \to \infty} \|\theta(t) - \theta_0\|^2 \leq d_0(0) \quad \textit{and} \qquad \limsup_{t \to \infty} \|\theta_c(t) - \theta(t)\|^2 = 0$$

where $d_0(0)$ *is given by equation* (31) *with* $\epsilon_1 = 0$.

2.

$$\limsup_{t \to \infty} \sum_{j=1}^{t} \lambda^{t-j} |y(j+1) - y^*(j+1)|^2 \leq ((1 + C_\gamma)\Sigma_D + \Sigma_1)^2$$

where Σ_1 *is defined in Assumption* A3 *and*

$$\Sigma_D^2 = \max\left\{\Sigma_\gamma(0)^2, \left[\frac{\lambda}{(1 + d_0(0))C_\theta} + \frac{1}{\mu\rho_1(0)}\right] d_0(0)k_\theta\right\} \exp\left(\frac{(1 + d_0(0)C_\theta}{\lambda\mu\rho_1(0)}\right)$$

where $\Sigma_\gamma(0)^2$ *is given by Assumption* A2, $d_0(0)$ *is given by equation* (31), *and* $\rho_1(0)$ *is given by Assumption* A2 *with* $\epsilon_1 = 0$.

3.

$$\limsup_{t \to \infty} \|\phi(t)\|_\lambda \leq C_{\phi 1}\Sigma_D + \frac{C_{\phi 1}k_{y^*} + C_{\phi 2}k_w}{(1 - \lambda)^{1/2}}$$

where $C_{\phi i}, i = 1, 2$, *are defined under statement* 3 *in Lemma* 4.1.

Proof. We first determine a difference inequality that describes the behaviour of the parameter estimation error. From the estimation algorithm (5) we obtain

$$V(t+1) \leq V(t) + 2\frac{\mu}{r(t)}\tilde{\theta}(t)'\phi(t)e(t+1) + \frac{\mu^2}{r(t)}e(t+1)^2 \tag{40}$$

where $V(t) = \|\tilde{\theta}(t)\|^2$. From (11), the definition of the normalizing sequence, and the above we then derive

$$V(t+1) \leq V(t) - 2\mu(1 - \frac{\mu}{2})\frac{z(t)^2}{r(t)}$$

$$+ \frac{1}{r(t)}[2\mu(1-\mu)|z(t)\gamma(t)| + \mu^2\gamma(t)^2 + 2\mu|\epsilon(t)u(t)| \cdot |z(t) + \gamma(t)|]. \tag{41}$$

We now show how global stability of the adaptive algorithm can be demonstrated by considering the following comparison function [13]:

$$S(t+1) = V(t+1) + \frac{W(t+1)}{r(t)} \tag{42}$$

where

$$\begin{aligned} W(t+1) = \ & \mu \sum_{j=1}^{t} \lambda^{t-j} \Big[\Big(1 - \frac{\mu}{2} + (1-\mu)C_\gamma + \frac{\mu}{2}C_\gamma^2 + \epsilon_1 C_u(1 + C_\gamma)\Big) z(j)^2 \\ & - 2(1-\mu)|z(j)\gamma(j)| - \mu\gamma(j)^2 - 2|\epsilon(t)u(t)| \cdot |z(t) + \gamma(t)|\Big] \end{aligned} \tag{43}$$

It will become clear in the analysis that follows that stability of the sequence $z(t)$ follows trivially whenever $W(t) \leq 0$. However, during these intervals the function $S(t)$ may increase, thus giving rise to the bursting of the sequence $z(t)$. As a consequence of this, the function $W(t)$ becomes positive, forcing $S(t)$ to converge, thus stabilizing $z(t)$. This mechanism does not rely on the use of external excitation; instead it is required that the parameters remain bounded and that singularities are avoided in order to prevent finite escape and ensure bounded growth of signals. □

We now define subsequences τ_k and $\sigma_k, k \geq 1$, as follows:

$$1 = \tau_1 < \sigma_1 < \tau_2 < \sigma_2 < \cdots < \tau_k < \sigma_k < \tau_{k+1} < \cdots$$

so that

$$W(t+1) \leq 0 \text{ for } t \in Q_k \quad \text{and} \quad W(t+1) > 0 \text{ for } t \in T_k$$

with the intervals T_k and Q_k defined so that

$$Q_k = [\tau_k, \sigma_k), \qquad T_k = [\sigma_k, \tau_{k+1}), \qquad k \geq 1.$$

If $W(2) > 0$ set $\tau_1 = 0$ and $\sigma_1 = 1$ and Q_k is defined for $k \geq 2$.

The proof is handled by considering three possible cases:

Case 1. For all finite k, we have $\tau_k < \infty$ and $\sigma_k < \infty$.

Case 2. There exists a finite $k_0 \geq 0$, such that $\tau_{k_0} < \infty$ and $\sigma_{k_0} = +\infty$.

Case 3. There exists a finite $k_1 \geq 1$ so that $\sigma_{k_1} < \infty$ and $\tau_{k_1+1} = \infty$.

In the first case $W(t+1)$ changes sign infinitely often. The main idea in analyzing this case is the following. During the time intervals Q_k the function $W(t+1)$ is nonpositive and the stability of the adaptive system follows directly from the definition of $W(t+1)$. During the intervals when $W(t+1) > 0$ the function $S(t)$ is nonincreasing or strictly decreasing. It follows that such intervals only have a finite duration since $V(t)$ is nonnegative. This is essentially the argument used in [16] to show that the normalization $r(t)$ can be replaced by $\|\phi(t)\|^2$. A similar development can be used here; however, in order to simplify the analysis we treat the case where the normalizing signal $r(t)$ is defined by equation (7) here and (8) under the heading of Theorem 5.2.

The two last cases are trivial since the sign of $W(t+1)$ eventually does not change and are analyzed at the end of the proof of this theorem.

Analysis of Case 1. Let us first consider the interval Q_k. Since $W(t+1) \leq 0$ for $t \in Q_k$, that we have from equation (43) and statement 1 of Lemma 4.1 that

$$\left(1 - \frac{\mu}{2} + (1-\mu)C_\gamma + \frac{\mu}{2}C_\gamma^2 + \epsilon_1 C_u(1+C_\gamma)\right) \|z(t)\|_\lambda^2$$

$$\leq 2(1-\mu)\|z(t)\|_\lambda(C_\gamma\|z(t)\|_\lambda + h_1(t)) + \mu(C_\gamma\|z(t)\|_\lambda^2 + h_1(t))^2$$

$$+2\epsilon_1\|u(t)\|_\lambda((1+C_\gamma)\|z(t)\|_\lambda + h_1(t))$$

where we used the fact that $|\epsilon(t)| \leq \epsilon_1$. By further manipulation it follows that we can write

$$(\rho_1(\epsilon_1) + 2\epsilon_1 C_u(1+C_\gamma))\|z(t)\|_\lambda^2 \leq 2\mu(1-\mu+\mu C_\gamma)\|z(t)\|_\lambda h_1(t)$$

$$+ \mu h_1(t)^2 + 2\epsilon_1(1+C_\gamma)\|u(t)\|_\lambda\|z(t)\|_\lambda + 2\epsilon_1\|u(t)\|_\lambda h_1(t).$$

Substituting statement 2 of Lemma 4.1 into this we have

$$\rho_1(\epsilon_1)\|z(t)\|_\lambda^2 \leq 2\|z(t)\|_\lambda[(1-\mu+\mu C_\gamma+\epsilon_1 C_u)h_1(t)+\epsilon_1(1+C_\gamma)h_2(t)]$$
$$+\mu h_1(t)^2+2\epsilon_1 h_1(t)h_2(t)$$
$$\leq 4\max\Big\{\|z(t)\|_\lambda(1-\mu+\mu C_\gamma+\epsilon_1 C_u)h_1(t)+\epsilon_1(1+C_\gamma)h_2(t));$$
$$\frac{1}{2}\mu h_1(t)^2+\epsilon_1 h_1(t)h_2(t)\Big\}.$$

From this we finally get, by applying simple inequalities,

$$\|z(t)\|_\lambda^2 \leq \Sigma_\gamma(\epsilon_1)^2+\xi(t), \qquad t\in Q_k. \tag{44}$$

We can now apply Lemma 4.3 together with the definition of the intervals Q_k to conclude that

$$\|\tilde{\theta}(t)\|^2 \leq d_0(\epsilon_1)+\xi(t), \qquad t\in[\tau_k+1,\sigma_k]. \tag{45}$$

We now analyze the intervals $T_k, k \geq 1$, where $W(t+1)$ is positive. Using the definition of $W(t+1)$ in equation (43) we obtain from equations (41), (42), and (43) that

$$S(t+1)\leq S(t)-\mu\rho_1(\epsilon_1)\frac{z(t)^2}{r(t)}, \qquad t\in T_k,$$

where the definiton of $\rho_1(\epsilon_1)$ is given in Assumption A2. After summing from $t = \sigma_k+1$ to $N < \tau_{k+1}$ we get

$$S(N+1)\leq S(\sigma_k+1)-\mu\rho_1(\epsilon_1)\sum_{t=\sigma_k+1}^{N}\frac{z(t)^2}{r(t)}. \tag{46}$$

From equation (43) it follows that

$$W(\sigma_k+1)=\mu\left(1-\frac{\mu}{2}+(1-\mu)C_\gamma+\frac{\mu}{2}C_\gamma^2+\epsilon_1 C_u(1+C_\gamma)\right)z(\sigma_k)^2$$

$$-2\mu(1-\mu)|z(\sigma_k)\gamma(\sigma_k)|-\mu^2\gamma(\sigma_k)^2+2|\epsilon(\sigma_k)u(\sigma_k)|\cdot|z(\sigma_k)+\gamma(\sigma_k)|+\lambda W(\sigma_k).$$

Since $W(\sigma_k)\leq 0$, we get, using equations (41) and (42), that

$$S(\sigma_k+1)\leq V(\sigma_k)-\mu\rho_1(\epsilon_1)\frac{z(\sigma_k)^2}{r(\sigma_k)}.$$

Using this in equation (46) we get

$$S(N+1)\leq V(\sigma_k)-\mu\rho_1(\epsilon_1)\sum_{t=\sigma_k}^{N}\frac{z(t)^2}{r(t)}, \qquad N\in T_k. \tag{47}$$

From the definition of $S(t)$ given by equation (42) and inequality (45) it then follows that

$$\|\tilde{\theta}(t+1)\|^2\leq d_0(\epsilon_1)+\xi(\sigma_k), t\in[\sigma_k,\tau_{k+1}). \tag{48}$$

Inequalities (45) and (48) imply that there exists a finite k_2 such that

$$\|\tilde{\theta}(t)\|^2 \leq d_0(\epsilon_1) + \rho_2 \tag{49}$$

for $t \in [\tau_k + 1, \tau_{k+1}], k \geq k_2$, or for all $t \geq \tau_{k_2} + 1$. Here ρ_2 is an arbitrarily small constant. From this we may conclude that

$$|\hat{b}_0(t)| \geq |b_0| - d_0(\epsilon_1)^{1/2} - \rho_2^{1/2} > 0 \quad \text{for } t \geq \tau_{k_2} + 1. \tag{50}$$

From equation (31) it follows that $d_0(\epsilon_1)$ is small if C_γ, k_w, and ϵ_1 are small relative to δ_1^*. This in turn implies that the estimate of b_0 stays close to its true value after a transient period that can be no longer than $\tau_{k_2} + 1$. After this we get $\epsilon(t) = 0$ in equation (10).

Equation (43) can now be written as

$$W(t+1) = \mu \sum_{j=1}^{t} \lambda^{t-j} \left[\left(1 - \frac{\mu}{2} + (1-\mu)C_\gamma + \frac{\mu}{2}C_\gamma^2\right) z(j)^2 \right.$$
$$\left. -2(1-\mu)|z(j)\gamma(j)| - \mu\gamma(j)^2\right] + \lambda^{t-\tau_{k_2}}\eta(\tau_{k_2}) \tag{51}$$

where

$$\eta(\tau_{k_2}) = \mu \sum_{j=1}^{\tau_{k_2}} \lambda^{\tau_{k_2}-j}[\epsilon_1 C_u(1+C_\gamma)z(j)^2 - 2|\epsilon(j)u(j)| \cdot |z(j) + \gamma(j)|].$$

Using inequality (44), which we write as

$$\|z(\tau_{k_2})\|_\lambda^2 \leq \Sigma_\gamma(\epsilon_1)^2 + \xi(\tau_{k_2}),$$

together with statements 1 and 2 of Lemma 4.1, this gives

$$\eta(\tau_{k_2})^2 \leq C_2 < \infty. \tag{52}$$

Similarly from equation (51) we get

$$\|z(t)\|_\lambda^2 \leq \Sigma_\gamma(0)^2 + \xi(t) \tag{53}$$

for all $t \in Q_k$, $k > k_2, k_2 < \infty$, where we used the fact that $\epsilon(t) = 0$ for all $t > \tau_{k_2}$. Appplying Lemma 4.3 one more time with (53) we conclude that

$$\|\tilde{\theta}(t+1)\|^2 \leq d_0(0) + \xi(t), \quad t \in Q_k, \quad k > k_2. \tag{54}$$

Following the same line of reasoning as that given in equations (46)–(47) and using the fact that $\epsilon(t) =$ 0 for all $t \geq \tau_{k_2} + 1$ we get immediately from equations (41), (42), and (51) that

$$S(N+1) \leq V(\sigma_k) - \mu\rho_1(0) \sum_{t=\sigma_k}^{N} \frac{z(t)^2}{r(t)}, \qquad N \in T_k, \quad k \geq k_2, \tag{55}$$

with $\rho_1(0)$ as defined under point 3 of Assumption A2. From (42), (54), and (55) it follows that

$$\limsup_{k\to\infty} \sup_{t\in T_k} \|\tilde{\theta}(t+1)\|^2 \leq d_0(0). \tag{56}$$

Statement 1 of Theorem 5.1 follows for the Case 1 scenario by combining (54) and (56). It now remains to show boundedness of all signals. Since the parameters are bounded, a number of different techniques can be applied. We follow the method developed by [13]. First, from (55) and (56) we have

$$\mu\rho_1(0)\sum_{t=\sigma_k}^{N}\frac{z(t)^2}{\tilde{r}(t)} \leq 1, \qquad N \in T_k, \quad k \geq k_2 \tag{57}$$

where

$$\tilde{r}(t) = (d_0(0) + \xi(\sigma_k))r(t). \tag{58}$$

From (49) we conclude that statement 4 of Lemma 4.1 can be applied for all $t \geq \tau_{k_2}+1$ with

$$f_\theta = \|\theta_0\| + d_0(0)^{1/2} + \rho_2^{1/2}.$$

Substituting the bound for $r(t)$ given by statement 4 of Lemma 4.1 into equation (58) then gives

$$\tilde{r}(t) \leq \max\{(1+d_0(0))C_\theta\|z(t-1)\|_\lambda^2, g(t)\}, \qquad t \in T_k, \quad k \geq k_2 \tag{59}$$

where

$$g(t) = (d_0(0) + \xi(\sigma_k))(k_\theta + \xi(t)). \tag{60}$$

Next we show that relationships (57) and (59) can be used together to establish stability during the intervals T_k, $k \geq k_2$. Specifically, on the one hand it follows from equation (57) that we can relate the magnitude of $z(t)$ for $t \in T_k$ to $\tilde{r}(t)$. On the other hand, from inequality (59) it follows that $\tilde{r}(t)$ is of the order of $g(t)$ or $(1+d_0(0))C_\theta\|z(t-1)\|_\lambda^2$.

We now introduce a further partitioning of the intervals T_k. Let $p_{ik} \in T_k$ and $l_{ik} \in T_k$ be defined so that

$$p_{0k} < l_{1k} < p_{1k} < \cdots < l_{ik} < p_{ik} < l_{(i+1)k} < \cdots,$$

so that for $\tilde{r}(t)$ defined by (59) the following inequality holds:

$$\tilde{r}(t) \leq (1+d_0(0))C_\theta\|z(t-1)\|_\lambda^2 \quad \text{for } t \in L_{ik}, i \geq 1, k \geq k_2, \tag{61}$$

and

$$\tilde{r}(t) \leq g(t) \quad \text{for } t \in D_{ik}, i \geq 1, k \geq k_2 \tag{62}$$

where the intervals L_{ik} and D_{ik} are defined so that

$$L_{ik} = [p_{(i-1)k}, l_{ik}) \text{ and } D_{ik} = [l_{ik}, p_{ik}).$$

We have $l_{ik} < \tau_{k+1}$ and $p_{ik} < \tau_{k+1}$ for $i \geq 1$.

$$\text{If } (1+d_0(0))C_\theta\|z(\sigma_k-1)\|_\lambda^2 \begin{cases} \geq g(\sigma_k), & \text{then } p_{0k} = \sigma_k; \\ < g(\sigma_k), & \text{then } p_{0k} = 0\,, l_{1k} = \sigma_k \\ & \quad \text{with intervals } L_{ik} \text{ defined for } i \geq 2. \end{cases}$$

If $(1+d_0(0))C_\theta\|z(t-1)\|_\lambda^2 \begin{cases} < g(t) \text{ for all } t \in T_k \text{ define } p_{0k} = 0, l_{1k} = \sigma_k, \text{ and } p_{1k} = \tau_{k+1}; \\ \geq g(t) \text{ for all } t \in T_k \text{ set } p_{0k} = \sigma_k \text{ and } l_{1k} = \tau_{k+1}. \end{cases}$

From (59) and (62) we now have

$$\|z(t-1)\|_\lambda^2 \leq \frac{g(t)}{(1+d_0(0))C_\theta} \quad \text{for } t \in D_{ik}, i \geq 1, k \geq k_2 \tag{63}$$

and relations (57), (59), and (62) imply that for $t \in D_{ik}$

$$z(t)^2 \leq \frac{\tilde{r}(t)}{\mu\rho_1(0)} \leq \frac{g(t)}{\mu\rho_1(0)}. \tag{64}$$

Since $\|z(t)\|_\lambda^2 = z(t)^2 + \lambda\|z(t-1)\|_\lambda^2$, relations (60), (63), and (64) yield

$$\|z(t)\|_\lambda^2 \leq \left(\frac{\lambda}{(1+d_0(0))C_\theta} + \frac{1}{\mu\rho_1(0)}\right) d_0(0)k_\theta + \xi(\sigma_k), \qquad t \in D_{ik}. \tag{65}$$

We now consider the intervals $L_{ik} \in T_k$ for $i \geq 1, k \geq k_2$. From (57) and (61) we obtain

$$R_{ik} = \sum_{t=p_{(i-1)k}}^{N} \frac{z(t)^2}{\|z(t-1)\|_\lambda^2} \leq \frac{(1+d_0(0))}{\mu\rho_1(0)} C_\theta, \qquad N \in L_{ik}. \tag{66}$$

From this it follows that we have

$$R_{ik} = \sum_{t=p_{(i-1)k}}^{N} \frac{\|z(t)\|_\lambda^2 - \lambda\|z(t-1)\|_\lambda^2}{\|z(t-1)\|_\lambda^2} \tag{67}$$

$$= \sum_{t=p_{(i-1)k}}^{N} \frac{\lambda}{\lambda\|z(t-1)\|_\lambda^2} \int_{\lambda\|z(t-1)\|_\lambda^2}^{\|z(t)\|_\lambda^2} dx,$$

which gives

$$R_{ik} \geq \lambda \sum_{t=p_{(i-1)k}}^{N} \int_{\lambda\|z(t-1)\|_\lambda^2}^{\|z(t)\|_\lambda^2} \frac{dx}{x} = \lambda \sum_{t=p_{(i-1)k}}^{N} (\log(\|z(t)\|_\lambda^2) - \log(\lambda\|z(t-1)\|_\lambda^2))$$

$$= \lambda \log \frac{\|z(N)\|_\lambda^2}{\|z(p_{(i-1)k} - 1)\|_\lambda^2} + \lambda(N - p_{(i-1)k}) \log \frac{1}{\lambda}. \tag{68}$$

From equations (66) and (68) we get

$$\|z(N)\|_\lambda^2 \leq \exp\left(\frac{1+d_0(0)}{\lambda\mu\rho_1(0)} C_\theta\right) \|z(p_{(i-1)k} - 1)\|_\lambda^2, \qquad N \in L_{ik}. \tag{69}$$

Since $p_{(i-1)k} - 1 \in D_{(i-1)k}$, we obtain using (65) for $t \in L_{ik}, i \geq 2$,

$$\|z(t)\|_\lambda^2 \leq d_0(0)k_\theta \left(\frac{\lambda}{(1+d_0(0))C_\theta} + \frac{1}{\mu\rho_1(0)}\right) \exp\left(\frac{1+d_0(0)}{\lambda\mu\rho_1(0)} C_\theta\right) + \xi(\sigma_k). \tag{70}$$

We now evaluate $\|z(t)\|_\lambda^2$ for the intervals $L_{1k}, k \geq k_2$. If $p_{0k} = \sigma_k$, we get from inequalities (53) and (69) that

$$\|z(t)\|_\lambda^2 \leq \Sigma_\gamma(0)^2 \exp\left(\frac{1+d_0(0)}{\lambda\mu\rho_1(0)}C_\theta\right) + \xi(\sigma_k). \tag{71}$$

In the case $p_{0k} = 0$ and $l_{1k} = \sigma_k$, the intervals L_{ik} are defined for $i \geq 2$. In equation (68), the case $\|z(N)\|_\lambda^2 \leq \|z(p_{(i-1)k} - 1)\|_\lambda^2$ is trivial and is covered by inequality (69). Therefore, from (65), (70), and (71) it follows that

$$\limsup_{k\to\infty} \sup_{t\in T_k} \|z(t)\|_\lambda^2 \leq \Sigma_D^2,$$

which together with (53) gives

$$\limsup_{t\to\infty} \|z(t)\|_\lambda^2 \leq \Sigma_D^2 \tag{72}$$

with Σ_D as given in Theorem 5.1. Statement 2 follows from the application of inequalities (33) and (72).

Analysis of Case 2. From the definition of this case it follows that relationship (44) is valid for all $t \geq \tau_{k_0}$. From this point on we apply Lemma 4.3, which demonstrates that (49) is valid for all $t \geq \tau_{k_0}$. Using the same technique as in the demonstration of (50), we conclude that there exists finite $t_1 \geq \tau_{k_0}$ such that this relation holds true for all $t \geq t_1$. This implies that $\epsilon(t) = 0$ in equation (10) for all $t \geq t_1$. $W(t+1)$ then is as in equation (51) with $\eta(\tau_{k_2})$ replaced by $\eta(t_1)$. Since $W(t+1) < 0$ for all $t \geq t_1$, we obtain as in inequality (53) that

$$\|z(t)\|_\lambda^2 \leq \Sigma_\gamma(0)^2 + \xi(t), \qquad t \geq t_1. \tag{73}$$

By application of Lemma 4.3 we then must conclude

$$\|\tilde{\theta}(t)\| \leq d_0(0) + \xi(t), \qquad t \geq t_1,$$

and we have proven statement 1 of the theorem for the Case 2 scenario.

Analysis of Case 3. There exists $k_1 < \infty$ so that $\sigma_{k_1} < \infty$ and $\tau_{k_1+1} = \infty$, and we conclude that $W(t+1) > 0$ for all $t \geq \sigma_{k_1}$ and consequently from (47) we have

$$\mu\rho_1(\epsilon_1) \sum_{t=\sigma_{k_1}}^{\infty} \frac{z(t)^2}{r(t)} < \infty.$$

This follows since $V(t) \geq 0$. Using statement 3 of Lemma 4.1 it is clear that there exists a finite $t_2 \geq \sigma_{k_1}$ such that

$$\|z(t)\|_\lambda^2 = o(1). \tag{74}$$

When we apply Lemma 4.3 together with this result we notice that

$$\|\tilde{\theta}(t)\| \leq d_0(\epsilon_1) + \xi(t), \qquad t \geq t_1 \tag{75}$$

for all $t \geq t_2$. Similarly as in the previous analysis we conclude that there exists a finite $t_3 \geq t_2$ such that $\epsilon(t) = 0$ and $|\hat{b}_0(t)| \geq \epsilon_1$ for all $t \geq t_3$. Setting $\epsilon_1 = 0$ in (75) gives what is needed to conclude that statement 1 is in fact correct.

Statements 2 and 3 of the theorem follow by application of equation (33) and (72) or (73) or (74), while statement 3 follows by application of Lemma 4.1, and Theorem 5.1 is established.

We now give the main results for normalization sequence (8).

THEOREM 5.2. *Let Assumptions* A1–A3 *hold, and assume that the algorithm gain sequence is updated using equation* (8). *Then there exists* $\delta_1^*, C_\gamma, k_w$, *and* ϵ_1 *so that*

1.

$$\limsup_{t\to\infty} \|\theta(t)-\theta_0\|^2 \le d_0(0) \quad \textit{and} \quad \limsup_{t\to\infty} \|\theta_c(t)-\theta(t)\|^2 = 0$$

where $d_0(0)$ *is given by equation* (31) *with* $\epsilon_1 = 0$.

2.

$$\limsup_{t\to\infty} |y(t+1)-y^*(t+1)|^2 \le \left(\left(1+\frac{C_\gamma}{(1-\lambda)^{1/2}}\right)\Sigma_{D1} + \frac{C_\gamma k_{y^*}+C_w k_w}{(1-\lambda)^{1/2}}\right)^2$$

where

$$\Sigma_{D1} = \max\{\Sigma_\gamma(0)^2; d_0(0)C_r\}$$

and C_r *is a constant that decreases when* C_γ *and* k_w *decrease, while* $\Sigma(0)$ *is given by equation* (17) *with* $\epsilon_1 = 0$.

Proof. The proof follows the same line of reasoning as for Theorem 5.1. Let us first consider the case when $\tau_\kappa < \infty$ and $\sigma_\kappa < \infty$ for all finite k. Then it is not difficult to see that starting from equation (33) up to equation (56), the analysis is exactly the same and holds for the case when $r(t)$ is defined by equation (8). This is because all that is required up to this point is that $r(t) \ge \lambda r(t-1)$. Therefore statement 1 of the theorem is valid.

Let us prove statement 2. First we show that $\max_{1\le\tau\le t}|z(\tau)|$ is bounded. From inequalities (54) and (55) it follows that for sufficiently large k

$$\frac{z(t)^2}{r(t)} \le d_0(0) + \beta_1 \text{ for all } t \in T_k \tag{76}$$

where $0 < \beta_1 << 1$ is a small number. On the other hand, from (8) and statement 3 of Lemma 4.1 we have

$$r(t)^{1/2} \le \frac{C_{\phi 1}}{(1-\lambda)^{1/2}} \max_{1\le\tau\le t}|z(\tau)| + C_{\phi 3} \tag{77}$$

where

$$C_{\phi 3} = r_0^{1/2} + \beta_2 + \frac{C_{\phi 1}k_{y^*} + C_{\phi 2}k_w}{(1-\lambda)^{1/2}}, \qquad 0 < \beta_2 < \infty.$$

Substituting (77) into equation (76) together with (53) gives

$$|z(t)| \le \begin{cases} (d_0(0)^{1/2} + \beta_1) r(t)^{1/2} \text{ for all } t \in T_k, \\ \Sigma_\gamma(0) + \xi(t) \text{ for all } t \in Q_k \end{cases} \tag{78}$$

and hence

$$|z(t)| \le (d_0(0)^{1/2}+\beta_1)\frac{C_{\phi 1}}{(1-\lambda)^{1/2}} \max_{1\le\tau\le t}|z(\tau)| + (d_0(0)^{1/2}+\beta_1)C_{\phi 3} + \Sigma_\gamma(0) + \xi(t) \tag{79}$$

Since β_1 can be made arbitrarily small, it follows from (31) with $\epsilon_1 = 0$ and the definition of $C_{\phi 1}$ in Lemma 4.1 that there exists numbers δ_1^*, C_γ, and k_w such that

$$(d_0(0)^{1/2} + \beta_1)\frac{C_{\phi 1}}{(1-\lambda)^{1/2}} < 1. \tag{80}$$

That is, if the level of excitation (δ_1^*) is sufficiently large relative to C_γ and k_w, then the parameter error $d_0(0)$ will be small and inequality (80) holds. From inequality (79) it then follows that

$$\max_{1\le\tau\le t} |z(\tau)| \le \frac{(d_0(0)^{1/2} + \beta_1)C_{\phi 3} + \Sigma_\gamma(0) + \xi(t)}{1 - (d_0(0)^{1/2} + \beta_1)C_{\phi 1}/(1-\lambda)^{1/2}}. \tag{81}$$

By using equation (77) we then have

$$r(t) \le C_r < \infty. \tag{82}$$

Note that when C_γ and k_w decrease then $d_0(0), C_{\phi 3}, \Sigma_\gamma(0)$, and $C_{\phi 1}$ decrease as well with the consequence that C_r decreases.

From (55) and (82) we obtain

$$\limsup_{k\to\infty} \sup_{N\in T_k} \sum_{t=\sigma_k}^{N} z(t)^2 \le d_0(0)C_r. \tag{83}$$

Combining (53) with inequality (83) gives

$$\limsup_{t\to\infty} z(t)^2 \le \Sigma_{D1}^2 \tag{84}$$

where Σ_{D1} is defined in Theorem 5.2. From (11), statement 1 of Lemma 4.1, and inequality (84) it follows that

$$\limsup_{t\to\infty} e(t+1)^2 \le \left(\left(1 + \frac{C_\gamma}{(1-\lambda)^{1/2}}\right)\Sigma_{D1} + \frac{C_\gamma k_{y^*} + C_w k_w}{(1-\lambda)^{1/2}}\right)^2.$$

Thus the theorem is proved for the case when $\sigma_k < \infty$ and $\tau_k < \infty$ for all finite k. The cases when there exists a finite κ_0 so that $\tau_{k_0} < \infty$ and $\sigma_{k_0} = \infty$ or there exists k_1 so that $\tau_{k_1+1} = \infty$ and $\sigma_{k_1} < \infty$ are trivial. □

Remarks.

1. The advantage of the normalization signal given by equation (8) over that given by equation (7) is that it does not require a knowledge of the characteristic time constant λ_0.
2. From Assumption A3 it follows that

$$\lim_{C_\gamma, k_w\to 0} \Sigma_1 = 0, \qquad \lim_{C_\gamma, k_w\to 0} \Sigma_\gamma(0) = 0, \qquad \lim_{C_\gamma, k_w\to 0} \delta_2^* = \delta_1^*/2.$$

From the definition of $d_0(\epsilon_1)$ in equation (31), the definition of Σ_D in statement 2 of Theorem 5.1, and the analysis given above it then follows that

$$\lim_{C_\gamma, k_w\to 0} d_0(0) = 0, \qquad \lim_{C_\gamma, k_w\to 0} \Sigma_D = 0.$$

Consequently from statements 1, 2 of Theorem 5.1

$$\lim_{C_\gamma, k_w \to 0} \sum_{j=1}^{t} \lambda^{t-j} |y(j+1) - y^*(j+1)|^2 = 0$$

and

$$\lim_{C_\gamma, k_w \to \infty} \|\theta(t) - \theta_0\| = 0.$$

Similar results can be derived from the expressions in Theorem 5.2. This implies that the tracking and parameter estimation errors are continuous with respect to the unmodeled dynamics and the disturbances in the sense that when the unmodeled dynamics and the external perturbations tend to zero, tracking and parameter estimation errors also tend to zero. Without the application of external excitation, the uniform convergence of the parameter error cannot be expected.

3. When persistent excitation is applied, the self-stabilization works in the following manner. Whenever the estimator produces parameters that cause the adaptive loop to become unstable, the controller stabilizes itself by producing excitation, which results in parameter tuning. This tuning is the result of "self-stabilization" and takes place even when the external signals are not excited. In fact, the analysis shows the adaptive control algorithm passes through two phases characterized by the intervals Q_k and T_k, defined in the proof of Theorem 5.1. In the intervals Q_k, the function $W(t+1)$ is nonpositive, which implies stability of the input and output signals. During these intervals of time the parameter estimates may drift unless external excitation is supplied. Furthermore, the analysis shows that the level of excitation needs to be higher, in some sense, than the intensities of the disturbances and the unmodeled dynamics. In this respect our result is equivalent to previous results that have been obtained using averaging analysis.
4. From a practical point of view it makes sense to turn the excitation and estimation algorithm off after the parameters have settled and acceptable performance has been achieved.
5. There exists a finite level of excitation and a finite time τ_{k_2} so that the estimate of the high frequency gain, $\hat{b}_0(t)$, stays away from zero, i.e., $|\hat{b}_0(t)| \geq \epsilon_1$ for all $t \geq \tau_{k_2}$. This implies that the modification introduced to avoid division with small numbers is only used during a transient period. This observation can be used to motivate the use of projection and persistent excitation to solve singularity problems associated with the solution of Bezout-like equations as well.
6. The results developed in this paper can be extended to indirect control and estimation. In particular, the obtained results lead in a natural way to the definiton of a robust controller based on the use of H_∞-like design techniques.
7. The burstings can only have finite duration. An estimate of the longest period of time a burst is tolerated by the adaptive control algorithm can be calculated from equation (68).

6. Conclusions. In this paper we show that excitation with fixed and finite energy can be used to stabilize the estimator in a direct adaptive control algorithm. We show that the parameter estimates and the input/output signals remain bounded when the level of excitation is sufficiently high relative to the magnitude of the external perturbations and the intensity of the unstructured unmodeled dynamics. We also show that the parameter estimates "converge" close to their optimal values. In other words, we show that it is possible to perform identification of open-loop unstable systems and that we can bound the parameter estimation error. The results apply to systems that have a stably invertible nominal model. It is quite straightforward to generalize the results to a broader class of systems and to apply more complex control laws. A few practical problems remain to be solved. First, it is not clear how to develop algorithms to monitor performance and turn the estimation algorithm on and off. Second, while we have been able to develop guidelines for choosing excitation level through the definition of δ_2^*, these are not so easy to implement because of the fact that bounds on certain system H_∞ norms have to be known in advance for us to be able to implement the approach.

REFERENCES

[1] P. DE LARMINAT AND H. F. REYNAUD, *A robust solution to the admissibility problem in indirect adaptive control*, Internat. J. Adaptive Control. Signal Proc., 2 (1988), pp. 95–110.

[2] B. D. O. ANDERSON, R. R. BITMEAD, C. R. JOHNSON, P. V. KOKOTOVIC, R. L. KOSUT, I. M. Y. MAREELS, L. PRALY, AND B. D. RIEDLE, *Stability of Adaptive Systems: Passivity and Averaging Analysis*, MIT Press, Cambridge, MA, 1986.

[3] B. EGARD, *Stability of adaptive controllers*, Springer-Verlag, New York, 1979.

[4] G. C. GOODWIN, P. J. RAMADAGE, AND P. E. CAINES, *Discrete time multivariable adaptive control*, IEEE Trans. Automat. Control, AC-35 (1990), pp. 449–456.

[5] R. LOZANO-LEAL, J. COLLADO, AND S. MONDIE, *Model reference robust adaptive control without a priori knowledge of the high frequency gain*, IEEE Trans. Aut. Control, AC-35 (1990), pp. 71–78.

[6] P. IOANNOU AND J. SUN, *Theory and design of robust direct and indirect adaptive control systems*, Internat. J. Control, 47 (1988), pp. 775–813.

[7] R. H. MIDDLETON, G. C. GOODWIN, D.·J. HILL, AND D. Q. MAYNE, *Design issues in adaptive control*, IEEE Trans. Automat. Control, 33 (1988), pp. 50–58.

[8] A. S. MORSE, D. Q. MAYNE, AND G. C. GOODWIN, *Applications of hysteresis switching in parameter adaptive control*, IEEE Trans. Automat. Contr., 37 (1992), pp. 1343–1354.

[9] S. J. NAIK, P. R. KUMAR, AND B. E. YDSTIE, *Robust Continuous Time Adaptive Control with Parameter Projection*, IEEE Trans. Automat. Control, T-AC 37, January 1992, pp. 182–197.

[10] K. S. NARENDRA AND A. M. ANNASWAMY, *Stable Adaptive Systems*, Prentice-Hall, Englewood Cliffs, NJ, 1989.

[11] L. PRALY, *Robustness of model reference adaptive control*, Proc. 2nd Yale Workshop on Adaptive Systems, New Haven, CT, 1983.

[12] ———, *Topological orbital equivalence with asymptotic phase for a two time-scales discrete time system*, Math. Control Systems Signals, 3 (1990), pp. 225–253.

[13] M. S. RADENKOVIC AND A. N. MICHEL, *Robust adaptive systems and self-stabilization*, IEEE Trans. Automat. Control, 37 (1992), pp. 1355–1369.

[14] B. D. RIEDLE AND P. V. KOKOTOVIC, *Integral manifolds of slow adaptation*, IEEE Trans. Automat. Control, 31 (1986), pp. 316–324.

[15] C. E. ROHRS, L. VALEVANI, M. ATHANS, AND G. STEIN, *Analytical verification of undesirable properties of direct model reference adaptive control algorithm*, Proc. 20th IEEE Conf. Decision and Control, 2, San Diego, CA, 1981, pp. 1272–1284.

[16] B. E. YDSTIE, *Transient performance and robustness of direct adaptive control*, IEEE Trans. Automat. Contr., 37 (1992), pp. 1091–1105.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 33, No. 4, pp. 1247–1255, July 1995

© 1995 Society for Industrial and Applied Mathematics

015

IDENTIFICATION OF $q(x)$ IN $u_t = \Delta u - qu$ FROM BOUNDARY OBSERVATIONS*

SERGEI AVDONIN[†] AND THOMAS I. SEIDMAN[‡]

Abstract. We consider the problem of recovering the coefficient $q(x)$ in the equation $u_t = \Delta u - qu$ from boundary observations. Uniqueness of q based on knowledge of the Neumann ↦ Dirichlet response operator is shown as an implication of (known) corresponding results concerning the inverse problem for the corresponding hyperbolic equation $w_{tt} = \Delta w - qw$. This is then reduced to use of the response to a single input with some consideration of computational approximation.

Key words. identification, parabolic, partial differential equation, uniqueness, approximation

AMS subject classifications. 35R30, 35K99, 35C99

1. Introduction.

We consider the problem of identifying the (unknown) coefficient $q = q(x)$ in the parabolic partial differential equation

$$u_t = \Delta u - qu \qquad \text{on } Q := (0,T) \times \Omega, \tag{1.1}$$

assuming input/output access only at the boundary $\Sigma = \Sigma_T := (0,T) \times \partial\Omega$. More precisely, we assume that we can specify the Neumann data for (1.1) with trivial initial data

$$\frac{\partial u}{\partial \nu} = f \qquad \text{on } \Sigma_T \qquad u\Big|_{t=0} = 0 \qquad \text{on } \Omega \tag{1.2}$$

and then observe the corresponding Dirichlet data,

$$g := u\Big|_{\Sigma}. \tag{1.3}$$

Formally, then, we have a linear input/output map (Neumann ↦ Dirichlet response operator)

$$\mathbf{R}_1 = \mathbf{R}_1(T;q) : f \mapsto g, \tag{1.4}$$

defined through (1.1), (1.2), and then the observation (1.3). Our principal result is that $\mathbf{R}_1$ (for any $T > 0$) uniquely determines the coefficient function $q(\cdot)$ appearing in (1.1), i.e., that

$$q \mapsto \mathbf{R}_1(T,q) \ \text{ is injective on } \boldsymbol{\mathcal{A}} \tag{1.5}$$

when considered for q in some suitable set $\boldsymbol{\mathcal{A}}$ of admissible functions.

We note that results like (1.5) are already available for the inverse problem for the corresponding hyperbolic equation

$$w_{tt} = \Delta w - qw \qquad \text{on } (0,\bar{T}) \times \Omega. \tag{1.6}$$

* Received by the editors May 28, 1993; accepted for publication (in revised form) February 22, 1994. This research was partially supported by the National Academy of Sciences under the National Academy of Sciences/National Research Council Project Development Program.

[†] Department of Applied Mathematics and Control, St. Petersburg State University, St. Petersburg, 198904, Russia (avdonin@apmath.lgu.spb.su).

[‡] Department of Mathematics and Statistics, University of Maryland–Baltimore County, Baltimore, Maryland 21228 (seidman@math.umbc.edu).

Our approach — exploiting the deep connection between (1.1) and (1.6) via transforms with respect to t — is stimulated by D. Russell's argument ([14]; see also [15]) showing how to deduce exact null controllability of the heat equation for a bounded region $\Omega \subset \mathbb{R}^N$ from a corresponding wave equation result. We may restate our description above to say that our primary result is the *implication*, under fairly general hypotheses, of (1.5) from

$$q \mapsto \mathbf{R}_2(\bar{T}, q) \text{ is injective on } \boldsymbol{\mathcal{A}}, \tag{1.7}$$

where $\mathbf{R}_2$ is the corresponding Neumann ↦ Dirichlet response operator for (1.6). This argument will be given in §2.

Parenthetically, we note that a quite different argument could alternatively obtain parabolic identifiability from corresponding results to the extent that these would be available for the elliptic rather than the hyperbolic case, i.e., deriving (1.5) from (cf., e.g., [11])

$$q \mapsto \mathbf{R}_0(q) \text{ is injective on } \boldsymbol{\mathcal{A}}, \tag{1.8}$$

where $\mathbf{R}_0(q) : f \mapsto g$ is the Neumann ↦ Dirichlet operator for the elliptic equation

$$-\boldsymbol{\Delta} v + qv = 0 \qquad \text{on } \Omega, \quad \frac{\partial v}{\partial \nu} = f(x) \qquad \left[g := v\Big|_{\partial\Omega} \right]. \tag{1.9}$$

To see this, one applies (1.1) and (1.2) to f constant in t which gives u analytic in t and, assuming $q > 0$, convergent to the steady-state solution v of (1.9) as $t \to \infty$. This analyticity implies that $\mathbf{R}_1(T,q)f$ uniquely determines $g(\cdot) := u\big|_{\partial\Omega}$ not only on $[0,T]$ but for all $t > 0$; compare the approach of [16]. The limit as $t \to \infty$ is then also uniquely determined so, for any such $f = f(x)$ and any $T > 0$, one sees that $\mathbf{R}_1(T,q)f = \mathbf{R}_1(T,\hat{q})$ implies $\mathbf{R}_0(q)f = \mathbf{R}_0(\hat{q})$; compare [10].

Whereas it seems that the entire response operator $\mathbf{R}_2$ may be needed for identifiability for (1.6), we will show in §3 that a single experiment, using a suitably chosen input f_* and observing the associated output

$$g_* = \boldsymbol{\Gamma}(q) := \mathbf{R}_1(T;q)f_*, \tag{1.10}$$

suffices to identify q in (1.1), i.e., that f_* can be chosen so that $\boldsymbol{\Gamma}$ is injective on $\boldsymbol{\mathcal{A}}$. Section 3 will also include some additional remarks on possible computational implementation.

2. Principal results.

We assume throughout that Ω is a bounded region in $\mathbb{R}^n$ with sufficiently smooth boundary $\partial\Omega$ for the relevant trace theory to apply for the operators $\mathbf{B} : u \mapsto u\big|_{\partial\Omega}$ and $\mathbf{C} : u \mapsto \frac{\partial u}{\partial \nu}$ and for the consideration of Neumann conditions. We also assume that the unknown coefficient q is in $L^\infty(\Omega)$; there is then no further loss of generality in assuming, as we shall do, that $q > 0$ since a substitution $v := e^{-\alpha t}u$ replaces q by $q + \alpha$ and f, g by $e^{-\alpha t}f$, $e^{-\alpha t}g$ so $q \mapsto R_1(T,q)$ will be injective if and only if $q \mapsto R_1(T, q + \alpha)$ is injective.

Let $\mathbf{A} = \mathbf{A}_q$ be the elliptic operator $\mathbf{A} = -\boldsymbol{\Delta} + q$ on $\mathcal{H} := L^2(\Omega)$ with domain $\mathcal{D} = \mathcal{D}(\mathbf{A}) := \{u \in H^2(\Omega) : \mathbf{C}u = 0\}$. We note at this point the existence of an orthonormal (with respect to $\mathcal{H}$) basis of eigenfunctions

$$\mathbf{A}e_k = \lambda_k e_k \tag{2.1}$$

with $0 < \lambda_1 \leq \lambda_2 < \cdots \to \infty$ since we have taken $q > 0$.

We introduce the Green's operator $\mathbf{G}$ defined by $\mathbf{G} : \varphi \mapsto u$ with

$$-\mathbf{\Delta} u + qu = 0 \quad \text{on } \Omega, \qquad \mathbf{C}u = \varphi \in \mathcal{X} := L^2(\partial\Omega). \tag{2.2}$$

We certainly have $u \in H^1(\Omega)$ for arbitrary $\varphi \in \mathcal{X} = L^2(\partial\Omega)$, so noting [7], [8], and the equivalence of $H^s(\Omega)$ and $\mathcal{D}(\mathbf{A}^\vartheta)$ for $\vartheta = 2s$, we have

$$\mathbf{A}^{1/2}\mathbf{G} : \mathcal{X} \overset{\text{cont.}}{\longrightarrow} \mathcal{H} \tag{2.3}$$

(with $\mathbf{A}^\vartheta \mathbf{G} : \mathcal{X} \to \mathcal{H}$ for any $\vartheta < \frac{3}{4}$ if $\partial\Omega$ is, e.g., in C^1). Then the solution u of

$$\dot{u} + \mathbf{A}u = 0, \quad \mathbf{C}u = f(t) \quad \text{with } u\Big|_{t=0} = 0 \tag{2.4}$$

has the representation [3]

$$u(t) = \int_0^t [\mathbf{A}^{1/2}\mathbf{S}(t-s)][\mathbf{A}^{1/2}\mathbf{G}]f(s)\, ds, \tag{2.5}$$

where $\mathbf{S}(\cdot)$ is the (analytic) semigroup on $\mathcal{H}$ generated by $-\mathbf{A}$ so

$$\|\mathbf{A}^\nu \mathbf{S}(t)\| \le Mt^{-\nu}. \tag{2.6}$$

From (1.3) and the form of (2.5), we then see that $\mathbf{R}_1$ is a convolution operator,

$$[\mathbf{R}_1 f](t) = g(t) := \mathbf{B}u(t) = \int_0^\infty \mathbf{K}_1(t-s)f(s)\, ds, \tag{2.7}$$

with the kernel $\mathbf{K}_1(\cdot) = \mathbf{K}_1(\cdot; q)$ given by

$$\mathbf{K}_1(t) := \begin{cases} 0 & \text{for } t \le 0, \\ \mathbf{BAS}(t)\mathbf{G} & \text{for } t > 0, \end{cases} \tag{2.8}$$

where, noting (2.3), (2.6), and

$$\mathbf{BA}^{-\gamma} : \mathcal{H} \overset{\text{cont.}}{\longrightarrow} \mathcal{X} \qquad (\text{any } \gamma > 1/4), \tag{2.9}$$

we may write

$$\mathbf{BAS}(t)\mathbf{G} = \left[\mathbf{BA}^{-\gamma}\right]\left[\mathbf{A}^{1/2+\gamma}\mathbf{S}(t)\right]\left[\mathbf{A}^{1/2}\mathbf{G}\right]$$

with $\frac{1}{4} < \gamma < \frac{1}{2}$ to see that $\|\mathbf{K}_1(\cdot)\|$ is integrable whence $\mathbf{R}_1$ is, e.g., a continuous operator from $\mathcal{F}_T := L^2((0,T) \times \partial\Omega)$ to itself.

At this point it is convenient to shift to the Fourier representation for the semigroup. Using (2.1) in (2.8) gives the series representation

$$\mathbf{K}_1(t)\xi = \sum_k \lambda_k e^{-\lambda_k t} \langle e_k, \mathbf{G}\xi \rangle \mathbf{B}e_k \tag{2.10}$$

for $t > 0$. What we will actually need is the Laplace transform of this:

$$\begin{aligned} \hat{\mathbf{K}}_1(s)\xi &:= \int_0^\infty e^{-st}\mathbf{K}_1(t)\xi\, dt \\ &= \sum_k \lambda_k \int_0^\infty e^{-st}e^{-\lambda_k t} dt \langle e_k, \mathbf{G}\xi \rangle \mathbf{B}e_k \\ &= \sum_k \frac{\lambda_k}{s+\lambda_k} \langle e_k, \mathbf{G}\xi \rangle \mathbf{B}e_k \\ &= \mathbf{BA}(s+\mathbf{A})^{-1}\mathbf{G}\xi \qquad \text{for } s > 0. \end{aligned} \tag{2.11}$$

Note that the final form of this easily gives boundedness on $\mathcal{X}$ of $\hat{\mathbf{K}}_1(s)$ for $s \geq 0$, so we have no difficulties justifying convergence for the series and our manipulations. More precisely, we observe that everything certainly works well for the core of the operator (specification for ξ in a suitable dense set of nice functions), and then we can extend by continuity, using the final form.

With boundary conditions and initial conditions, the wave equation (1.6) now becomes

$$\ddot{w} + \mathbf{A}w = 0, \quad \mathbf{C}w = f \quad \text{with } w = 0 = \dot{w} \text{ at } t = 0, \tag{2.12}$$

and the response operator is the map

$$\mathbf{R}_2 = \mathbf{R}_2(\bar{T}, q) : f \mapsto \mathbf{B}w,$$

with w defined by (2.12) for the time interval $(0, \bar{T})$. It is well known that this $\mathbf{R}_2$ is a bounded operator from, e.g., $\mathcal{F}_{\bar{T}} := L^2((0,\bar{T}) \times \partial\Omega)$ to itself.

We proceed directly to the separation-of-variables solution, again expanding with respect to the orthonormal basis $\{e_k\}$,

$$w = \sum_k y_k(t)e_k, \qquad \mathbf{G}f = \sum_k \varphi_k(t)e_k.$$

One easily verifies from (2.12) that each $y_k(\cdot)$ is the solution of the ordinary differential equation

$$\ddot{y} + \lambda_k y = \lambda_k \varphi_k(t) \quad \text{with } y(0) = 0 = \dot{y}(0)$$

whence, noting that the assumed positivity $q > 0$ gives $\lambda_k > 0$, one has

$$y_k(t) = \mu_k \int_0^t [\sin \mu_k(t-s)]\varphi_k(s)\, ds \qquad \left(\mu_k := \sqrt{\lambda_k}\right).$$

Substituting, this gives the series representation

$$[\mathbf{R}_2 f](t) = \mathbf{B}w(t) = \int_0^t \sum_k \mu_k [\sin \mu_k(t-s)] \langle e_k, \mathbf{G}f(s)\rangle \mathbf{B}e_k\, ds, \tag{2.13}$$

so we see that $\mathbf{R}_2$ is a convolution operator$f \mapsto \mathbf{K}_2 * f$ with the kernel $\mathbf{K}_2(\cdot)$ given, corresponding to (2.10), by the series

$$\mathbf{K}_2(t)\xi = \sum_k \mu_k [\sin \mu_k t] \langle e_k, \mathbf{G}\xi\rangle \mathbf{B}e_k. \tag{2.14}$$

Again, we need the Laplace transform of this:

$$\begin{aligned} \hat{\mathbf{K}}_2(s)\xi &:= \int_0^\infty e^{-st}\mathbf{K}_2(t)\xi\, dt \\ &= \int_0^\infty e^{-st} \sum_k \mu_k [\sin \mu_k t] \langle e_k, \mathbf{G}\xi\rangle \mathbf{B}e_k\, dt \\ &= \sum_k \mu_k \int_0^\infty e^{-st} [\sin \mu_k t]\, dt\, \langle e_k, \mathbf{G}\xi\rangle \mathbf{B}e_k \\ &= \sum_k \frac{\lambda_k}{s^2 + \lambda_k} \langle e_k, \mathbf{G}\xi\rangle \mathbf{B}e_k \\ &= \mathbf{B}\mathbf{A}(s^2 + \mathbf{A})^{-1}\mathbf{G}\xi \qquad \text{for } s > 0. \end{aligned} \tag{2.15}$$

Again, we think of these manipulations as performed for nice ξ, with the result then extended by continuity, using the final form. Comparing (2.15) with (2.11) gives our key identity:

$$\hat{\mathbf{K}}_2(s;q) \equiv \hat{\mathbf{K}}_1(s^2;q) \qquad \text{for } s > 0. \tag{2.16}$$

Returning to (2.8), we observe that, since $\mathbf{S}(\cdot)$ is an analytic semigroup, the operator function $t \mapsto \mathbf{K}_1(t;q)$ is itself analytic in t (for complex t with positive real part). It follows that specification of $\mathbf{R}_1(T;q)$ implies specification of the kernel $\mathbf{K}_1(t;q)$ for $0 < t < T$ and so, by analyticity, uniqueness of the determination of $\mathbf{K}_1(\cdot;q)$ on $(0,\infty)$. This means that the Laplace transform $\hat{\mathbf{K}}_1(\cdot;q)$ is uniquely determined, as is the Laplace transform $\hat{\mathbf{K}}_2(\cdot;q)$, by the identity (2.16). By the standard uniqueness results for Laplace transforms, this means that $\mathbf{K}_2(t;q)$ is determined for $t > 0$, so the convolution operator $\mathbf{R}_2(\bar{T};q)$ is uniquely determined for arbitrary $\bar{T} > 0$. We have thus proved[1] the asserted implication in the following theorem.

THEOREM 2.1. *Suppose it is known, for some bounded Ω in $\mathbb{R}^n$ and a set $\boldsymbol{A}$ of bounded functions on Ω, that* (1.7) *holds for some $\bar{T}$. Then* (1.5) *holds for arbitrary $T > 0$.* (*Equivalently, if $q, \hat{q} \in \boldsymbol{A}$ with $q \not\equiv \hat{q}$, then $\mathbf{R}_1(T,q) \not\equiv \mathbf{R}_1(T,\hat{q})$ for all $T > 0$.*)

From [12] we have, restated in our notation, the following[2] result.

THEOREM 2.2 (R-S). *Let Ω be a bounded region in $\mathbb{R}^n$ with C^1 boundary $\partial\Omega$. Then* (1.7) *holds for $\boldsymbol{A} = L^\infty(\Omega)$.*

Combining this with Theorem 2.1, we immediately obtain the desired identifiability result for (1.1).

COROLLARY 2.3. *Let Ω be a bounded region in $\mathbb{R}^n$ with C^1 boundary $\partial\Omega$. Assume it is known that the coefficient q in* (1.1) *is in $\boldsymbol{A} = L^\infty(\Omega)$. Then q is uniqely determined by $\mathbf{R}_1(T,q)$.*

The observation that verification of our manipulations on a dense set is sufficient could become more significant if we wished to consider variations on the operator. In particular, if we wished to use Dirichlet data as input instead and then observe the corresponding Neumann data (reversing the roles of $\mathbf{B}$ and $\mathbf{C}$), then the regularity results would not be as cooperative, and it is useful to observe that equality on a core

[1] The argument also provides a partial converse to the implication. If one could independently show uniqueness of the correspondence $\mathbf{R}_1(T;q) \leftrightarrow q$ (for some T and some class of q), then one would necessarily have uniqueness for $\mathbf{R}_2(\cdot;q) \leftrightarrow q$, with observation now needed on all of $\mathbb{R}_+$ since analyticity in t is unavailable for $\mathbf{R}_2$ to obtain uniqueness from an interval without further information.

[2] We are indebted, for the reference to [12], to a referee of a previous version of this paper which referred, instead, to a sequence of recent papers, [1], [2], [4]–[6] by M. Belishev and others, which provide a reconstruction algorithm for q, justified under a control-theoretic hypothesis that the pair $[\Omega, q]$ is normal, i.e., $0 .< t < T_*$ the set of approximately reachable states by boundary control on $(0,t)$ is all of $\mathcal{H}_t := \{v \in \mathcal{H} : v(x) = 0 \text{ if } |x - \partial\Omega| > t\}$. Using duality, a sufficient condition for this normality is that $\partial\Omega$ and q be analytic for applicability of the classic Holmgren–John uniqueness theorem, although we note that this has quite recently been extended to the nonanalytic case by Tataru [17] (see, also, related results by Robbiano [13] and by Hörmander [9]). In comparison with [12], we observe that considerable regularity may be needed for normality in the reconstruction but not for the (nonconstructive) injectivity of $q \mapsto R_2(\bar{T}, q)$. On the other hand, using the results in [17] one can obtain uniqueness results applying to observation on a part of the boundary, while the results in, e.g., [6] consider more general wave equations

$$\rho(x)w_{tt} = \boldsymbol{\nabla} \cdot (\mu(x)\boldsymbol{\nabla} w) - qw, \tag{2.17}$$

where any two of the three coefficient functions ρ, μ, and q are assumed to be known, with the third coefficient to be recovered. The argument in §2 for our key identity (2.16) is valid also for these settings, so one would obtain corresponding identifiability results for the parabolic case. We view these as directions for future extensions of our present results.

suffices. Alternatively, one could obtain continuity using other boundary operators $\mathbf{B}$ and $\mathbf{C}$ by suitable adjustment of the spaces, perhaps admitting different $\mathcal{X}_{\text{in}}$ for f and $\mathcal{X}_{\text{out}}$ for g so that one can then proceed exactly as we have done. For such possible generalization, we also note that we do not need the full strength of the present self-adjointness of $\mathbf{A}$, giving orthonormality of the eigenfunctions in (2.1) but only, e.g., that $\Sigma\alpha_k e_k \mapsto \left[\Sigma|\alpha_k|^2\right]^{1/2}$ is an equivalent norm.

3. Identification with a single input. Theorem 2.1 and its corollary require complete knowledge of $\mathbf{R}_1$ in order to determine q. Interpreted directly, this would mean that one would need knowledge of all possible input/output pairs $[f, g]$ corresponding to (1.1), (1.2), and (1.3), requiring an infinite number of input/output experiments. Using the form of $\mathbf{R}_1$ given in (2.7) and (2.8), together with the regularity associated with (1.1), we now wish to show that a single experiment, observing the output g_* for a single properly chosen input f_*, will suffice to determine $\mathbf{K}_1(\cdot)$ and thus q.

Taking any total set (e.g., an orthonormal basis) $\{\xi_k\}$ for $\mathcal{X} = L^2(\partial\Omega)$ and a sequence of times $0 = t_1 < t_2 < \cdots \to T$, we may set

$$f_*(t) := \sum_{t_k<t} c_k\xi_k \qquad (0 < t < T) \tag{3.1}$$

with, e.g., $c_k := 2^{-k}$ ensuring convergence in $\mathcal{F}_T$. We set $\mathcal{I}_k := (t_k, t_{k+1})$, $\hat{\mathcal{I}}_k := (0, t_{k+1} - t_k)$ for $k = 1, 2, \dots$ so $\mathcal{I} := \bigcup_k \mathcal{I}_k = [0, T) \setminus \{t_1, t_2, \dots\}$. Clearly, $g_* := \mathbf{R}_1 f_*$ will be continuous and piecewise analytic in t on $\mathcal{I}$ with

$$\dot{g}_*(t) = \sum_{t_k<t} c_k\mathbf{K}_1(t - t_k)\xi_k \qquad (t \in \mathcal{I}). \tag{3.2}$$

Although g_* certainly depends on q, we note that no a priori information about q is needed for this construction of f_*.

From (3.2) one first notes that knowledge of g_* on $\mathcal{I}_1$ just gives $\mathbf{K}_1(\cdot)\xi_1$ on $\hat{\mathcal{I}}_1$ by differentiation and therefore determines $\mathbf{K}_1(t)\xi_1$ for all $t > 0$ by analyticity. Next, knowing g_* on $\mathcal{I}_2$ we may subtract the now-known $\frac{1}{2}\mathbf{K}_1(t)\xi_1$ from $\dot{g}_*$ to obtain $\mathbf{K}_1(\cdot)\xi_2$ on $\hat{\mathcal{I}}_2$ whence, again by analyticity, $\mathbf{K}_1(t)\xi_2$ would be known for all $t > 0$. Recursively, we similarly obtain each $\mathbf{K}_1(\cdot)\xi_k$ on $\hat{\mathcal{I}}_k$ and so on $\mathbb{R}_+$ for $k = 3, 4, \dots$. Thus, a single pair $[f_*, g_*]$ constructed in this fashion will uniquely determine $\mathbf{K}_1(t)\xi_k$ for each k and all $t > 0$ and hence will determine q.

The input function f_* is here piecewise constant in t, but we note that replacing f_* as input by its time integral just produces the time integral of g_* as output and thus also determines the original $\dot{g}_*$ of (3.2). Iterating this idea, we can use an input which is C^m in t for arbitrary m. We can get any desired spatial regularity by a suitable choice of $\{\xi_k(\cdot)\}$ as smooth functions on $\partial\Omega$.

THEOREM 3.1. *Given Ω and any $T > 0$ one can select a suitable* (*smooth*) *function $f_* \in \mathcal{F}_T$ such that the corresponding map $\mathbf{\Gamma}$ of* (1.10) *is injective when considered on $\boldsymbol{\mathcal{A}} \subset L^\infty(\Omega)$.*

Fixing Ω, T, and f_* as above, the injectivity of $\mathbf{\Gamma}$ in Theorem 3.1 means that (exact) observation of the output $g_* = \mathbf{\Gamma}(q)$ uniquely determines q. The obvious next question is whether this determination can be realized computationally: We would like an implementable procedure to recover q to any desired degree of accuracy, provided we are able to compute to arbitrary accuracy and to produce the input and measure the output with arbitrary accuracy. This is far from obvious in view of the ill-posedness of the problem for any reasonable topologies.

The argument for justification of any computational schema for the problem sets this in the context of a sequence of increasingly accurate approximating problems and then asserts the convergence of the computed approximants q_j to the true coefficient q. We begin by writing our a priori information about q in the form

$$q \in \mathcal{K} \subset \boldsymbol{\mathcal{A}} \subset L^\infty(\Omega). \tag{3.3}$$

Our principal assumptions here are that $\mathcal{K}$ is a closed subset of $L^\infty(\Omega)$ and that $\boldsymbol{\Gamma} : \mathcal{K} \overset{\text{cont.}}{\longrightarrow} \mathcal{G}$ for some suitable $\mathcal{G}$ topology with respect to which we can assume an increasingly accurate sequence of measurements $g_j \to g_*$. Standard techniques of numerical analysis enable us to provide computational solutions for the defining equations, giving a sequence of approximations $\boldsymbol{\Gamma}_j \to \boldsymbol{\Gamma}$. We assume here that this is uniform convergence on $\mathcal{K}$ but note that the convergence need only be at q if, instead, we would have uniform equicontinuity on $\mathcal{K}$ of the $\boldsymbol{\Gamma}_j$. The various approaches to ill-posed problems now each provide some selection procedure: Given $g_j, \boldsymbol{\Gamma}_j$ (with some accuracy estimate), there is a way to select $q_j \in \mathcal{K}$ so that $\boldsymbol{\Gamma}_j(q_j) \approx g_j$, and we may assume this is done in such a way as to have

$$[\boldsymbol{\Gamma}_j(q_j) - g_j] \to 0 \quad \text{as } k \to \infty. \tag{3.4}$$

If $\mathcal{K}$ is compact, the generic argument is to obtain (for a subsequence) convergence $q_j \to \bar{q}$ for some $\bar{q}$. We then have

$$\boldsymbol{\Gamma}(\bar{q}) - g_* = [\boldsymbol{\Gamma}(\bar{q}) - \boldsymbol{\Gamma}(q_j)] + [\boldsymbol{\Gamma}(q_j) - \boldsymbol{\Gamma}_j(q_j)] + [\boldsymbol{\Gamma}_j(q_j) - g_j] + [g_j - g_*],$$

and since each term on the right goes to 0, we conclude that $\boldsymbol{\Gamma}(\bar{q}) = g_*$. By our uniqueness theorem, we must then have $\bar{q} = q$. Finally, uniqueness of the limit makes the subsequence extraction irrelevant so, as desired, one has convergence of the sequence of computed approximants to the true solution $(q_j \to q)$ in the sense of the $\mathcal{K}$ topology.

As a variant of this, suppose one were to know a priori only that $q \in L^\infty(\Omega)$ but did not know any specific bound. We then propose the selection procedure: Choose

$$\|q_j\|_{L^2(\Omega)} + \|q_j\|_{L^\infty(\Omega)} \leq \min + \varepsilon_j \tag{3.5}$$

subject to a constraint on the residual error

$$\|\boldsymbol{\Gamma}_j(q_j) - g_j\| \leq \varepsilon'_j. \tag{3.6}$$

We make the assumptions that $\varepsilon_j \to 0$ and also that $\varepsilon'_j \to 0$, giving (3.4) but with ε'_j large enough (in comparison to the error estimates for the computational map $\boldsymbol{\Gamma}_j$ and for the observation g_j) that q itself is permitted to compete in the minimization, i.e., that (3.6) is satisfied with $q_j = q$.

THEOREM 3.2. *The computational procedure determined by* (3.5) *and* (3.6) *provides a sequence* (q_j) *which converges strongly to the true* q *in* $L^p(\Omega)$ *for all finite* p.

Proof. If we set

$$\alpha_j := \|q_j\|_{L^2(\Omega)}, \;\alpha := \|q\|_{L^2(\Omega)}, \;\beta_j := \|q_j\|_{L^\infty(\Omega)}, \;\beta := \|q\|_{L^\infty(\Omega)},$$

then (3.5), with the admissibility of q in (3.6), gives

$$\limsup[\alpha_j + \beta_j] \leq [\alpha + \beta]. \tag{3.7}$$

Since this means $\{\alpha_j\}$ is bounded, we must have, for a subsequence, weak convergence in $L^2(\Omega)$, i.e., $q_j \rightharpoonup \hat{q}$. Further, convexity gives

$$\|\hat{q}\|_{L^2(\Omega)} =: \hat{\alpha} \leq \liminf \alpha_j, \qquad \|\hat{q}\|_{L^\infty(\Omega)} =: \hat{\beta} \leq \liminf \beta_j. \tag{3.8}$$

The hypotheses, together with (3.6), ensure that

$$\lim \mathbf{\Gamma}(q_j) = \lim \mathbf{\Gamma}_j(q_j) = \lim g_j = g_* := \mathbf{\Gamma}(q).$$

Now let u_j be the solution of

$$u_t = \mathbf{\Delta} u - q_j u, \quad u_\nu = f_*, \quad u\Big|_{t=0} = 0 \tag{3.9}$$

and observe that the uniform L^∞ bound on q_j gives the standard (uniform) bound on u_j in $L^2([0,T] \to H^1(\Omega))$ and so also a uniform bound on $\dot{u}_j$ in $L^2([0,T] \to H^{-1}(\Omega))$. Using the Aubin compactness theorem, we may extract a further subsequence to have $u_j \to \hat{u}$ in, say, $L^2([0,T] \to H^s(\Omega))$ for any $s < 1$. From the weak formulation of the problem, one easily sees that for $q_j \rightharpoonup \hat{q}$, one has $\hat{u}$ satisfying the limit equation. Since the boundary trace is closed when applied to solutions of (3.9) and we already know that $\mathbf{C}u_j = g_j \to g_*$, it follows that $\mathbf{\Gamma}(\hat{q}) = g_*$, i.e., $\mathbf{\Gamma}(\hat{q}) = \mathbf{\Gamma}(q)$. Since (3.8) gives $\hat{q} \in L^\infty(\Omega)$, Theorem 3.1 now gives $\hat{q} = q$, and uniqueness of this limit means that we may ignore the previous extractions of subsequences. Since this gives $\hat{\alpha} = \alpha$ and $\hat{\beta} = \beta$, it follows from (3.7) and (3.8) that $\alpha_j \to \alpha$. This, together with the weak convergence $q_j \rightharpoonup q$, gives strong convergence $q_j \to q$ in the Hilbert space $L^2(\Omega)$. As Ω is bounded, this immediately gives $L^p(\Omega)$ convergence for $p \leq 2$, and the presence of an $L^\infty(\Omega)$ bound also gives $L^p(\Omega)$ convergence for all $p < \infty$. □

REFERENCES

[1] S. A. Avdonin, M. I. Belishev, and S. A. Ivanov, *Boundary control and a matrix inverse problem for the vector equation* $u_{tt} - u_{xx} + V(x)u = 0$, Mat. Sb., 182 (1991), pp. 307–331. (In Russian.) Math. USSR Sb., 72 (1992), pp. 287–310.

[2] ———, *Dirichlet boundary control in filled domains for the multidimensional wave equation*, Avtomatika, 2 (1991), pp. 86–90. (In Russian.) Eng. trans. in Soviet Automatic Control.

[3] A. V. Balakrishnan, *Applied Functional Analysis*, Springer-Verlag, New York, 1976.

[4] M. I. Belishev, *On an approach to a multidimensional inverse problem for the wave equation*, Dokl. Akad. Nauk SSSR, 297 (1987), pp. 524–527. (In Russian.). Soviet Math. Dokl. 36 (1988).

[5] ———, *Boundary control and wave field continuation*, preprint P-I-90 Leningrad Ordel Math. Inst. Steklov, Leningrad, 1990. (In Russian.)

[6] M. I. Belishev and Ya. V. Kurylev, *Boundary control, wave field continuation, and inverse problems for the wave equation*, Comput. Math. Appl., 22 (1991), pp. 27–52.

[7] D. Fujiwara, *Concrete characterization of the domains of fractional powers of some elliptic differential operators of the second order*, Proc. Japan Acad. Ser. A Math. Sci., 43 (1967), pp. 82–86.

[8] P. Grisvard, *Caractérisation de quelques espaces d'interpolation*, Arch. Rational Mech. Anal., 25 (1967), pp. 40–63.

[9] L. Hörmander, *A uniqueness theorem for second order hyperbolic differential equations*, Comm. Partial Differential Equations, 17 (1992), pp. 696–714.

[10] B. Lowe and W. Rundell, *Unique recovery of a coefficient in an elliptic equation from input sources*, Texas A & M University, 1993, preprint.

[11] A. NACHMAN, *Reconstructions from boundary measurements*, Ann. Math., 128 (1988), pp. 531–576.

[12] RAKESH AND W.W. SYMES, *Uniqueness for an inverse problem for the wave equation*, Comm. Partial Differential Equations, 13 (1988), pp. 87–96.

[13] L. ROBBIANO, *Théorème d'unicité adapté au contrôle des solutions des problèmes hyperboliques*, Comm. Partial Differential Equations, 16 (1991), pp. 789–800.

[14] D. L. RUSSELL, *A uniform boundary control theory for hyperbolic and parabolic partial differential equations*, Stud. Appl. Math., 52 (1973), pp. 189–211.

[15] T. I. SEIDMAN, *Exact boundary controllability for some evolution equations,* SIAM J. Control Optim., 16 (1978), pp. 979–999.

[16] ———, *Determination of the nonlinearity in a parabolic equation from boundary measurements,* in Control of Distributed Parameter Systems, 1989: Selected Papers from the Fifth IFAC Symposium, Perpignan, France, 26–29 June 1989, M. Amouroux and A. El Jai, eds., Pergamon, New York, 1990, pp. 181–186.

[17] D. TATARU, *Unique continuation for solution to pde's: Between Hormander's and Holmgren's theorems*, Dept. of Mathematics, Northwestern University, 1993, preprint.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 33, No. 4, pp. 1256–1269, July 1995

© 1995 Society for Industrial and Applied Mathematics

016

EXACT OBSERVABILITY OF THE TIME-VARYING HYPERBOLIC EQUATION WITH FINITELY MANY MOVING INTERNAL OBSERVATIONS*

A. YU. KHAPALOV†

Abstract. The problem of exact observability of the linear hyperbolic equation with time-varying coefficients under finitely many internal observations is considered. The question with which we are concerned in this paper is a sharp correspondence between the internal regularity of the solutions and a type of observation required to provide $L^\infty(0,T;R^{n+1})$- or $C([0,T[;R^{n+1})$-exact observability with respect to the energy norm. Two types of observations are considered: pointwise and spatially averaged, for which the existence of needed observation curves (continuous on $[0,T[$ for $n=1$) and set-valued maps (continuous on $[0,T[$ with respect to Lebesgue measure) is established. The techniques involved are related to the construction of suitable skeletons for these curves and maps.

Key words. time-varying hyperbolic equation, exact observability, moving observation

AMS subject classifications. Primary, 35L20; Secondary, 93C20

1. Introduction and problem formulation.

Let Ω be an open, bounded domain in R^n with boundary $\partial\Omega$. We consider the following initial-boundary value problem:

(1.1)
$$\frac{\partial^2 y}{\partial t^2} = \sum_{i,j=1}^{n} \frac{\partial}{\partial x_i}\left(a_{ij}(x,t)\frac{\partial y}{\partial x_j}\right) - \sum_{i=1}^{n} a_i(x,t)\frac{\partial y}{\partial x_i} - a(x,t)y \quad \text{in } Q = \Omega\times(0,T),$$

$$y = 0 \quad \text{in } \Sigma = \partial\Omega\times(0,T),$$

$$y\;|_{t=0} = y_0, \qquad y_t\;|_{t=0} = y_1.$$

We assume that the operator in the right-hand side of (1.1) is uniformly coercive:

$$\nu_1\sum_{i=1}^{n}\xi_i^2 \;\le\; \sum_{i,j=1}^{n} a_{ij}\,(x,t)\,\xi_i\xi_j \;\le\; \nu_2\sum_{i=1}^{n}\xi_i^2 \quad \text{for } \forall\xi_i\in R \quad \text{a.e. in } Q,$$

$$a_{ij}(x,t) = a_{ji}(x,t), \;\; i,j=1,\dots,n, \qquad \nu_1 = \text{const} > 0, \quad \nu_2 = \text{const} > 0.$$

The aim of this paper is to study the exact observability of (1.1) under finitely many scalar observations (as it generally occurs in applications) with respect to the energy space $H_E = H_0^1(\Omega)\times L^2(\Omega)$, which is of physical importance. The basic assumptions on the regularity of the solutions of the time-reversible problem (1.1) are

(1.2)
$$\{y,y_t\}\in C([0,T];H_E),\; dE^{1/2}(y(\cdot,t)) \le E^{1/2}(y(\cdot,0)) \le c\,E^{1/2}(y(\cdot,t)), \;\; \forall t\in[0,T],$$

* Received by the editors August 26, 1992; accepted for publication (in revised form) March 7, 1994.

† Department of Electrical and Computer Engineering, Oregon State University, Corvallis, Oregon 97331, and International Institute for Applied Systems Analysis, A-2361 Laxenburg, Austria.

where $d(= d(T))$, $c(= c(T)) > 0$ (e.g., $d = c^{-1}$) are given and $E^{1/2}(\cdot)$ is the energy norm,

$$E^{1/2}(y(\cdot,t)) = \left(\int_\Omega (\mid \nabla y(x,t) \mid^2 + y_t^2(x,t))\, dx\right)^{1/2},$$

$$\nabla = \left(\frac{\partial}{\partial x_1}, \dots, \frac{\partial}{\partial x_n}\right), \quad \mid \cdot \mid = \| \cdot \|_{R^n}.$$

(Recall that, due to Poincaré's inequality, $E^{1/2}(y(\cdot,t))$ is equivalent to the standard H_E-norm.) The general structure of the observations of our further interest is derived from (1.2):

$$(1.3) \qquad [0,T] \ni t \rightarrow z(t) = \mathbf{G}(t)\{\nabla y(\cdot,t), y_t(\cdot,t)\} \in R^{n+1}.$$

The question of primary concern in this paper is a sharp correspondence between assumptions on the internal regularity of the solutions of (1.1), which may vary (e.g., in the context of pointwise observations) with the growth of the space dimension, and a type of finite-dimensional operator (observation) in (1.3) that is required to provide $L^\infty(0,T;R^{n+1})$- or $C([0,T[;R^{n+1})$-exact observability in a given space dimension.

DEFINITION 1.1. *Given a normed space B and a linear manifold $H \subset H_E$, the system* (1.1)–(1.3) *is said to be* B-exactly observable (*this paper deals with* $B = L^\infty(0,T;R^{n+1})$ *or* $C([0,T[;R^{n+1})$) *on H with respect to the energy norm* (*we omit the latter in the text*) *if*

$$(1.4) \qquad \exists \gamma > 0 \;\; \textit{such that} \;\; \| \mathbf{G}\{\nabla y, y_t\} \|_B \geq \gamma\, E^{1/2}(y(\cdot,0))$$

for any y that satisfies (1.1)–(1.2) *and such that* $\{y(\cdot,0), y_t(\cdot,0)\} \in H$.

Remark 1.1. This definition takes into account a situation typically arising in the context of infinite-dimensional studies, namely that the domain of the observation operator may not match the regularity of the solutions of the system considered (while being, say, densely defined). Note that it treats the well-posedness of the observation (1.3) as the enclosure of the output z in (1.3) into B when $\{y(\cdot,0), y_t(\cdot,0)\}$ ranges all over H, while the continuity with respect to this pair in the energy norm is not required. On the other hand, it is clear that (1.4) (if it holds) generates some topology on H. The interrelation between Definition 1.1 and the *dual* issue of *exact controllability* is briefly discussed at the end of the next section.

In the context of the time-invariant setting the problem of exact observability for (1.1)–(1.3), being of traditional practical interest, has received considerable attention in the literature. In particular, how the Hilbert Uniqueness Method can be linked with the static pointwise sensor structure was discussed by Lions [13] (see also El Jai and others [3]) for the wave equation under the following observation: $z(t) = y(\bar{x},t)$, $t \in [0,T]$, when $B = L^2(0,T)$. For the same system the results of Triggiani and Tataru [21], [22] on exact controllability imply that for $n = 2,3$ $L^2(0,T)$-exact observability with respect to the energy norm is not possible. The techniques of the above-mentioned papers include those of harmonic and nonharmonic analysis. The $L^2(0,T)$-exact observability of the one-dimensional wave equation with *static* point observation as in (2.1) (i.e., when $\bar{x}(\cdot) \equiv \bar{x}$) was established for $T > 2 \times \max\{1 - \bar{x}, \bar{x}\}$ by Ho [5], who used the multipliers techniques. In [8], by making use of the integral formula for the general solution of the wave equation, it is shown that the infimum of the just-mentioned observation time can indeed be achieved. We stress that the techniques used in all the above works are based heavily on the time-invariant properties of the problems considered. In contrast to the time-invariant case, very little is known about the general *time-varying* problem (1.1)–(1.3).

In the present paper we establish the existence of moving (this is natural for the time-varying processes) $(n+1)$-dimensional observations (1.3) of two types: the point and the spatially averaged observations (2.1) and (2.6), which are able to ensure the $L^\infty(0,T;R^{n+1})$- or $C([0,T[;R^{n+1})$-exact observability of (1.1)–(1.3) on manifolds that are sharply linked with the assumptions on the internal regularity of the equation (1.1). The techniques employed are related to the construction of suitable skeletons for observation curves and set-valued maps, associated with (2.1) and (2.6), and based on a priori energy estimates (1.2). We remind the reader that (1.1) is *not* conservative. This approach was applied to the wave equation (in an arbitrary space dimension) with moving point observation in [6]–[8] under the assumption $\{\nabla y, y_t\} \in [C([0,T] \times \bar{\Omega})]^n$ ("$-$" stands for the closure). In this paper we show that the same approach is capable of handling the linear time-varying hyperbolic equation (1.1) under rather general and, in a certain sense, minimal assumptions on its regularity.

The paper is organized as follows. Section 2 states the main exact observability results, Theorems 2.1, 2.2, 2.4, and 2.6, which are then proven in §§3 and 5. Section 4 discusses auxiliary properties of the spatially averaged observations. In the appendix we refine the result of Theorem 2.6.

2. Main results.

We begin by the point observation

$$z(t) = \{\nabla y(\bar{x}(t),t), y_t(\bar{x}(t),t)\}, \quad t \in [0,T], \tag{2.1}$$

where $\bar{x}(t) \in \bar{\Omega}$ a.e. in $[0,T]$ (in fact, we can consider only internal curves; see Remark 3.1), is a given function (curve) measurable with respect to Lebesgue measure. In general, this observation is ill defined on the solutions of the system (1.1)–(1.2).

Assumption 2.1. $H \subset H_E \cap \{\{y(\cdot,0), y_t(\cdot,0)\} \mid \{\nabla y, y_t\} \in L^\infty(0,T;[C(\bar{\Omega})]^{n+1})\}$ is a linear manifold *separable* with respect to the following norm:

$$\| \{y(\cdot,0), y_t(\cdot,0)\} \| = \| \{y, y_t\} \|_{C([0,T];H_E)} + \operatorname*{ess\,sup}_{t\in[0,T]} \| \{\nabla y(\cdot,t), y_t(\cdot,t)\} \|_{[C(\bar{\Omega})]^{n+1}} . \tag{2.2}$$

Remark 2.1. The condition $\{\nabla y, y_t\} \in L^\infty(0,T;[C(\bar{\Omega})]^{n+1})$ in the above is to ensure the enclosure of the output in (2.1) into $L^\infty(0,T;R^{n+1})$, whereas *separability* is due to the techniques applied in this paper (recall along these lines that $L^\infty(Q)$ is not separable). A number of requirements (beginning with those providing the classical solutions) on the system (1.1) that imply Assumption 2.1 can be found in the literature (see, e.g., [15], [17], [9]).

THEOREM 2.1 (point observation). *Let $H \subset H_E$ satisfy Assumption* 2.1 *and $T > 0$ be given. Then there exists a class of measurable curves $\bar{x}(\cdot)$, which make the system* (1.1)–(1.2), (2.1) *$L^\infty(0,T;R^{n+1})$-exactly observable on H.*

THEOREM 2.2 (the one-dimensional case). *Let $\Omega =]0,1[$, $T > 0$ be given and Assumption* 2.1 *be fulfilled for H. Assume, in addition, that $y \in H^2(Q)$ when $\{y(\cdot,0), y_t(\cdot,0)\} \in H$. Then the observation curves can be selected in Theorem* 2.1 *to be arbitrarily smooth on $[0,T[$ and to lie entirely in Ω.*

The condition

$$\operatorname*{ess\,sup}_{Q} \mid a_{11t}, a_{11tt}, a_{11x}, a_1, a_{1t}, a, a_t \mid \leq \text{ const}, \quad n = 1 \tag{2.3}$$

(where all the derivatives are understood in the generalized sense), given in Ladyzhenskaya [9, pp. 162, 164], ensures the fulfillment of the assumptions of Theorem 2.2 with

$$H = (H^2(\Omega) \cap H_0^1(\Omega)) \times H_0^1(\Omega). \tag{2.4}$$

COROLLARY 2.3. *Let* (2.3) *be verified. Then Theorem* 2.2 *holds for* H *as in* (2.4).

Theorem 2.2 admits a straightforward extension to the following type of observations:

$$z(t) = \begin{pmatrix} \text{meas}^{-1}\{S_{h(t)}(\bar{x}(t))\} \int_{S_{h(t)}(\bar{x}(t))} y_x(x,t)\,dx \\ \text{meas}^{-1}\{S_{h(t)}(\bar{x}(t))\} \int_{S_{h(t)}(\bar{x}(t))} y_t(x,t)\,dx \end{pmatrix}, \quad t \in [0,T], \tag{2.5}$$

where $S_h(\bar{x}) = \{x \mid \mid x - \bar{x} \mid < h,\ x \in]0,1[\}$ and meas$\{\cdot\}$ here and elsewhere stands for Lebesgue measure.

THEOREM 2.4. [1] *Let* $\Omega =]0,1[$. *Given* $T > 0$, *there exist arbitrarily smooth functions* $\bar{x}(t) \in \Omega$, $h(t) > 0$, $t \in [0,T[$, *which make the system* (1.1)–(1.2), (2.3), (2.5) $C([0,T[;R^2)$*-exactly observable on* H *as in* (2.4).

When the space dimension is higher than one we do not manage to extend Theorem 2.4 to the general multidimensional case under the assumptions (1.2): this would require, from our point of view, an additional pointwise regularity of the solutions (e.g., as in Assumption 2.1; see Remark 3.3), which in this paper in the context of the spatially averaged observations seems to us redundant. Therefore, we enlarge the class of the spatially averaged observations (2.5) as follows:

$$z(t) = \frac{1}{\text{meas}\{\Omega(t)\}} \begin{pmatrix} \int_{\Omega(t)} v_1^0(x,t) y_{x_1}(x,t)dx \\ \vdots \\ \int_{\Omega(t)} v_n^0(x,t) y_{x_n}(x,t)dx \\ \int_{\Omega(t)} v^1(x,t) y_t(x,t)dx \end{pmatrix}, \quad t \in [0,T], \tag{2.6}$$

where $[0,T] \ni t \longrightarrow \Omega(t) \subset \Omega$ is a given set-valued map from $[0,T]$ into the set of all measurable subsets of Ω and $\{v_p^0\}_{p=1}^n$, v^1 are given measurable functions of a sign-type

$$\mid v_p^0(x,t) \mid, \mid v^1(x,t) \mid = +1 \text{ or } 0 \quad \text{a.e. in } Q, \quad p = 1,\ldots,n.$$

To avoid a misunderstanding, it is assumed here and everywhere below that if $\Omega(t)$ is of zero-measure at some t^*, then $\mathbf{G}(t^*)$ is the zero-operator. The last type of observations may be considered a generalization of the concept of a moving point observation to the case when a curve cannot be associated with a well-defined finite-dimensional observation of interest. The map $\Omega(\cdot)$ then plays a "pure" role (compared with (2.5)) of an observation curve. By virtue of (1.2), the observation (2.6) is well defined for any $t \in [0,T]$ and all the outputs are bounded when $\{y(\cdot,0), y_t(\cdot,0)\} \in H_a$, where

$$H_a = H_E \cap \{\{y(\cdot,0), y_t(\cdot,0)\} \mid \{\nabla y, y_t\} \in L^\infty(Q; R^{n+1})\}. \tag{2.7}$$

DEFINITION 2.5. *We shall say that a set-valued map* $[0,T] \ni t \longrightarrow F(t) \subset \Omega$ *is continuous with respect to Lebesgue measure at* $t = t^*$ *if*

$$\text{meas}\{F(t^*)\ \Delta\ F(t^* + \Delta t)\} \longrightarrow 0, \qquad \Delta t \longrightarrow 0,$$

[1] In the context of $L^\infty(0,T;R^2)$-exact observability this result was announced in [6].

where $A\Delta B$ stands for the symmetric difference $A\Delta B = (A\backslash B) \cup (B\backslash A)$.

Remark 2.2. The continuity of $\Omega(t)$ at $t = t^*$ in the sense of Definition 2.1 implies the continuity of the function $f(t) = \text{meas}\{\Omega(t)\}$ at $t = t^*$. Moreover, if $\text{meas}\{\Omega(t^*)\} > 0$, then all the outputs of (1.1)–(1.2), (2.6) are continuous at $t = t^*$.

THEOREM 2.6 (Spatially averaged observation). *Let $H_a \neq \emptyset$ and $T > 0$ be given. Then there exists a class of set-valued maps $\Omega(\cdot)$ continuous with respect to Lebesgue measure on $[0, T[$ and associated functions $\{v_p^0\}_{p=1}^n, v^1 \subset C([0, T[; L^2(\Omega))$ that make the system* (1.1)–(1.2), (2.6) *$C([0, T[; R^{n+1})$-exactly observable on $H = H_a$.*

We emphasize that the proof of this theorem, given in §5, is constructive (as well as those of Theorems 2.1, 2.2, and 2.4) and does not involve the constraint (2.7), which is to outline an a priori largest class of those solutions that are consistent with C (or L^∞) space for the (thus, *a priori bounded*) outputs. Once the set-valued map and functions satisfying Theorem 2.6 are found (see (5.2), (5.3)), the set H in (2.7) can be extended (due to (5.7)) to the following "a posteriori" largest set (see also §6):

$$(2.8) \qquad H_{ap} = H_E \cap \{\{y(\cdot, 0), y_t(\cdot, 0)\} \mid z \in C([0, T[; R^{n+1}), \\ \text{where } z \text{ is due to } (1.1), (1.2), (2.6)\}.$$

COROLLARY 2.7. *Let the observation* (2.6) *be defined by* (5.2), (5.3) *in §5. Then the assertion of Theorem* 2.6 *holds for $H = H_{ap}$.*

Remark 2.3. The following assumption ensures (1.2) [9, p. 167]:

$$(2.9) \qquad \operatorname*{ess\,sup}_{Q} \mid a_{ijt}, a_{ijtt}, a_{ijx}, a_i, a, \mid \leq \text{ const},$$

where the derivatives are understood in the generalized sense.

COROLLARY 2.8. *Let* (2.9) *be verified. Then Theorem* 2.6 *and Corollary* 2.7 *hold.*

Note that Theorems 2.1, 2.2, 2.4, and 2.6 deal with observations that employ the only curve or set-valued map. In the appendix we show that the usage of independent maps for each of $y_{x_p}, p = 1, \ldots, n, y_t$ can considerably improve the value of γ in (1.4) obtained by Theorem 2.6; see (5.7) and (A.4).

Controllability. Let $\mathbf{K} = \mathbf{GS}, \mathbf{K} : H \to B$, where $\mathbf{S}(t)\{y_0, y_1\} = \{\nabla y(\cdot, t), y_t(\cdot, t)\}$, $t \in [0, T]$. Then the B-exact observability property is equivalent to the bounded invertibility of the operator $\mathbf{K}$ on its range with respect to the energy norm (see Definition 1.1) and, hence, the maximal value of γ in (1.4) is equal to $(\| \mathbf{K}^{-1} \|)^{-1}$. A *direct* (duality) method (see, e.g., [19], [2], [18], [12], [20], [10], [11] and the bibliography therein) implies that if a linear manifold H is fundamental (dense) in H'_E, then B-exact observability is equivalent to exact controllability of the dual control system in $H'_E = L^2(\Omega) \times [H_0^1(\Omega)]'$ with controls from B', the dual of B. Otherwise, one has exact controllability in H', the dual of H as a linear topological manifold in H_E. A detailed study of the issue of exact controllability of the wave equation in an arbitrary space dimension under moving point control dual to (2.1) when $B = L^\infty(0, T; R^{n+1})$ or $L^2(0, T; R^{n+1})$ is given in [7].

3. Point observation.

Proof of Theorem 2.1. *Step* 1. Let Y_H be the set of all the solutions of (1.1)–(1.2) such that $\{y(\cdot, 0), y_t(\cdot)\} \in H$. We show first how for any given $y \in Y_H$ one can find a curve that ensures the estimate (1.4) for this particular solution.

Fix y and let $\beta > 0$ be given. By virtue of Assumption 2.1, both $\nabla y(x, t)$ and $y_t(x, t)$ are of Carathéodory type and, hence, for any measurable function $\bar{x}(t)$, $t \in$

$[0,T]$, the observation (2.1) is well defined. Let r be an arbitrary subinterval of $[0,T]$, $\text{meas}\{r\} > 0$. Next we set

(3.1)
$$e = \left\{ (x,t) \in \bar{\Omega} \times r \;\middle|\; |\nabla y(x,t)|^2 + y_t^2(x,t) \geq \operatorname*{ess\,sup}_{(x,t)\in\Omega\times r} (|\nabla y(x,t)|^2 + y_t^2(x,t)) - \beta \right\}.$$

Consider the set-valued map $r^* \ni t \longrightarrow F(t) = \{x \mid (x,t) \in e\}$, where $r^* = \text{dom } F(t) = \{t \mid \{x| (x,t) \in e\} \neq \emptyset\}$. By Assumption 2.1, the sets $F(t)$ are closed almost everywhere in r^*. Applying the measurable selection theorem (see, e.g., [4], [1]) then yields the existence of a measurable function $\bar{x}(t)$, $t \in [0,T]$ such that

$$\bar{x}(t) \in F(t) \quad \text{a.e. in } r^*. \tag{3.2}$$

Note next that by (1.2)

$$c^{-2}E(y(\cdot,0)) \leq \inf_{t\in r^*} E(y(\cdot,t)) \leq \text{meas}^{-1}\{r^*\} \int_{r^*}\int_{\Omega} (|\nabla y(x,t)|^2 + y_t^2(x,t))dxdt \tag{3.3}$$
$$\leq \text{meas}\,\{\Omega\} \operatorname*{ess\,sup}_{(x,t)\in\Omega\times r^*} (|\nabla y(x,t)|^2 + y_t^2(x,t)).$$

Combining (3.1)–(3.3) yields the following *basic* estimate:

$$E^{1/2}(y(\cdot,0)) \leq c\,\text{meas}^{1/2}\{\Omega\} (\| \{\nabla y(\bar{x}(\cdot),\cdot), y_t(\bar{x}(\cdot),\cdot)\} \|^2_{L^\infty(r^*;R^{n+1})} + \beta)^{1/2}. \tag{3.4}$$

Step 2. Take any $\delta > 0$. Select next in Y_H a *countable* (this can be done by Assumption 2.1) δ-net $\{y^k\}_{k=1}^\infty$ in Y_H as follows (see (2.2)): for any $y \in Y_H$ there is an element y^k for which

$$E^{1/2}(y(\cdot,t) - y^k(\cdot,t)) \leq \delta, \quad \forall t \in [0,T], \tag{3.5a}$$

$$\operatorname*{ess\,sup}_{t\in[0,T]} \| \{\nabla y(\cdot,t) - \nabla y^k(\cdot,t), y_t(\cdot,t) - y_t^k(\cdot,t)\} \|_{[C(\bar{\Omega})]^{n+1}} \leq \delta. \tag{3.5b}$$

Step 3. We proceed now with the construction of an observation curve that ensures the estimate (1.4). Let $\{t_k\}_{k=1}^\infty$ be an arbitrary strictly increasing sequence in $]0,T[$. The estimate (3.4) implies that for each k there is a curve $\{\bar{x}^k(\cdot)\}$ such that

(3.6)
$$E^{1/2}(y^k(\cdot,0)) \leq c\,\text{meas}^{1/2}\{\Omega\}(\| \{\nabla y^k(\bar{x}^k(\cdot),\cdot), y_t^k(\bar{x}^k(\cdot),\cdot)\} \|^2_{L^\infty(t_k,t_{k+1};R^{n+1})} + \beta)^{1/2}.$$

Let $\bar{x}(\cdot)$ be an arbitrary measurable trajectory defined on $[0,T]$ such that

$$\bar{x}(t) = \bar{x}^k(t), \quad t \in\,]t_k, t_{k+1}[, \qquad k = 1, \ldots. \tag{3.7}$$

Step 4. Let us show that $\bar{x}(\cdot)$ satisfies the requirements of Theorem 2.1. Take any $y \in Y_H$ and an element $y^k \in Y^\delta$ for which (3.5) is verified. In particular,

$$E^{1/2}(y(\cdot,0)) \leq E^{1/2}(y^k(\cdot,0)) + \delta.$$

This and (3.6)–(3.7) imply

$$\begin{aligned} E^{1/2}(y(\cdot,0)) &\leq c\,\text{meas}^{1/2}\{\Omega\} (\| \{\nabla y^k(\bar{x}(\cdot),\cdot), y_t^k(\bar{x}(\cdot),\cdot)\} \|^2_{L^\infty(t_k,t_{k+1};R^{n+1})} + \beta)^{1/2} + \delta \\ &\leq c\,\text{meas}^{1/2}\{\Omega\} (\| \{\nabla y^k(\bar{x}(\cdot),\cdot), y_t^k(\bar{x}(\cdot),\cdot)\} \|_{L^\infty(t_k,t_{k+1};R^{n+1})} + \sqrt{\beta}) + \delta. \end{aligned}$$

Now, by virtue of (3.5b), we arrive at

$$E^{1/2}(y(\cdot,0)) \leq c\,\mathrm{meas}^{1/2}\{\Omega\}(\| \{\nabla y(\bar{x}(\cdot),\cdot), y_t(\bar{x}(\cdot),\cdot) \|_{L^\infty(t_k,t_{k+1};R^{n+1})} + \delta + \sqrt{\beta}) + \delta. \tag{3.8}$$

In other words, we obtain the estimate (1.4) with $\gamma = (c\,\mathrm{meas}^{1/2}\{\Omega\}(1+\delta+\sqrt{\beta})+\delta)^{-1}$. This completes the proof of Theorem 2.1.

Remark 3.1. In fact, the assertion of Theorem 2.1 holds true in the class of those curves that lie entirely in the interior of Ω. Indeed, given β, y, one can replace Ω in the right-hand side of (3.1) by any strictly interior subdomain, which still preserves nonemptiness of such a modified set e $(= e(y,\beta))$. The rest of the proof is much the same as in the above.

Proof of Theorem 2.2. Let $\lim_{k\to\infty} t_k = T$. Since $y \in H^2(Q)$, due to the embedding theorem [15], [17], both $y_x(x,t)$ and $y_t(x,t)$ are continuous in x for almost all $t \in [0,T]$ and in t for almost all $x \in [0,1]$. Therefore, one can obtain the required assertion while avoiding use of the measurable selection theorem. Indeed, let $y \in Y_H$ be fixed. In Step 1 of the proof of Theorem 2.1 take any instant $t_* \in r^*$ such that $y_x(x,t_*), y_t(x,t_*) \in C[0,1]$. Without loss of generality we can assume that $F(t_*)$ contains a nontrivial interval. Then there is a point $x_* \in]0,1[$ such that $(x_*,t_*) \in e$ (see (3.1)) and $y_x(x_*,t), y_t(x_*,t) \in C[0,T]$. Using continuity of y in t, x, as was mentioned in the above, one then comes to the conclusion that (3.4) is verified for any continuous curve passing through x_* at time t_*, which is constant in some neighborhood of t_*. This allows us to compose $\bar{x}(\cdot)$ in (3.7) to be arbitrarily smooth on $[0,T[$. The rest of the proof follows along Steps 2–4 in the proof of Theorem 2.1.

Remark 3.2. The proof of Theorem 2.1 deals with a net in the set Y_H. However, by making use of the linearity of (1.1) and (2.1), this set can be replaced by its subset, which consists of those solutions whose energy norms at $t=0$ are equal to 1 (this can be applied to the proofs of Theorems 2.2, 2.4, and 2.6 as well). The estimate (3.8), being then derived only for the latter, implies (1.4) with $\gamma = (1 - c\,\mathrm{meas}^{1/2}\{\Omega\}(\delta + \sqrt{\beta}) - \delta)(c\,\mathrm{meas}^{1/2}\{\Omega\})^{-1}$ for β and δ appropriately small.

Proof of Theorem 2.4. The scheme of the proof is as much the same as that of Theorems 2.1 and 2.2.

Step 1. From the proofs of Theorem 2.1 (Step 1), Theorem 2.2, and Remark 2.2 it immediately follows that, given $\beta^* > 0$ and an interval $r \subset [0,T]$, for any solution y of (1.1)–(1.2), (2.3)–(2.4) there exists a pair of (arbitrarily smooth) functions $h(t) > 0$, $\bar{x}(t) \in]0,1[$, $t \in [0,T[$, and a nontrivial interval $r^* \subseteq r$ such that

$$\begin{aligned} &\left(\mathrm{meas}^{-1}\{S_{h(t)}(\bar{x}(t))\} \int\limits_{S_{h(t)}(\bar{x}(t))} y_x(x,t)dx\right)^2 \\ &+ \left(\mathrm{meas}^{-1}\{S_{h(t)}(\bar{x}(t))\} \int\limits_{S_{h(t)}(\bar{x}(t))} y_t(x,t)dx\right)^2 \\ &\geq \text{ ess sup}_{(x,t)\in\Omega\times r}(y_x^2(x,t) + y_t^2(x,t)) - \beta^*, \quad \forall t \in r^*. \end{aligned} \tag{3.9}$$

Step 2. Since (3.9) plays a role similar to the relations (3.1)–(3.2), we can establish an estimate analogous to (3.4) for the observation (2.5) as well. Taking into account that

$$\text{meas}^{-1}\{S_{h(t)}(\bar{x}(t))\} \int\limits_{S_{h(t)}(\bar{x}(t))} 1\, dx = 1,$$

(3.5b) implies

$$\| \mathbf{G}(\cdot)\{y_x - y_x^k,\ y_t - y_t^k\} \|_{L^\infty(t_k,t_{k+1};R^2)} \leq \sqrt{2}\,\delta,$$

where $\mathbf{G}(\cdot)$ stands for the observation operator in (2.5). Since the latter plays a crucial role in the derivation of (3.8) along (3.6)–(3.7), by taking into account Remark 2.2 one can obtain the conclusion of Theorem 2.4 with respect to $C([0,T[;R^2)$-exact observability in a similar way, constructing (as in (3.7)) a needed set-valued map $[0,T] \ni t \rightarrow S_{h(t)}(\bar{x}(t))$ to be continuous and of positive measure on $[0,T[$.

Remark 3.3. It is readily seen that the proof of Theorem 2.4 admits an extension to the general multidimensional case under Assumption 2.1.

4. Generalized spatially averaged observation. We begin by studying the properties of the observation (2.6).

Assumption 4.1. Let r be a given subinterval of $[0,T]$. A set-valued map $[0,T] \ni t \rightarrow \Omega(t) \subset \Omega$ satisfies Assumption 4.1 on r if the set $\{(x,t) \mid x \in \Omega(t),\ t \in r\}$ is measurable with respect to Lebesgue measure on $\Omega \times (0,T)$ and

$$\operatorname*{ess\,inf}_{t\in r}\ \text{meas}\{\Omega(t)\} > 0. \tag{4.1}$$

It is clear that if $\Omega(t)$ satisfies Assumption 4.1 on r, then all the outputs of the system (1.1)–(1.2), (2.6) lie in $L^\infty(r;R^{n+1})$.

The following class of set-valued maps plays a crucial role in the proof of Theorem 2.6. Let y be an arbitrary solution of the system (1.1)–(1.2) and $\beta > 0$ be given. Set

(4.2)
$$r(y,t) = \{x \in \Omega \mid |\nabla y(x,t)|^2 + y_t^2(x,t) > \text{meas}^{-1}\{\Omega\}\ E(y(\cdot,t)) - \beta^2\},\quad t \in [0,T],$$

$$\begin{aligned} r^0_{p+}(y,t) &= \{x \in r(y,t) \mid y_{x_p}(x,t) > 0\},\\ r^0_{p-}(y,t) &= \{x \in r(y,t) \mid y_{x_p}(x,t) < 0\},\quad p = 1,\dots, \end{aligned} \tag{4.3a}$$

(4.3b)
$$r^1_+(y,t) = \{x \in r(y,t) \mid y_t(x,t) > 0\},\qquad r^1_-(y,t) = \{x \in r(y,t) \mid y_t(x,t) < 0\}.$$

It is not hard to see that $r(y,t)$ is of positive measure for any $t \in [0,T]$. Furthermore, from (1.2) it follows that for any $\varepsilon > 0$

(4.4)
$$\text{meas}\{x \mid |\nabla y(x,t+\Delta t) - \nabla y(x,t)| \geq \varepsilon\} \rightarrow 0,\quad \text{when}\quad \Delta t \rightarrow 0,\quad \forall t \in [0,T],$$

$$\text{meas}\{x \mid |y_t(x,t+\Delta t) - y_t(x,t)| \geq \varepsilon\} \rightarrow 0,\quad \text{when}\quad \Delta t \rightarrow 0,\quad \forall t \in [0,T]. \tag{4.5}$$

DEFINITION 4.1. *We shall say that a set-valued map $[0,T] \ni t \rightarrow F(t) \subset \Omega$ is lower semicontinuous (see, e.g., [1] and the bibliography therein for various definitions of continuity of set-valued maps) at $t = t^*$ with respect to Lebesgue measure if $\forall \varepsilon_1 > 0$ $\exists\, \varepsilon_2 > 0$ such that*

$$\text{meas}\{F(t^*) \setminus F(t^* + \Delta t)\} \leq \varepsilon_1,\quad \forall\, \Delta t:\ |\Delta t| \leq \varepsilon_2. \tag{4.6}$$

The relations (4.4), (4.5) and the fact that the inequality in (4.2) is strict imply that $r(y,t)$ satisfies (4.6) everywhere in $[0,T]$. Hence, the function $f(t) = \text{meas}\{r(y,t)\}$ is lower semicontinuous on $[0,T]$ and, therefore, reaches its minimum on $[0,T]$ (compare with (4.1)).

Remark 4.1. Let y be an arbitrary solution of the system (1.1)–(1.2). Then one can deduce from (4.4)–(4.6) that for any sequence $\{y_i\}_{i=1}^{\infty}$ that converges to y in the $C([0,T];H_E)$-norm the following estimate holds:

$$\liminf_{i\to\infty}\left(\min_{t\in[0,T]}\text{meas}\{r(y_i,t)\}\right) \geq \min_{t\in[0,T]}\text{meas}\{r(y,t)\} > 0.$$

Given $\Omega(t)$ in (2.6), set for any $t \in [0,T]$

$$\Omega^0_{p+}(t) = \{x \mid v^0_p(x,t) = 1, \ x \in \Omega(t)\}, \ \ \Omega^0_{p-}(t) = \{x \mid v^0_p(x,t) = -1, \ x \in \Omega(t)\},$$

$$\Omega^1_{+}(t) = \{x \mid v^1(x,t) = 1, \ x \in \Omega(t)\}, \ \ \ \Omega^1_{-}(t) = \{x \mid v^1(x,t) = -1, \ x \in \Omega(t)\}.$$

LEMMA 4.2. *Let y be an arbitrary solution of the system* (1.1)–(1.2) *and $\beta > 0$, $t^* \in [0,T]$ be given. Then any set-valued map $\Omega(\cdot)$ and functions $\{v^0_p)\}_{p=1}^n$, v^1 in* (2.6) *such that*

$$\Omega(t^*) = r(y,t^*), \quad \Omega^0_{p\pm}(t^*) = r^0_{p\pm}(y,t^*), \quad \Omega^1_{\pm}(t^*) = r^1_{\pm}(y,t^*), \quad p = 1,\dots,n, \tag{4.7}$$

ensure the estimate

(4.8)

$$E(y(\cdot,0)) \leq c^2(n+2)\,\text{meas}\{\Omega\}\left(\sum_{p=1}^n\left(\text{meas}^{-1}\{\Omega(t^*)\}\int_{\Omega(t^*)} v^0_p(x,t^*)\,y_{x_p}(x,t^*)\,dx\right)^2\right.$$

$$\left. + \left(\text{meas}^{-1}\{\Omega(t^*)\}\int_{\Omega(t^*)} v^1(x,t^*)\,y_t(x,t^*)dx\right)^2 + \beta^2\right).$$

Proof. (1.2) yields

$$E(y(\cdot,0)) \leq c^2\,E(y(\cdot,t^*)).$$

Then, by (4.2)–(4.3),

$$E^{1/2}(y(\cdot,0)) \leq c\,\text{meas}^{1/2}\{\Omega\}\ \operatorname*{ess\,inf}_{x\in r(y,t^*)}(\mid \nabla y(x,t^*)\mid^2 \ + y_t^2(x,t^*) + \beta^2)^{1/2} \tag{4.9}$$

$$\leq c\,\text{meas}^{1/2}\{\Omega\}\ \operatorname*{ess\,inf}_{x\in r(y,t^*)}\left(\sum_{p=1}^n \mid y_{x_p}(x,t^*)\mid + \mid y_t(x,t^*)\mid + \beta\right).$$

From (4.9) we immediately obtain the needed result. □

5. Proof of Theorem 2.6. We stress that the constraint (2.7) is not involved in the further construction procedure.

Step 1. Let Y be the set of all the solutions of the system (1.1)–(1.2). Select an arbitrary monotone sequence $\{\delta_j\}_{j=1}^{\infty}$, $\delta_j \to 0+$, $j \to \infty$. Specify next for each j an arbitrary δ_j-net $Y^{\delta_j} = \{y_{kj}\}_{k=1}^{\infty}$ in Y uniformly with respect to the energy norm. In other words, $\{Y^{\delta_j}\}_{j=1}^{\infty}$ is dense in $Y \subset C([0,T]; H_E)$: for any $y \in Y$ there exists a sequence of elements y_{kj} such that

$$E^{1/2}(y(\cdot,t) - y_{kj}(\cdot,t)) \le \delta_j, \quad \forall t \in [0,T], \quad j = 1,\dots, \quad k = k(y,j). \tag{5.1}$$

Step 2. Select $\beta > 0$. Take an arbitrary countable set of all distinct strictly monotone sequences $\{t_k^j\}_{k=1}^{\infty} \subset]0,T[$, $j = 1,\dots$, with the *only* limit point $t = T$, that is, $\lim_{k\to\infty} t_k^j = T$, $j = 1,\dots$. Let $[0,T[\ni t \to \Omega(t)$ be an arbitrary set-valued map of positive measure, continuous with respect to Lebesgue measure and such that (see (4.2))

$$\Omega(t_k^j) = r(y_{kj}, t_k^j), \quad k,j = 1,\dots. \tag{5.2}$$

Let $\{v_p^0\}_{p=1}^n$, v^1 be arbitrary functions of a sign-type from $C([0,T[; L^2(\Omega))$ such that (see (4.3))

$$v_p^0(x,t_k^j) = \begin{cases} +1, & x \in r_{p+}^0(y_{kj},t_k^j), \\ -1, & x \in r_{p-}^0(y_{kj},t_k^j), \\ 0, & \text{otherwise}, \end{cases} \qquad k,j = 1,\dots, \qquad p = 1,\dots,n, \tag{5.3a}$$

$$v^1(x,t_k^j) = \begin{cases} +1, & x \in r_{+}^1(y_{kj},t_k^j), \\ -1, & x \in r_{-}^1(y_{kj},t_k^j), \\ 0, & \text{otherwise}. \end{cases} \qquad k,j = 1,\dots, \tag{5.3b}$$

Note that, due to Remark 2.2, all the outputs of (1.1)–(1.2), (2.6), (5.2)–(5.3) are continuous on $[0,T[$. In the next two steps we show that the constructed observation (2.6), (5.2)–(5.3) satisfies the requirements of Theorem 2.6.

Step 3. This step is to show that the net $\{y_{kj}\}_{k,j=1}^{\infty}$ specified in Step 1 generates via (2.6), (5.2)–(5.3) a certain "pointwise" net along the sequence $\{t_k^j\}_{k,j=1}^{\infty}$ in the set of all outputs, namely, in the sense of the relation (5.5a).

Take any $y \in Y$. Then, as was discussed in §4,

$$\min_{t\in[0,T]} \text{meas}\{r(y,t)\} = r(y) > 0. \tag{5.4}$$

Observe that for any $j = 1,\dots$, $k = k(y,j)$ such that (5.1) is fulfilled the following chain of estimates holds:

$$\left| \frac{1}{\text{meas}\{\Omega(t_k^j)\}} \int_{\Omega(t_k^j)} \cdot v_p^0(x,t_k^j) y_{x_p}(x,t_k^j)dx - \frac{1}{\text{meas}\{\Omega(t_k^j)\}} \int_{\Omega(t_k^j)} v_p^0(x,t_k^j) y_{kjx_p}(x,t_k^j)dx \right|$$

$$= \left| \frac{1}{\text{meas}\{r(y_{kj},t_k^j)\}} \int_{r(y_{kj},t_k^j)} v_p^0(x,t_k^j)(y_{x_p}(x,t_k^j) - y_{kjx_p}(x,t_k^j))dx \right|$$

$$\le \text{meas}^{-1/2}\{r(y_{kj},t_k^j)\}\ E^{1/2}(y(\cdot,t_k^j) - y_{kj}(\cdot,t_k^j))$$

$$\le \delta_j \,\text{meas}^{-1/2}\{r(y_{kj},t_k^j)\}, \quad p = 1,\dots,n.$$

In a similar way one can derive an analogous estimate for y_t. From these estimates, (5.2), (5.4), and Remark 4.1 it follows that for any given $y \in Y$, $\alpha > 0$ there exists an element y_{kj} such that simultaneously the following two estimates hold (compare with Assumption 2.1):

$$\| \mathbf{G}(t_k^j)\{\nabla y(\cdot, t_k^j) - \nabla y_{kj}(\cdot, t_k^j), y_t(\cdot, t_k^j) - y_{kjt}(\cdot, t_k^j)\} \|_{R^{n+1}} \leq \alpha, \tag{5.5a}$$

$$E^{1/2}(y(\cdot, t) - y_{kj}(\cdot, t)) \leq \alpha, \quad \forall t \in [0, T], \tag{5.5b}$$

where $\mathbf{G}(\cdot)$ stands for the observation operator in (2.6) under (5.2)–(5.3).

Step 4. Fix any $y \in Y$. Let $\{\alpha_i\}_{i=1}^{\infty}$ be an arbitrary sequence of positive numbers converging to zero and $\{y_{k_i j_i}\}_{i=1}^{\infty}$ be an associated sequence of solutions of (1.1)–(1.2) such that (5.5) is fulfilled for α_i, $i = 1, \ldots.$ Then (5.5b) and Lemma 4.2 imply

$$E^{1/2}(y(\cdot, 0)) \leq E^{1/2}(y_{k_i j_i}(\cdot, 0)) + \alpha_i \leq c \text{ meas}^{1/2}\{\Omega\} \sqrt{n+2} \tag{5.6}$$

$$\times(\beta^2 + \| \mathbf{G}(t_{k_i}^{j_i})\{\nabla y_{k_i j_i}(\cdot, t_{k_i}^{j_i}), y_{k_i j_i t}(\cdot, t_{k_i}^{j_i})\} \|_{R^{n+1}}^2)^{1/2} + \alpha_i, \quad i = 1, \ldots.$$

Combining (5.6) and (5.5a) yields

$$E^{1/2}(y(\cdot, 0))$$
$$\leq c \text{ meas}^{1/2}\{\Omega\} \sqrt{n+2}\ (\beta + \| \mathbf{G}(t_{k_i}^{j_i})\{\nabla y(\cdot, t_{k_i}^{j_i}), y_t(\cdot, t_{k_i}^{j_i})\} \|_{R^{n+1}} + \alpha_i) + \alpha_i,$$

and, further, with $i \to \infty$

(5.7)
$$E^{1/2}(y(\cdot, 0)) \leq c \text{ meas}^{1/2}\{\Omega\}\sqrt{n+2} \left(\beta + \sup_{t \in [0, T[} \| \mathbf{G}(t)\{\nabla y(\cdot, t), y_t(\cdot, t)\} \|_{R^{n+1}} \right),$$

which under (2.7) or (2.8) implies (1.4) with $\gamma = (c \text{ meas}^{1/2}\{\Omega\} \sqrt{n+2}\ (\beta + 1))^{-1}$. This completes the proof of Theorem 2.6 (and Corollary 2.7).

Remark 5.1.

(i) Note that the estimate (5.7) formally holds for any solution of (1.1)–(1.2), that is, even if the condition (2.7) (or (2.8)) is not verified.

(ii) In order to obtain exact observability in a prescribed subspace spanned by a finite number of solutions of (1.1)–(1.2) (as it may occur in applications), it suffices to specify a *finite* skeleton of type (5.2)–(5.3).

(iii) For the construction of suitable δ-nets in Step 1 one can use the Galerkin scheme (see also Remark 3.2).

(iv) The observation map defined by (5.2) becomes lower semicontinuous on $[0, T]$ if at $t = T$ it is of zero-measure.

6. Concluding remarks. The problem of exact observability under finitely many moving internal observations was discussed for the linear time-varying hyperbolic equation. Two types of observations linked sharply with the internal regularity of the solutions were considered, and the existence of observation curves and set-valued maps required for $L^\infty(0, T; R^{n+1})$- or $C([0, T[; R^{n+1})$-exact observability was established. The approach that was applied is related to the construction of suitable skeletons for the observations and deals with establishing a *certain countable net* in the pair of linear manifolds of the solutions and their associated outputs in the topology that is consistent with the well-posedness of the observations. It was shown that this can be achieved either a priori, when a given internal regularity implies the existence of a required net for any admissible observation operator (Assumption 2.1 and Theorems 2.1, 2.2, and 2.4, Corollary 2.3) or a posteriori, when, in order to ensure the existence of a suitable net in the set of the outputs, a particular observation operator has to be constructed (Theorem 2.6, Corollaries 2.7 and 2.8, and Theorem A.1).

Appendix A. Observe that the constant γ obtained in Theorem 2.6 (see (5.7)) is considerably "worse" than that in Theorem 2.1; see (3.8). We show now that one can improve the estimate (5.7), namely, get rid of the multiplier $\sqrt{n+2}$ if instead of (2.6) the following observation is employed:

$$\text{(A.1)} \qquad z(t) = \begin{pmatrix} \frac{1}{\text{meas}\{\Omega_1^0(t)\}} \int_{\Omega_1^0(t)} v_1^0(x,t)\; y_{x_1}(x,t)dx \\ \vdots \\ \frac{1}{\text{meas}\{\Omega_n^0(t)\}} \int_{\Omega_n^0(t)} v_n^0(x,t)\; y_{x_n}(x,t)dx \\ \frac{1}{\text{meas}\{\Omega^1(t)\}} \int_{\Omega^1(t)} v^1(x,t)\; y_t(x,t)dx \end{pmatrix}, \qquad t \in [0,T],$$

where the set-valued maps $\{\Omega_p^0(\cdot)\}_{p=1}^n$, $\Omega^1(\cdot)$ and functions $\{v_p^0\}_{p=1}^n$, v^1 are defined as in (2.6).

Let y be an arbitrary solution of the system (1.1)–(1.2) and $\beta > 0$ be given. Set for $\forall t \in [0,T]$:

(A.2a)

$$r_p^0(y,t) = \{x \in \Omega \mid \mid y_{x_p}(x,t) \mid^2 > \text{meas}^{-1}\{\Omega\} \parallel y_{x_p}(\cdot,t) \parallel^2_{L^2(\Omega)} -\beta^2\}, \quad p = 1,\dots,n,$$

$$\text{(A.2b)} \qquad r^1(y,t) = \{x \in \Omega \mid \mid y_t(x,t) \mid^2 > \text{meas}^{-1}\{\Omega\} \parallel y_t(\cdot,t) \parallel^2_{L^2(\Omega)} -\beta^2\}.$$

Let the sets $r_{p\pm}^0(y,t)$, $p = 1,\dots$, $r_\pm^1(y,t)$ be defined as in (4.3) with accordingly $r_p^0(y,t)$ and $r^1(y,t)$ substituted for $r(y,t)$.

THEOREM A.1. *Let* Y^{δ_j} *and* $\{t_k^j\}_{k=1}^\infty$, $j = 1,\dots$, *be defined as in Steps* 1–2 *in* §5 *and* $[0,T[\ \ni t \to \Omega_p^0(t)$ $(p = 1,,\dots,n)$, $\Omega^1(t)$ *be arbitrary set-valued maps of positive measure, continuous with respect to Lebesgue measure and such that* (*due to* (A.2)):

$$\text{(A.3)} \qquad \Omega_p^0(t_k^j) = r_p^0(y_{kj},t_k^j), \quad p = 1,\dots,n, \quad \Omega^1(t_k^j) = r^1(y_{kj},t_k^j), \quad k,j = 1,\dots .$$

Let $\{v_p^0\}_{p=1}^n$, v^1 *be arbitrary functions of a sign-type from* $C([0,T[;L^2(\Omega))$ *satisfying* (5.3) *under* (A.2). *Then for any solution of the system* (1.1)–(1.2), (A.1)–(A.3) *the following estimate is verified:*

(A.4)

$$E^{1/2}(y(\cdot,0)) \le c\,\text{meas}^{1/2}\{\Omega\} \left(\sqrt{n+1}\beta + \sup_{t\in[0,T[} \parallel \mathbf{G}(t)\{\nabla y(\cdot,t), y_t(\cdot,t)\} \parallel_{R^{n+1}} \right),$$

where $\mathbf{G}(\cdot)$ *is due to* (A.1)–(A.3).

Proof. The proof of this theorem follows the lines of §§4 and 5 with only one exception. Namely, the estimate (4.8) is replaced by

(A.5)

$$E(y(\cdot,0)) \le c^2\,\text{meas}\{\Omega\} \left(\sum_{p=1}^n \left(\text{meas}^{-1}\{\Omega_p^0(t^*)\} \int_{\Omega_p^0(t^*)} v_p^0(x,t^*) y_{x_p}(x,t^*)dx \right)^2 \right.$$

$$+\left(\operatorname{meas}^{-1}\{\Omega^1(t^*)\}\int\limits_{\Omega^1(t^*)} v^1(x,t^*)y_t(x,t^*)\mid dx\right)^2 + (n+1)\,\beta^2\Bigg).$$

Indeed, instead of the chain (4.9) we have now

$$E^{1/2}(y(\cdot,0)) \leq c\,\operatorname{meas}^{1/2}\{\Omega\}$$
$$\cdot\left(\sum_{p=1}^{n} \operatorname*{ess\,inf}_{x\in r_p^0(y,t^*)} \mid y_{x_p}(x,t^*)\mid^2 + \operatorname*{ess\,inf}_{x\in r^1(y,t^*)} |y_t(x,t^*)|^2 + (n+1)\beta^2\right)^{1/2},$$

from which we immediately obtain (A.5) and, eventually, the assertion of Theorem A.1. □

REFERENCES

[1] J.-P. AUBIN, H. FRANKOWSKA, *Set-Valued Analysis,* Birkhäuser, Cambridge, MA, 1990.

[2] SZ. DOLECKI AND D. L. RUSSELL, *A general theory of observation and control*, SIAM J. Control Optim., 15 (1977), pp. 185–219.

[3] A. EL JAI, A. BELFEKIH, AND J. BOUYAGHROUMNI, *Observability and sensors for hyperbolic systems*, Internat. J. Systems Sci., 22 (1991), pp. 1255–1265.

[4] A. F. FILIPPOV, *On certain questions in the theory of optimal control*, SIAM J. Control, 1 (1962), pp. 76–84.

[5] L. F. HO, *Exact controllability of the one-dimensional wave equation with locally distributed control*, SIAM J. Control Optim., 28 (1990), pp. 733–748.

[6] A. YU. KHAPALOV, *Observability of hyperbolic systems with interior moving sensors*, in Lecture Notes in Control and Inform. Sci. 185, R. F. Curtain, A. Bensoussan, and J. L. Lions, eds., Springer-Verlag, Berlin, Heidelberg, and New York, 1993, pp. 489–499.

[7] ———, *Controllability of the wave equation with moving point control*, Appl. Math. Optim., 31 (1995), pp. 155–175.

[8] ———, *Pointwise control of the wave equation*, TRECE 93.002, Oregon State University, Corvallis, Oregon, 1993.

[9] O. H. LADYZHENSKAYA, *The Boundary Value Problems of Mathematical Physics,* Springer-Verlag, New York, 1985.

[10] I. LASIECKA, *Controllability of a viscoelastic Kirchhoff plate*, Internat. Ser. Numer. Math. 91, Birkhäuser, Basel, 1989, pp. 237–247.

[11] I. LASIECKA AND R. TRIGGIANI, *Exact controllability of the wave equation with Neumann boundary control*, Appl. Math. Optim., 19 (1989), pp. 243–290.

[12] G. LEUGERING, *Exact boundary controllability of an integro-differential equation*, Appl. Math. Optim., 15 (1987), pp. 223–250.

[13] J.-L. LIONS, *Exact controllability, stabilization and perturbations for distributed systems*, John Von Neumann Lecture, Boston, MA, July, 1986; SIAM Rev., 30 (1988), pp. 1–68.

[14] ———, *Contrôlabilité Exacte et Stabilisation de Systémes Distribués,* I, *Contrôlabilité Exacte*, Masson, Paris, 1988.

[15] J.-L. LIONS AND E. MAGENES, *Non-homogeneous Boundary Value Problems and Applications*, I–III, Springer-Verlag, New York, 1972.

[16] Y. MEYER, *Etude d'un modéle mathématique issu du contrôle des structures spatiales déformables*, in Non Linear Partial Differential Equations and Their Applications, Collège de France Seminar, Vol. VII, H. Brezis and J. L. Lions, eds., Pitman, Boston, 1985, pp. 234–242.

[17] V. P. MIKHAILOV, *Partial Differential Equations,* Mir, Moscow, 1978.

[18] G. SCHMIDT AND N. WECK, *On the boundary behavior of solutions to elliptic and parabolic equations—with applications to boundary control for parabolic equations*, SIAM J. Control Optim., 16 (1978), pp. 593–598.

[19] T. I. SEIDMAN, *Observation and prediction for the heat equation* II, J. Math. Anal. Appl., 38 (1972), pp. 149–166.

[20] R. TRIGGIANI, *Exact boundary controllability on* $L_2(\Omega)\times H^{-1}(\Omega)$ *of the wave equation with Dirichlet boundary control acting on a portion of the boundary* $\partial\Omega$*, and related problems*, Appl. Math. Optim., 18 (1988), pp. 241–277.

[21] R. Triggiani, *Interior and boundary regularity of the wave equation with interior point control*, Differential Integral Equations, 6 (1993), pp. 111–129.

[22] R. Triggiani and D. Tataru, private communication.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 33, No. 4, pp. 1270–1276, July 1995

© 1995 Society for Industrial and Applied Mathematics
017

RENDEZVOUS SEARCH ON THE LINE WITH DISTINGUISHABLE PLAYERS*

STEVE ALPERN[†] AND SHMUEL GAL[‡]

Abstract. Two players are placed on the real line at a distance d with a distribution F known to both. Neither knows the direction of the other, nor do they have a common notion of a positive direction on the line. We seek the least expected *rendezvous time* $R = R(F)$ in which they can meet, given maximum speeds of one. We consider the cases where F is a bounded, point, discrete, or finite mean distribution. We obtain upper bounds or exact values for R and in one case an optimality condition for search strategies. A connection with Beck's linear search problem is established.

Key words. rendezvous, search

AMS subject classifications. 90B40, 90D05

1. Introduction.

We consider the problem faced by two players placed randomly on the real line, who can move at unit speed and wish to meet as soon as possible. They know the probability distribution of the distance between them, but neither knows the direction of the other. They are pointed in a random direction when placed, so they have no common notion of a positive direction on the line. We consider what we call the case of distinguishable players, which means that we allow them to use different strategies. The interpretation is that they have previously agreed which of the two roles each will take prior to the start of play. This corresponds to the *asymmetric rendezvous problem* which was defined for general spaces in a recent paper [1] by the first author. (That paper deals mainly with the *symmetric rendezvous problem,* where the players must adopt a common mixed strategy.)

The strategy space for both players is the set of speed one paths

$$P = \left\{p : \Re^+ \to \Re,\ p(0) = 0,\ |p(s) - p(t)| \leq |s - t|\right\}.$$

A player placed at a point x who chooses strategy p will have the time paths $x \pm p(t)$ equiprobably, depending on which way he is initially pointed. Without loss of generality we may fix a coordinate system where player I starts at the point 0 and player II starts equiprobably at $\pm d$, where the initial distance d between the players is drawn from the known cumulative probability distribution F. We assume that they meet in the first moment they occupy the same point. (Similar results can be obtained if the time of meeting is the first moment when their distance is smaller than some detection distance r, where r is small.) If I chooses $g \in P$ and II chooses $h \in P$, then the expected meeting time T is given by

$$T = T(g,h) = \int_0^\infty \frac{1}{4} \sum_{i,j=\pm 1} \min\{t : g(t) = id + j\,h(t)\}\ dF(d). \tag{1}$$

The integrand in the formula above represents the expected meeting time given that the initial distance is d, considering the uncertainty about whether II is placed to the right ($i = 1$) and whether II is pointed to the right ($j = 1$). For this problem we seek

* Received by the editors December 15, 1993; accepted for publication (in revised form) March 17, 1994.

[†] Mathematics Department, London School of Economics, London WC2A 2AE, England.

[‡] Statistics Department, University of Haifa, Haifa, Israel.

the *asymmetric rendezvous value*

$$R = R(F) = \min_{g,h \in P} T(g,h) \tag{2}$$

or bounds on this value and, if possible, the strategies for which the minimum is achieved. This strategy pair can be considered to be a Nash equilibrium of a game where the players have identical interests, although we tend to see rendezvous as a team problem rather than a game. In the symmetric version of rendezvous on the line, considered in [1], it was shown why mixed strategies were required to achieve the symmetric rendezvous value R^s. However, in the asymmetric version considered here it is clear that the minimum is achieved for pure strategies.

The paper is organized as follows. In §2 we consider bounded distributions F and obtain an upper bound $R(F) \leq 9D/8 + \mu/2$ in terms of the mean μ and maximum D of F. This compares favorably with the bound $R^s(F) \leq 2D + \mu/2$ obtained for the symmetric rendezvous value in [1]. In §3 we consider the case where F is a point distribution; that is, where the initial distance d between the players is known. In this case we show that $R = 13d/8$, which in fact agrees with the general upper bound obtained in the previous section. In §4 we consider discrete distributions F, where the initial distance between the players can take only countably many values. In this case we obtain a simple optimality condition. In the case of a finite distribution, this condition reduces the search for the optimal strategy pair to a finite problem. Finally, in §5 we consider the relation between our problem and the celebrated linear search problem introduced by Bellman [3] and extensively studied by Anatole Beck (whose latest is [2]) and others (including the second author [4]). We compare the rendezvous problem to the problem faced by a single searcher starting at 0 who wants to minimize the expected time $L = L(F)$ taken to find an object at distance d with distribution F, which is located equiprobably in either direction. We show that $L/2 \leq R \leq L$, for any distibution F.

2. Bounded distributions. In this section we assume that the initial distance between the players is bounded above by some least number D for which $F(D) = 1$. For such distributions there is a simple strategy pair $(\hat{g}, \hat{h})$ (depending on D) which guarantees rendezvous for all distributions with $F(D) = 1$ and gives a uniform upper bound for all such F in terms of D and the mean $\mu = \int_0^D x \, dF(x)$. As we shall see in the next section, this bound is sometimes equal to the rendezvous value R.

The strategies referred to above are defined by the formulae

$$\hat{g}(t) = \begin{cases} t & \text{if } 0 \leq t \leq D, \\ 2D - t & \text{if } D \leq t \leq 3D, \end{cases} \qquad \hat{h}(t) = \begin{cases} t & \text{if } 0 \leq t \leq D/2, \\ D - t & \text{if } D/2 \leq t \leq 2D, \\ t - 3D & \text{if } 2D \leq t \leq 3D. \end{cases}$$

The graphs of $\hat{g}(t)$ and the four paths $\pm d \pm \hat{h}(t)$ are drawn in Fig. 1 for $d = 2D/3$. It is now easy to establish the following.

THEOREM 2.1. *Let F be any distribution with mean μ and maximum D (that is, $F(D) = 1$). Then the rendezvous value $R = R(F)$ satisfies $R \leq (4\mu + 9D)/8$.*

Proof. It is straightforward to verify, from Fig. 1 or the definitions of $\hat{g}$ and $\hat{h}$, that the meeting times forming the integrand of 1 are given by

$$\min\left\{t : \hat{g}(t) = id + j\,\hat{h}(t)\right\} = \begin{cases} d/2 & \text{if } i = 1, j = -1, \\ (D+d)/2 & \text{if } i = 1, j = 1, \\ D + (D+d)/2 & \text{if } i = -1, j = -1, \\ 2D + (D+d)/2 & \text{if } i = -1, j = 1. \end{cases}$$

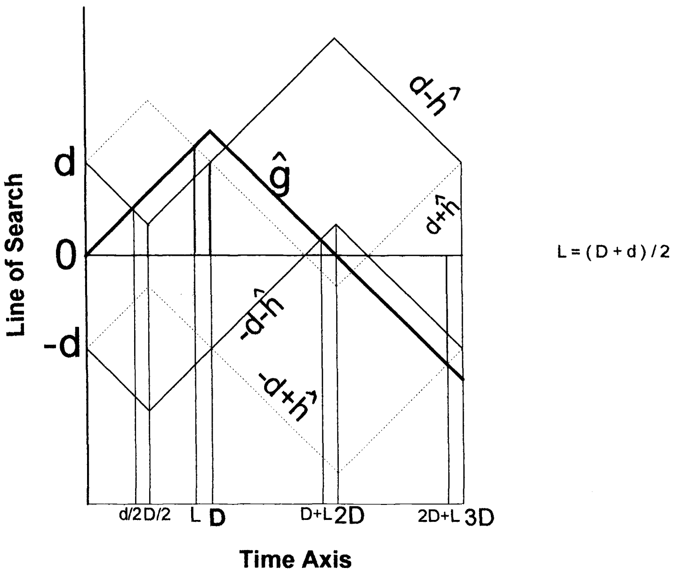


FIG. 1. *Graphs of* $\hat{g}$, $\pm d \pm \hat{h}$, *and their intersections.*

It then follows from (1) that the expected meeting time is given by

$$\begin{aligned}T\left(\hat{g},\hat{h}\right) &= \frac{1}{4}\int_0^D \left(\frac{x}{2}\right) + \left(\frac{D+x}{2}\right) + \left(D + \frac{D+x}{2}\right) + \left(2D + \frac{D+x}{2}\right)\, dF\left(x\right)\\ &= \frac{1}{8}\int_0^D \left(4x + 9D\right)\, dF\left(x\right) = \frac{1}{8}\left(4\int_0^D x\, dF\left(x\right) + 9D\right)\\ &= \frac{4\mu + 9D}{8}.\end{aligned}$$

The theorem then follows from the definition of the rendezvous value R. □

The strategy pair $(\hat{g}, \hat{h})$ seems to be effective when μ/D is close to one (indeed it is optimal at one; see Theorem 3.2). However, when $\mu/D < \frac{1}{4}$, the better estimate of $R \leq \mu + D$ is obtained by having one player stationary and the other searching D in one direction and then $2D$ in the other. Actually we can obtain a better bound than $\mu + D$, as shown in the next theorem.

THEOREM 2.1′. $R \leq \frac{3}{4}\mu + D$.

Proof. Assume that player I uses $\hat{g}$ and that player II rests until time D and then uses $\hat{h}$ (for $t \geq D$). If the initial distance is x, then the meeting time will be x with probability $\frac{1}{2}$, $D + (D + x)/2$ with probability $\frac{1}{4}$, and $2D + (D + x)/2$ with probability $\frac{1}{4}$. This gives an expected meeting time of $3x/4 + D$. To compute T we integrate with respect to the initial distance x, giving $T = 3\mu/4 + D$. □

Note that the bound given in Theorem 2.1′ is better than the bound given in Theorem 2.1 if and only if $\mu/D < \frac{1}{2}$. (For the uniform distribution these two strategy pairs yield the same expected meeting time $T = \frac{11}{8}$.)

3. Point distributions (*d* known). We now consider a special case of rendezvous on the line, where the players know the distance d to the other player but not the direction. We consider, as before, that player I starts at 0 and follows the path $g\left(t\right)$. However, we now take the view that there are four agents of player II, following the linked paths $\pm d \pm h\left(t\right)$. The time T is now simply the average of the times $t_1 \leq$

$t_2 \leq t_3 \leq t_4$ taken by I to meet the four agents of II. To obtain a lower bound on $T(g,h)$ for all g and h in P, we derive lower bounds on the numbers $t_i(g,h)$.

LEMMA 3.1. *If the initial distance between the players is d, then for all $g, h \in P$ we have that*

$$\begin{aligned} t_1 &\geq d/2, \\ t_2 &\geq d, \\ t_3 &\geq 2d, \\ t_4 &\geq 3d, \end{aligned}$$

and so $R \geq T(g,h) = (t_1 + t_2 + t_3 + t_4)/4 \geq (d/2 + d + 2d + 3d)/4 = 13d/8.$

Proof. Fix arbitrary strategies $g, h \in P$, and label the four agent II paths $\pm d \pm h$ as $L_i(t)$, indexed so that the meeting times t_i defined by $\min\{t : g(t) = L_i(t)\}$ are nondecreasing. By symmetry considerations we may assume that $L_1 = d - h$. The lower bound on t_1 is obvious, the relevant estimate being

$$L_1(t) - g(t) \geq d - (g(t) - h(t)) \geq d - 2t,$$

which is positive for $t \leq d/2$. The remainder of the proof splits into two cases depending on whether or not $L_1(0) = L_2(0)$.

First suppose that the condition is true; that is, $L_2 = d + h$. Since in this case we have $L_2(t_1) = d + h(t_1) = d + (d - g(t_1)) = 2d - g(t_1)$ and $L_2(t_2) = d + h(t_2) = g(t_2)$, we have

$$(3) \qquad 2d - g(t_1) - g(t_2) = h(t_1) - h(t_2) \leq t_2 - t_1.$$

Solving for t_2 and using the fact that $t_i \geq g(t_i)$, we have $t_2 \geq d$ because

$$(4) \qquad t_2 \geq 2d - g(t_2) + (t_1 - g(t_1)) \geq 2d - g(t_2).$$

We now consider how long after time t_2 the path $g(t)$ can intersect the nearer of the two agent paths which started at $-d$; that is, $L_3 = -d - h$. (The estimates for the case $L_3 = -d + h$ are higher and will not be given here.) Assuming that the intersection takes place at a point $g(t_3) \leq 0$ (otherwise both t_3 and t_4 will be larger than the estimates given here), we have that $t_3 - t_2 \geq g(t_2) - g(t_3)$, or

$$(5) \qquad t_3 \geq t_2 + g(t_2) - g(t_3) \geq 2d - g(t_3) \geq 2d,$$

using the result that $t_2 + g(t_2) \geq 2d$ from (4). The last estimate required for this case is for the time t_4 of the meeting of $g(t)$ with the last remaining player II agent, $L_4 = -d + h$. To this end observe that

$$t_4 - t_3 \geq |g(t_3) - L_4(t_3)|/2 = -h(t_3) = d + g(t_3), \text{ so}$$

$$t_4 \geq d + (t_3 + g(t_3)) \geq d + 2d = 3d, \text{ by (5)}.$$

We now consider the remaining case where $L_1(0) \neq L_2(0)$. With our choice of $L_1 = d - h$, this means that $L_2 = -d \pm h$. To obtain the required estimate on t_2, consider the speed one path $f(t)$ which equals L_1 until t_1 and then follows g until t_2. Since $|f(0) - L_2(0)| = 2d$, and $f(t_2) = L_2(t_2)$, we have $t_2 \geq d$. Since both f and L_2 can reach $g(t_2)$ in time t_2, we have

$$(6) \qquad t_2 \geq d + |g(t_2)|.$$

We first restrict ourselves to the case where $L_2 = -d + h$, as the alternative leads to larger estimates. Since $g(t_2) = L_2(t_2) = -d+h(t_2)$, we have $h(t_2) = g(t_2)+d$. Hence, regardless of the indexing, we have $|L_3(t_2)| = |L_4(t_2)| = 2d+|g(t_2)|$, $g(t_2)$ is between $L_3(t_2)$ and $L_4(t_2)$, and the latter two paths must move in opposite directions . Hence by symmetry we may assume without loss of generality that $g(t_2) \geq 0$ and that $L_3 = d+h$. Since $L_3(t_2)-g(t_2) = 2d$ and $L_3(t_3)-g(t_3) = 0$, it follows that $t_3- t_2 \geq d$, and hence $t_3 \geq 2d$. The final agent is $L_4 = -d-h$. Thus $L_3(t_2)-L_4(t_2) = 4d+2g(t_2)$. The path which follows L_3 from time t_2 to time t_3 and then follows g until time t_4 is a continuous path with speed bounded by one which goes from L_3 to L_4. Hence we have $t_4- t_2 \geq \frac{1}{2}[4d + 2g(t_2)]$ so that $t_4 \geq t_2 + 2d \geq 3d$ by (6). □

THEOREM 3.2. *If the distance d between the players is known, then the rendezvous value is given by $R = 13d/8$.*

Proof. By Lemma 3.1, $13d/8$ is a lower bound for R, and since $\mu = D = d$ in this case it is also an upper bound by Theorem 2.1. □

4. Discrete distributions. We now consider distributions supported on a countable set of points $0 \leq x_1 \leq x_2 \leq \ldots$, where the probability that the initial distance is x_k is p_k. For any strategy h of player II we may label the possible paths of II, the paths of the form $\pm x_k \pm h(t)$, as $L_m(t)$, $m = 1, 2, \ldots$. We call $L_m(t)$ the path of the mth agent of player II. Given a strategy g for player I such that $T(g,h) < \infty$, we may further assume that the player II paths are numbered so that player I first meets $L_m(t)$ at location y_m at time t_m, where $t_1 \leq t_2 \leq \ldots$. It turns out that any optimal strategy pair (g,h) must have a very specific and simple behaviour on each of the time intervals $[t_m, t_{m+1}]$.

THEOREM 4.1. *Suppose (g,h) is an optimal strategy pair. For $m = 1, 2, \ldots$ we have*

$$|L_{m+1}(t_m) - g(t_m)| = 2|t_{m+1} - t_m| \text{ and hence}$$

$$y_{m+1} = \frac{1}{2}g(t_m) + \frac{1}{2}L_{m+1}(t_m).$$

In other words player I *and agent $m+1$ approach each other at speed one as soon as agent m has been met.*

Proof. Suppose the first condition of the theorem is not satisfied for some times $t_m < t_{m+1}$. In this case $|L_{m+1}(t_m) - g(t_m)| < 2|t_{m+1} - t_m|$. Then change g and h to new strategies $\tilde{g}$ and $\tilde{h}$ by modifying them *only on the closed interval* $[t_m, t_{m+1}]$ so that I and L_{m+1} move at speed one to the midpoint $\frac{1}{2}g(t_m) + \frac{1}{2}L_{m+1}(t_m) = \tilde{g}(\tilde{t}_{m+1}) = \tilde{L}_{m+1}(\tilde{t}_{m+1})$, reaching it at time $\tilde{t}_{m+1} = t_m + |L_{m+1}(t_m) - g(t_m)|/2 < t_{m+1}$, and return to their original positions $g(t_{m+1})$ and $L_{m+1}(t_{m+1})$ at time t_{m+1}. (This is possible because, under the original strategy, at least one of the players reaches the midpoint at time $t \geq \tilde{t}_{m+1}$.) We will then have that $\tilde{t}_n \leq t_n$ for all n and that $\tilde{t}_{m+1} < t_{m+1}$. It follows that

$$T\left(\tilde{g}, \tilde{h}\right) = \sum_n p_n \tilde{t}_n < \sum_n p_n t_n = T(g,h). \quad □$$

In the case where there is a finite distribution concentrated on K points, the above condition reduces the search for an optimal strategy to a finite problem. Each ordering of the $4K$ agents of player II leads to at most one possible optimal strategy pair, and hence one of these strategy pairs must be optimal. Of course one need not check all $(4K)!$ orderings because some (most) are impossible. For example, if $x_i < x_j$, then I cannot meet any agent $\pm x_j \pm h(t)$ before the corresponding agent (i.e., with the same choice of signs) $\pm x_i \pm h(t)$. This means that agents with the same pair of signs must be met in increasing values of x_j. This reduces the number of cases to be checked to $(4K)!/K!^4$. This number can then be divided by 4 if we assume without loss of generality that the first agent to be met is $x_1 + h(t)$. Thus when $K = 1$ only six cases need be checked, among which two cases are dominated and the group of the other four can be split into two groups of (symmetric) equivalent cases ($h(t)$ replaced by $-h(t)$ by the third and fourth agents). Thus we actually have to check just two cases corresponding to the two optimal solutions. (Note that in Theorem 2.1 we actually have another solution with $g(t)$ and $h(t)$ interchanged.) Indeed this provides an easy alternative proof of Lemma 3.1 concerning a known initial distance.

5. Relation to the linear search problem. The problem of this paper is related to the following symmetric form of the linear search problem. A searcher with speed one and initial position 0 seeks to find an object hidden at distance d drawn from a distribution F and placed equiprobably at $\pm d$. The least expected time to find the object is denoted by $L = L(F)$. It can easily be shown that by using a geometric search pattern of, say, doubling the successive searches to the right and left, L is finite if and only if F has a finite mean. A similar approach will also give the same result for the rendezvous value R, but we prefer to employ a comparison of $R(F)$ and $L(F)$. Since $L(F)$ has been extensively studied, our comparison immediately extends those results to $R(F)$.

To begin this analysis, we first consider a variant of the rendezvous problem defined in the introduction. Instead of assuming that the players are placed on the line facing in an equiprobable direction, we will assume both are placed facing in the same direction. Alternatively, we are assuming that they have a common notion of a positive direction on the line. In this case the expected meeting time corresponding to a strategy pair $g, h \in P$ is given by

$$T^* = T^*(g,h) = \int_0^\infty \frac{1}{2} \sum_{i=\pm 1} \min\{t : g(t) = id + h(t)\}\, dF(d), \tag{7}$$

and the corresponding rendezvous value is given by

$$R^* = R^*(F) = \min_{g,h \in P} T^*(g,h). \tag{8}$$

LEMMA 5.1. *$R^* = L(F)/2$, and the minimizing $g, h \in P$ of* (8) *satisfy $g = -h$ and $|g'| \equiv 1$.*

Proof. First observe from (7) that T^* depends only on the difference $f = g - h$, which belongs to the space $2P$ of functions with Lipschitz constant (maximum speed) 2, and $f(0) = 0$. Any such function f has a derivative almost everywhere, and we may define the total distance travelled by f up to time t by $\theta(t) \equiv \int_0^t |f'(t)|\, dt \leq 2t$. Then the function $\hat{f}$ defined by the equation $\hat{f}(\theta(t)/2) = f(t)$ satisfies $|\hat{f}'(t)| \equiv 2$ almost everywhere and reaches every point on the line not later than f does. Thus there is always a minimizing function f with $|f'(t)| \equiv 2$ which therefore must be the difference of functions g and h as stated in the lemma. Furthermore we have from (8) that

$$R^* = \min_{f \in 2P} \int_0^\infty \frac{1}{2} \sum_{i=\pm 1} \min\{t : f(t) = id\}\, dF(d) = L(F)/2. \qquad \square$$

THEOREM 5.2. *For any distribution F, we have*

$$L(F)/2 \leq R(F) \leq L(F).$$

Proof. In the version of rendezvous search where the players have a common notion of direction, they could choose to ignore this information and thus play the version of the game with no common direction. Thus $R^* \leq R$, and hence $L/2 \leq R$ by the previous lemma. To obtain the right inequality, simply restrict the game to strategy pairs where one the of players doesn't move, so that the other player is faced with the linear search problem.. □

As an example, consider the point distribution where the distance between the players is 1. Here we found that $R = \frac{13}{8}$. It is easily seen that in this case $L = (1+3)/2 = 2$, and the inequalities of the theorem are strict in that $1 < \frac{13}{8} < 2$.

REFERENCES

[1] S. ALPERN, *The rendezvous search problem*, SIAM J. Control Optim., 33 (1995), pp. 673–683.

[2] A. BECK AND M. BECK, *The revenge of the linear search problem*, SIAM J. Control Optim., 30 (1992), pp. 112–122.

[3] R. BELLMAN, *An optimal search problem*, SIAM Rev., 5 (1963), p. 274.

[4] S. GAL, *Search Games, Part* II, Academic Press, San Diego, 1980.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 33, No. 4, p. 1277, July 1995

© 1995 Society for Industrial and Applied Mathematics

018

ERRATUM
Observability and Observers for Nonlinear Systems*

J.-P. GAUTHIER† AND I. A. K. KUPKA‡

On page 991, assumption (A2),

—(A_2) each of the maps

$$\frac{\partial F^i}{\partial x^{i+1}}, \quad 0 \leq i \leq d-2,$$

is globally Lipschitz with respect to $\underline{x^i}$ uniformly with respect to u and x^{i+1} (denoting $\underline{x^i} = (x^0, \ldots, x^i)$) in (C),

should be replaced by the following:

—(A_2) each of the maps F^i, $0 \leq i \leq d-1$, is globally Lipschitz with respect to $\underline{x^i}$ uniformly with respect to u and x^{i+1} (denoting $\underline{x^i} = (x^0, \ldots, x^i)$) in (C).

* Received by the editors October 15, 1994; accepted for publication November 18, 1994. SIAM J. Control Optim., 32 (1994), pp. 975–994.

† Institut Universitaire de France, Departement de Mathématiques, Institut National des Sciences Appliquées de Rouen, Unité de Recherche Associée 1378 de Centre National de la Recherche Scientifique, B.P. 08, Place Emile Blondel, 76131 Mont Saint-Aignan cedex, France.

‡ Department of Mathematics, University of Toronto, 100 St. George Street, Toronto, Ontario M5S 1A1, Canada.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 33, No. 5, pp. 1279–1311, September 1995

© 1995 Society for Industrial and Applied Mathematics
001

TOWARD A GEOMETRIC THEORY IN THE TIME-MINIMAL CONTROL OF CHEMICAL BATCH REACTORS*

B. BONNARD† AND J. DE MORANT‡

Abstract. In this article we outline a geometric theory for the time-minimal control of chemical batch reactors by analyzing the equations from Pontryagin's maximum principle applied to the optimal control problem. This theory is used for computing the optimal feedback law for a batch reactor in which three species X, Y, Z are reacting according to the scheme $X \longrightarrow Y \longrightarrow Z$ and every reaction in the sequence obeys first-order kinetics. The control variable is the derivative of the temperature in the reactor, and the terminal condition is a specified ratio of concentrations of species X and Y.

Key words. time-optimal control, optimal synthesis, chemical systems

AMS subject classifications. 49B10, 93C10

1. Introduction. The choice of best temperature schedule in a batch reactor to maximize the yield per year is one of the main problems in chemical engineering. Until now, in practise, *the reaction temperature is held constant* over the duration of a batch and hence an optimal law is computed among all the constant temperatures. Clearly, by varying the temperature of the reactions we may improve the yield. Therefore, many researchers have recently concentrated their efforts on the optimisation of chemical reactors; see for instance [6], [16]. Their studies are *mainly numerical* and based on *Pontryagin's maximum principle* (PMP) or the *Hamilton–Jacobi–Bellman equation* (HJB). If PMP is used, the optimal law is computed as an open-loop function. Moreover only the "classic" optimal control developed in the 1960s presented in [6] is used. On the other hand, the HJB approach will not lead to an optimal control, even for a small-dimensional state space, for reasons discussed in [16].

In this article, we outline a geometric theory for the time-minimal control of a chemical reactor based on the *analysis of the equations coming from the maximum principle*. In order to be implemented, the optimal law is computed as a *feedback law* (closed-loop function). Our study is in the spirit of similar approaches developed by Sussmann and Tang [30] for *mechanical systems* and uses recent results outlining a *geometric theory* of solutions of the maximum principle and providing a methodology to analyze optimal control problems and solving time-optimal control problems of reasonable complexity. It must be noted that this methodology can be successfully applied to chemical systems because in many situations—although the reaction scheme is, in general, complicated—*it can be reduced to a few reactions*. Among the numerous contributions to this theory, we shall make an intensive use of the results from Schättler and Sussmann concerning the parametrizations of the *boundary of the* (*small time*) *accessibility set* and their applications to the construction of the optimal syntheses [26]–[28].

Indeed, this set plays the role of the unit ball in Riemaniann geometry, and its complexity explains the difficulty of solving optimal control problems. Other useful tools in our problem are the results concerning the classification of extremals by Kupka [18]. Moreover, the technique established by Bonnard and Kupka in [2] allows us to obtain evaluations of the accessibility set along a singular extremal and therefore to define the concept of *conjugate points* and to obtain C^0-optimality conditions along such a trajectory. This technique will be used to show the *existence of conjugate points for even simple reaction schemes* and hence indicates the nontriviality of the control problem. Now, of course all the available tools are

* Received by the editors December 9, 1992; accepted for publication (in revised form) March 1, 1994.

† Université de Bourgogne, Laboratoire de Topologie, B.P. 138, 21004 Dijon Cedex, France.

‡ Shell Recherche S.A., 76530 Grand Couronne, France.

not sufficient to solve our problem automatically. We have to develop the theory in different directions. Moreover the specific geometry of our problem has to be taken into account. *The analysis developed in this article has been used to implement a feedback optimal law in a* 5*-liter reactor in the department of chemical engineering at Institut National des Sciences Appliquées Rouen*; see [12], [23].

The main contributions of this article are the following. First, we have to develop the theory in the following direction. In the batch reactor problem, the terminal condition in the state space belongs to a hypersurface, which corresponds physically to a desired repartition of the concentrations at the end of the batch. Therefore we initialize a *generic classification of the optimal feedback law* for the time-optimal control problem for *planar systems*, when *the terminal set is a manifold of codimension one*. This classification is similar to the one by Sussmann for the fixed end-point problem [27], although the techniques to analyze the problem are different. Then, using [2], we develop an algorithm to determine if a reference singular extremal is optimal, for the fixed end-point problem, with respect to all trajectories contained in a (C^0-small) neighborhood of the given trajectory. This algorithm is in fact a *method for computing the conjugate points* along a singular trajectory, under generic conditions, *without any integration*. It can be modified in order to deal with the optimal problem regardless of the terminal condition (concept of *focal point*). Finally all these results are applied to solve the time-optimal control problem for a scheme of two consecutive reactions.

This article is organized as follows. In §2, we give the mathematical model and describe the optimal control problem. In §3, we briefly recall the maximum principle and some basic properties of its solutions. In §4, we outline a classification of optimal feedback laws near the terminal set. In §5, we give an algorithm to compute the conjugate points along a singular trajectory under generic and codimension-one conditions. It is presented for a system in $\mathbf{R}^3$ but can be straightforwardly extended to $\mathbf{R}^n$. In §6, we patch together all these results to compute the optimal feedback law for a batch reactor in which three species X, Y, Z are reacting according to the scheme $X \to Y \to Z$ and every reaction is irreversible and of first order, the target being a specified ratio of the concentrations of X and Y.

2. Mathematical model of a chemical batch reactor and description of the control problem.

2.1. Batch reactor. Batch reactors are used for the production of many chemicals, e.g., polymers and fine chemicals. The reactants are initially loaded into the reactor, mixed well, and allowed to react for a certain time. The ambient temperature is allowed to vary by a heat exchanger designed around the reactor. We assume that the cooling fluid has a constant flow rate. Its temperature is controlled with a regulator.

2.2. Mathematical model. The mathematical model for the process has to be divided into two distinct parts. First, we have a model describing the *chemical reactions* in the reactor, coming from experimental and physical laws of *chemical kinetics* [8]. Second, we have a model describing the *thermodynamic phenomena* obtained from standard laws and models coming from chemical thermodynamics [13] and heat exchanger models in chemical reaction engineering [23].

2.2.1. Chemical kinetics. We briefly recall some standard results from chemical kinetics (see [8] for more details).

Consider first a single chemical reaction (e.g., $2N0 + 2H_2 \to N_2 + 2H_20$):

$$\sum_{i=1}^{n_1} \alpha_i R_i \to \sum_{i=1}^{n_2} \beta_i P_i,$$

where the R_i's are called the *reactants* and the P_i's are called the *products*. The coefficients α_i, β_i are the *stoichiometric coefficients* given with the convention $\alpha_i > 0$ and $\beta_i < 0$, and since they are only defined up to a factor, one can set $\alpha_1 = 1$. Let X_i be a species R_i or P_i, with stoichiometric coefficient γ_i. Initially we have $n_i(0)$ moles of constituent X_i and $n_i(t)$ moles at time t. The *molar extent* of species X_i is given by

$$\zeta_i(t) = \frac{n_i(t) - n_i(0)}{\gamma_i},$$

and from the law of mass conservation all the ζ_i's are equal to a same number ζ called the *molar extent of the reaction.* If more than one chemical reaction is possible and ζ_k is the extent of X_i due to the kth reaction and γ_i^k stoichiometric coefficient of the species X_i in the kth reaction, the total change in the number of moles of species X_i because of p reactions is

$$n_i(t) - n_i(0) = \sum_{k=1}^{p} \gamma_i^k \zeta_k(t).$$

Now we have to model the kinetics of the reactions. For that, consider a single reaction between n species X_i with stoichiometric coefficient γ_i and assume that the reaction is at *constant volume* V. The rate of evolution of species X_i is

$$v_i = \frac{dn_i}{dt},$$

from the law of mass conservation we have

$$\gamma_j v_i = \gamma_i v_j,$$

and its *specific rate* is

$$r_i = \frac{v_i}{V} = \frac{1}{V}\frac{dn_i}{dt} = \frac{dc_i}{dt},$$

where $c_i = \frac{n_i}{V}$ is the *molar concentration* of species X_i. Let $r = r_i/\gamma_i$.

The rate r is given at constant temperature by an *empirical law* of Gulberg and Waage,

$$r = -k \prod_{i=1}^{n} c_i^{\delta_i}, \tag{2.1}$$

the reaction is said to be of the δ_i*th order* with respect to species X_i, and the overall order is $m = \sum_{i=1}^{n} \delta_i$. The numbers δ_i *are not generally related to the corresponding stoichiometric coefficients* γ_i *and have to be determined experimentally.*

The coefficient k depends on the temperature T of the reaction and is given by a *physical law* called *Arrhenius' law*,

$$k = A_r e^{-E_r/RT}, \tag{2.2}$$

where the parameters A_r and E_r are the *frequency factor* and the *activation energy*, respectively, of the reaction and R is the gas constant.

We can now model every reaction network occurring at constant volume. In this article we consider the case where $(n+1)$ species X_i whose concentration is c_i are reacting according to the scheme

$$X_1 \xrightarrow{k_1} X_2 \xrightarrow{k_2} \cdots \xrightarrow{k_n} X_n.$$

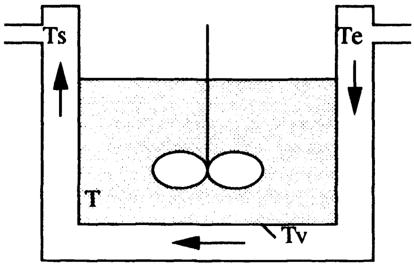


FIG. 1. *T = temperature in the reactor, T_e = temperature of the fluid at the entrance of the exchanger, T_s = temperature of the fluid at the exit, T_v = temperature of the inner shell.*

Moreover, let us assume that in this network of *consecutive* and *irreversible* reactions every reaction $X_i \overset{k_i}{\longrightarrow} X_{i+1}$ obeys first-order kinetics:

$$\frac{dc_i}{dt} = -k_i\, c_i,$$

where from (2.1), $k_i = A_i e^{-Ei/RT}$. Since we have

$$\sum_{i=1}^{n+1} c_i(0) = \sum_{i=1}^{n+1} c_i(t),$$

if we introduce $x^i = c_i / \sum_{i=1}^{n+1} c_i(0)$ and $v = k_1$, the reaction scheme is modelled by

$$\text{(2.3a)} \qquad \frac{dx}{dt} = K(v)x,$$

$$\text{(2.3b)} \qquad \frac{dv}{dt} = h(v)u,$$

where $x =^t (x^1, \ldots, x^n)$, $u = \frac{dT}{dt}, h(v) = \frac{R}{E_1} v \ln^2(v/A_1)$, K is the matrix

$$K = \begin{bmatrix} -k_1 & & \\ k_1 & 0 & \\ 0 & k_{n-1} & -k_n \end{bmatrix},$$

and x, v satisfy the inequalities

$$0 \leq x^i, \sum_{i=1}^{n} x^i \leq 1, \qquad 0 < v < A_1.$$

Moreover x^{n+1} can be computed using the law of mass conservation.

2.2.2. Heat exchanges. To get a complete mathematical model of the process represented by Fig. 1 and in which the chemical reactions are at constant volume, we have to model the heat exchanges. To analyze the optimal control problem and for reasons explained later, we don't need a mathematical model and indicate only the nature of the expected equations; see [23] for a complete description.

The equation describing the evolution of the temperature T in the reactor is of the form

$$\frac{dT}{dt} = f\left(T_v, T, \frac{dx}{dt}\right), \tag{2.4}$$

where T_v is the temperature of the shell containing the chemical species. This equation models two different properties. First, in every chemical reaction, there is a heat transfer Q, which can be positive (*exothermic reaction*) or negative (*endothermic reaction*). Hence the variation of temperature depends on the variation of concentrations, which measures the number of reactions. Second, the temperature T is modified by heat transfer with the inner shell of the reactor and, hence depends on T_v.

Now, there are relations describing the heat transfer between the inner shell whose temperature is T_v and the temperature T_f of the fluid in the jacket and the way of tuning the temperature at the entrance T_e, and the equations are of the form

$$g\left(T_v, \frac{dT_v}{dt}, \dots, T_e, \frac{dT_e}{dt}, \dots, T_f, \frac{dT_f}{dt}, \dots, u\right) = 0, \tag{2.5}$$

where $u(.)$ is the *physical control.* (The reactor is assumed to be isolated from the outside world.)

2.3. The optimal control problem. The optimal control problem is the following. Given the reaction scheme in the reactor and the *desired product*, *maximize the production of the reactor over a year.* This objective can be clearly translated into a time-optimal control problem as follows. *Fix a desired product quantity for a batch, and minimize the batch time.*

Now, for the optimal control problem, it is reasonable to assume that the control is T or $\frac{dT}{dt}$. This means that for the optimal problems, equations (2.4) and (2.5) are not taken into account and *the optimal law is computed for a given chemical reactions network and does not depend on the transfer devices.* Of course, to be implemented, *the optimal law has to be tracked* with a *regulator* (like a proportional integral-derivative feedback) whose parameters are determined using equations (2.4) and (2.5). (See [23] for such a study.) Moreover these equations are imposing the bounds u_- and u_+ (depending in general on the state variables) such that $u_- \leq \frac{dT}{dt} \leq u_+$.

The choice of T or $\frac{dT}{dt}$ as the control variable depends on the ability of the regulator to track the computed optimal law. If T can be chosen as the control variable, the bounds on T, like 20°C $\leq T \leq$ 90°C, imposed since the shell of the reactor is in glass (this can also impose constraints such as $|T - T_v| \leq M$) are then the bounds of the control variable and the optimal control problem can seem to be better posed. However, it is not true, since the control system is not linear with respect to the input and an *admissible time-optimal law may not exist.* (This is due to the well-known relaxation phenomenon; see [5] or [19].)

Hence, in this article, we shall assume that $\frac{dT}{dt}$ *is the control variable. Moreover we shall suppose that the bounds* u_- *and* u_+ *such that* $u_- \leq \frac{dT}{dt} \leq u_+$ *are fixed. The bounds on the state variable* T, $T_{\text{min}} \leq T \leq T_{\text{max}}$, *will not be taken into account.* (See [21] for a discussion of this problem.)

The optimal control problem denoted by $\mathcal{P}$ is, then, as follows: Consider system (2.3), the class of admissible controls $\mathcal{U}$ being the set of all measurable functions $u(.) = \frac{dT}{dt}(.)$ defined on $[0, t_u]$ and taking their values in the fixed interval $[u_-, u_+]$, with $u_- < 0 < u_+$. Let N be an analytic manifold in the space of concentrations $\{x\}$ and with codimension one, and let $m_0 = (x_0, v_0)$ be an initial condition in the (physical) state space. Among all the solutions of the system and starting from m_0, find the ones such that the time duration of control to reach the target is minimal. The manifold N is called the *terminal manifold.*

For practical reasons the optimal law has to be computed for each initial state m_0 to provide an optimal feedback law $m \rightarrow u^*(m)$ (*synthesis problem*). The implementation of this feedback law requires the reconstruction of the state using an *observer*. (In practise, we have only a few observations.) The observation and the estimation problems will not be studied in this article, although they are crucial in practise.

3. Pontryagin's maximum principle and some geometric properties of extremals. To make this article self-contained we shall summarize in this section some geometric properties of extremals which will be used extensively to analyze the optimal control problem.

3.1. Pontryagin's maximum principle. Consider a system of the form

$$\frac{dx}{dt}(t) = f(x(t), u(t)), \qquad x \in \mathbf{R}^n, \tag{3.1}$$

where f is an analytic mapping from $\mathbf{R}^n \times \mathbf{R}^m$ into $\mathbf{R}^n$ and the set of admissible controls $\mathcal{U}$ is the set of bounded measurable mappings $u(.)$ defined on an interval $[0, T_u]$ of $\mathbf{R}^+$ and taking their values in a subset Ω of $\mathbf{R}^m$. Let N be a regular analytic submanifold of $\mathbf{R}^n$. The PMP tells us that if $u^*(t), t \in [0, t^*]$, is an optimal control for the time-minimal control problem with terminal manifold N, then there exists an adjoint vector $p^*(t) \in \mathbf{R}^n \backslash \{0\}$, absolutely continuous (a.c.), such that the following equations are satisfied almost everywhere (a.e.) on $[0, t^*]$:

$$\frac{dx^*}{dt} = \frac{\partial H}{\partial p}(x^*, p^*, u^*), \qquad \frac{dp^*}{dt} = -\frac{\partial H}{\partial x}(x^*, p^*, u^*), \tag{3.2}$$

$$H(x^*, p^*, u^*) = \operatorname{Max} H_{u \in \Omega}(x^*, p^*, u), \tag{3.3}$$

where $H(x,p,u) = \langle p, f(x,u)\rangle$, $\langle , \rangle$ being the standard inner product in $\mathbf{R}^n$. Moreover

(3.4) the mapping $t \rightarrow \operatorname{Max} H_{u \in \Omega}(x^*(t), p^*(t), u)$ is constant everywhere and positive.

The vector p^* can be selected to satisfy the *transversality condition.*

(3.5) $p^*(t^*)$ orthogonal to $T_{x^*(t)}N$, where T_xN is the tangent space to N at x.

The system (3.2) is called the *Hamiltonian lift* of (3.1), and H is the *Hamiltonian*. A triple (x, p, u) solution of (3.2), (3.3), and (3.4) is called an *extremal*. (Sometimes we omit p.) An extremal (x,p,u) such that $H(x,p,u) = 0$ almost everywhere is called *exceptional*. Assume now that Ω is a convex polyhedron. An extremal (x,p,u) is called *regular* if, for almost every t, $u(t)$ lies on the vertices of Ω and (totally) *singular* if, for each t, $u(t)$ lies in the interior of Ω; hence $\frac{\partial H}{\partial u}(x,p,u) = 0$.

3.2. Singular extremals. We briefly recall some concepts and results concerning the singular extremals; see [1] and [2] for details.

By definition, a singular extremal (x,p,u) is a solution of (3.2), (3.3), and (3.4) contained in the variety $\frac{\partial H}{\partial u} = 0$, and from (3.3) it has to satisfy the *Legendre condition* $\partial^2 H/\partial u^2_{|(x,p,u)} \leq 0$. If for all t, this inequality is strict, it is called the *strong Legendre condition*. Then, by the implicit function theorem, the singular control u can be computed locally as a mapping $\hat{u} : (x,p) \rightarrow \hat{u}(x,p)$ by solving the equation $\frac{\partial H}{\partial u} = 0$. Consider now a *single-input affine* analytic system

$$\frac{dx}{dt} = X(x) + uY(x), \qquad x \in \mathbf{R}^n, \qquad u \in [-1, +1], \tag{3.6}$$

and let (x,p,u) be a singular extremal. The associated Hamiltonian is $H(x,p,u) = \langle p, X + uY\rangle$. Hence the equation $\frac{\partial H}{\partial u} = 0$ is equivalent to $\langle p, Y(x)\rangle = 0$, and the singular controls have to be computed as follows.

DEFINITION 3.1. *The Lie bracket of two vector fields* Z_1, Z_2 *is computed with the convention* $[Z_1, Z_2](x) = \frac{\partial Z_2}{\partial x}(x)Z_1(x) - \frac{\partial Z_1}{\partial x}(x)Z_2(x)$, *and let* ad Z_1 *be the mapping defined by* ad $Z_1(Z_2) = [Z_1, Z_2]$. *A point* (x,p) *is called* ordinary *if* $\langle p, \operatorname{ad}^2 Y(X)\rangle \neq 0$, *and let* θ *be the set of nonordinary points. A singular extremal* (x,p,u) *such that* $(x(t),p(t)) \in \boldsymbol{R}^{2n}\backslash\theta$ *is called of* order 2. *Let* Σ *be the variety* $\{(x,p);\ \langle p, Y(x)\rangle = 0\}$, *and let* $\Sigma' = \{(x,p);\ \langle p, Y(x)\rangle = \langle p, [X,Y](x)\rangle = 0\}$. *Let* $\hat{H}$ *be the restriction to* $\Sigma'\backslash\theta$ *of the mapping* $(x,p) \rightarrow \langle p, X(x) + \hat{u}(x,p)Y(x)\rangle$, *where*

$$\hat{u}(x,p) = \frac{\langle p, \operatorname{ad}^2 X(Y)(x)\rangle}{\langle p, \operatorname{ad}^2 Y(X)(x)\rangle}. \tag{3.7}$$

PROPOSITION 3.2. *The singular extremals* (x,p,u) *of order* 2 *are defined by*

(i) $u(t) = \hat{u}(x(t),p(t))$;

(ii) (x,p) *is a solution of*

$$\frac{dx}{dt} = \frac{\partial \hat{H}}{\partial p}(x,p), \qquad \frac{dp}{dt} = -\frac{\partial \hat{H}}{\partial x}(x,p); \tag{3.8}$$

and in order to be admissible they have to satisfy the constraint

(iii) $(x,p) \in \{(x,p);\ |\hat{u}(x,p)| \leq 1\}$.

DEFINITION 3.3. *Let* (x,p,u) *be a singular extremal of order* 2 *and h be the value of t* $\rightarrow$ $\operatorname{Max}_{u\in\Omega} H(x,p,u)$. *According to* 3.1, *it is called* exceptional *if* $h = 0$. *If* $h \neq 0$, *it is called* hyperbolic *if* $\langle p(t), \operatorname{ad}^2 Y(X)(x(t))\rangle < 0$ *and* elliptic *if* $\langle p(t), \operatorname{ad}^2 Y(X)(x(t))\rangle > 0$.

3.3. Time-optimality problem. Let (x,p,u) be a singular extremal of *order* 2, defined on $[0,T]$. Since the maximum principle is only a necessary condition for optimality, the main problem when the solutions of PMP are analyzed is to determine their optimality. We make the following assumptions.

Assumption 3.4. First, let us make the following assumptions:

(H0) $t \rightarrow x(t)$ is one to one. Then one may set $u \equiv 0$, and moreover, let us assume the following:

(H1) $\forall t \in [0,T]$, the $(n-1)$ vectors $\{\operatorname{ad}^k X(Y)(x(t)); k = 0,\dots,n-2\}$ are linearly independent.

(H2) $\forall t \in [0,T]$, $\operatorname{ad}^2 Y(X)(x(t)) \notin$ Span $\{\operatorname{ad}^k X(Y)(x(t)); k = 0,\dots,n-2\}$.

(H3) If $n = 2$, $X(x(t))$ and $Y(x(t))$ are linearly independent $\forall t \in [0,T]$, and if $n \geq 3$, $X(x(t)) \notin$ Span $\{\operatorname{ad}^k X(Y)(x(t)), k = 0,\dots,n-3\}\, \forall t \in [0,T]$.

THEOREM 3.5. *Let* (x,u) *be a trajectory defined on* $[0,T]$ *and satisfying* (H0) – (H3). *Then there exists a* C^o*-neighborhood U of x such that x is a time-minimizing* (*resp., maximizing*) *trajectory with respect to all solutions of* (3.6) *contained in U and joining* $x(0)$ *to* $x(T)$ *if* (x,u) *is a hyperbolic or exceptional* (*resp., elliptic*) *extremal and* $T < t_{1c}$, *where* t_{1c} *is the first conjugate time along* x.

In [2], we give an algorithm to compute t_{1c}. It is based on the *evaluation of the accessibility set* along the reference trajectory x using a *seminormal form* for the action of the *feedback group*. The computations of conjugate points in [2] require linear transformations and integration of the vector field Y and the reference trajectory. A different algorithm without integration will be described in §5. This theorem solves the time-optimality problem of singular trajectories satisfying (H0) – (H3) when the terminal manifold is a point. It has to be adapted to deal with the situation encountered with the control of batch reactors where the terminal manifold is of codimension one.

3.4. Connection between singular extremals in the affine and nonaffine case. For a batch reactor, the system is affine if $\frac{dT}{dt}$, denoted $\dot{T}$, is the control (T is then a state variable) and nonaffine if the control is T. The object of this section is to relate singular extremals in both cases.

Let us consider a general system of the form (3.1). This system can be interpreted as an affine system with respect to a new control v if we set $\dot{u} = v$, that is, if we introduce an integrator. Let us study the converse transformation, in optimal control called Goh's transformation.

DEFINITION 3.6. Let us consider an affine single-input system of $\mathbf{R}^n$, $\dot{x} = X + uY$, and let us assume $n \geq 2$. Take $x_0 \in \mathbf{R}^n$ such that $Y(x_0) \neq 0$. Hence there exists an open set U containing x_0 such that $Y_{|U} = \frac{\partial}{\partial x^n}$, $(x^1, \dots, x^n)$ are the coordinates of $\mathbf{R}^n$, and the restriction of the system to U can be written as

$$\dot{x}' = X'(x', x^n), \qquad \dot{x}^n = X^n(x) + u,$$

where $x' =^t (x^1, \dots, x^{n-1})$ and $X = X' \frac{\partial}{\partial x'} + X^n \frac{\partial}{\partial x^n}$. The system $\dot{x}' = X'(x', x^n)$, where x^n is the control variable and which is defined on an open set U' of $\mathbf{R}^{n-1}$, is called the *reduced system* associated with system (X, Y). If $H = \langle p, X + uY \rangle$ is the Hamiltonian of the original system, we set $H'(x', p', x^n) = \langle p', X'(x', x^n) \rangle$, where $p' = (p_1, \dots, p_{n-1})$ is the dual variable of x'.

LEMMA 3.7. *The pair (x, p) is the projection on the space $\{(x, p)\}$ of a solution (x, p, u) of $\dot{x} = \frac{\partial H}{\partial p}$, $\dot{p} = -\frac{\partial H}{\partial x}$, $\frac{\partial H}{\partial u} = 0$ if and only if (x', p', x^n) is a solution of $\dot{x}' = \frac{\partial H'}{\partial p'}$, $\dot{p}' = -\frac{\partial H'}{\partial x'}$, $\frac{\partial H'}{\partial x^n} = 0$. Moreover the following relations are satisfied*:

$$\text{(i)} \qquad \left(\frac{d}{dt} \frac{\partial H}{\partial u} \right)_{|(x,p,u)} = \langle p, [X, Y](x) \rangle = -\frac{\partial H'}{\partial x^n}_{|(x',p',x^n)},$$

$$\text{(ii)} \qquad \frac{\partial}{\partial u} \frac{d^2}{dt^2} \frac{\partial H}{\partial u}_{|(x,p,u)} = -\langle p, \operatorname{ad}^2 Y(X)(x) \rangle = -\frac{\partial^2 H'}{\partial {x^n}^2}_{|(x',p',x^n)}.$$

For the proof of this result, see [2].

DEFINITION 3.8. *Let (x, p, u) be an singular extremal of* (3.1). *The condition*

$$\frac{\partial}{\partial u} \frac{d^2}{dt^2} \frac{\partial H}{\partial u}_{|(x,p,u)} \geq 0$$

is called the Legendre–Clebsch condition.

COROLLARY 3.9. *The Legendre–Clebsch condition along a singular extremal is equivalent to the Legendre condition along the corresponding extremal for the reduced system.*

3.5. Projected problems. According to the theory developed in [1], symmetry properties of a control system have to be coded by symmetry properties for the differential equation whose solutions are singular trajectories. For batch reactors where every reaction is of first order, this will imply a nice projection property.

3.5.1. Statement of the problem (every object is real analytic). Let M and M' be two manifolds, π be a submersion from M into M', and N (resp., N') be a regular submanifold of M (resp., M') with $N = \pi^{-1}(N')$. Consider now the system on M : $\frac{dx}{dt} = f(x, u), u \in \Omega$, and the associated time-optimal control problem with terminal manifold N, denoted $\mathcal{P}$. Let us assume that for each fixed $u \in \Omega$, the vector field $x \longrightarrow f(x, u)$ can be π-*projected* and is complete for each admissible control $u(.)$. Hence, one may define the *projected system* on

M', $\frac{dx'}{dt} = f'(x',u), u \in \Omega$, where $x' = \pi(x)$ and $f'(x',u) = d\pi(f(x,u))$, $d\pi$ being the differential of π, and the associated time-optimal control problem $\mathcal{P}'$ with terminal manifold N' is called the *projection* of problem $\mathcal{P}$.

Our aim is to compare both extremals and optimal trajectories. Let us denote by $x(t,x_0,u)$ the solution of $\frac{dx}{dt} = f(x,u)$ starting at $t = 0$ from $x_0 \in M$ and by $x'(t,x'_0,u)$ the solution of $\frac{dx'}{dt} = f'(x',u)$ starting at $t = 0$ from $x'_0 = \pi(x_0)$. We have the following lemma.

LEMMA 3.10. *The trajectory $x^*(t,x_0,u^*)$ defined on $[0,t^*]$ is a solution of $\mathcal{P}$ if and only if $x(t,x'_0,u^*)$ is a solution of $\mathcal{P}'$.*

Proof. By completeness, we have for each t : $x'(t,x'_0,u) = \pi(x(t,x_0,u))$, and by definition $N = \pi^{-1}(N')$.

LEMMA 3.11. *Every extremal (x',p',u) defined on $[0,T]$ of the projected problem $\mathcal{P}'$ can be lifted to an extremal (x,p,u) of the original problem $\mathcal{P}$. Moreover if (x',p') satisfies the boundary conditions $x'(T) \in N'$ and $p'(T)$ orthogonal to $T_{x'(T)}N'$ imposed by $\mathcal{P}'$, then (x,p) can be selected to satisfy the boundary conditions imposed by $\mathcal{P}$.*

Proof. The system $\frac{dx}{dt} = f(x,u)$ can be written locally as

$$\frac{dx'}{dt} = f'(x',u), \qquad \frac{dx''}{dt} = f''(x',x'',u).$$

Hence every extremal (x',p',u) of $\mathcal{P}'$ can be lifted into $(x,p,u) = ((x',x''),(p',0),u)$, where x'' is any solution of the second equation. Clearly (x,p,u) is an extremal of $\mathcal{P}$, and we have $x'(T) \in N', p'(T)$ orthogonal to $TN' \Leftrightarrow x(T) \in N$ and $p(T)$ orthogonal to TN.

Conclusion 3.12. Not every extremal of $\mathcal{P}$ can be projected onto an extremal of $\mathcal{P}'$. *Hence, although $\mathcal{P}'$ is equivalent to $\mathcal{P}$, from Lemma* 3.10, *its analysis using PMP is simpler because we have fewer extremals.*

3.6. Regular extremals. Below, we shall briefly recall some useful results concerning the behavior of regular extremals for a single-input affine system (3.6) with $u \in [-1,+1]$ (which will be used extensively later). (See [18] for details.)

DEFINITION 3.13. *Let (z,u), where $z = (x,p)$, be an extremal defined on $[0,T]$. A time $s \in [0,T]$ is called a* switching time *if s belongs to the closure of the set of $t's \in [0,T]$, where z is not C^1. The set $\{z(s)\}$, z being any extremal, where s is a switching time, is called the* switching set. *Observe that this set is a subset of $\Sigma = \{z = (x,p); \langle p, Y(x)\rangle = 0\}$.*

3.7. Classification of regular extremals. Let $z = (x,p)$ be a *smooth* solution of $\dot{x} = \frac{\partial H}{\partial p}$, $\dot{p} = -\frac{\partial H}{\partial x}$ defined on $[0,T]$, where $H = \langle p, X + uY\rangle$ and corresponding to the control $u(.)$. Let us introduce the *switching function* $\Phi : t \to \langle p(t), Y(x(t))\rangle$ evaluated along z. If $u \equiv +1$ (resp., -1), we set $z = z^+$ and $\Phi = \Phi^+$ (resp., $z = z^-$ and $\Phi = \Phi^-$). By differentiating Φ twice with respect to t we get

$$\begin{aligned} \dot{\Phi}(t) &= \langle p(t), [X,Y](x(t))\rangle, \\ \ddot{\Phi}(t) &= \langle p(t), \operatorname{ad}^2 X(Y)(x(t)) - u(t) \operatorname{ad}^2 Y(X)(x(t))\rangle. \end{aligned} \tag{3.9}$$

3.7.a. Normal switching points. Let $z_0 = (x_0,p_0) \in \Sigma$, and let us assume that $Y(x_0) \neq 0$ and $z_0 \in \Sigma \backslash \Sigma'$. (Recall that $\Sigma' = \{(x,p); \langle p, Y(x)\rangle = \langle p, [X,Y](x)\rangle = 0\}$.) The point z_0 is then called *normal*. The behavior of regular extremals near z_0 is given by Fig. 2, where Σ^+ (resp., Σ^-) $= \{(x,p); \langle p, Y(x)\rangle > 0\}$ (resp., < 0). From (3.9) we have the following lemma.

LEMMA 3.14. *Let t_0 be the switching time given by $z^+(t_0) = z^-(t_0) = z_0$. Then we have*

$$\dot{\Phi}^+(t_0) = \dot{\Phi}^-(t_0) = \langle p_0, [X,Y](x_0)\rangle \quad (\textit{reflexion law}). \tag{3.10}$$

Moreover, let $z = (x,p)$ be an extremal passing through z_0; then we have

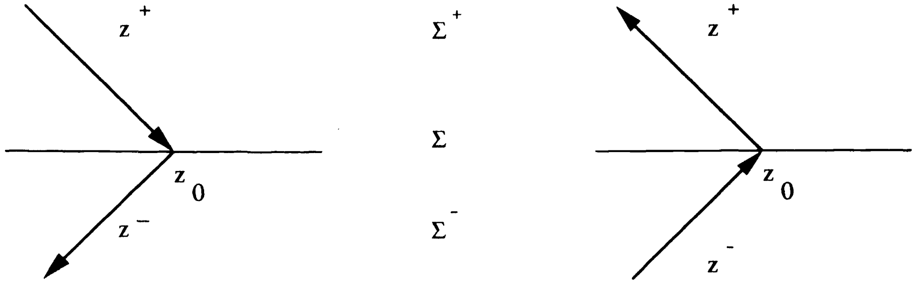


FIG. 2.

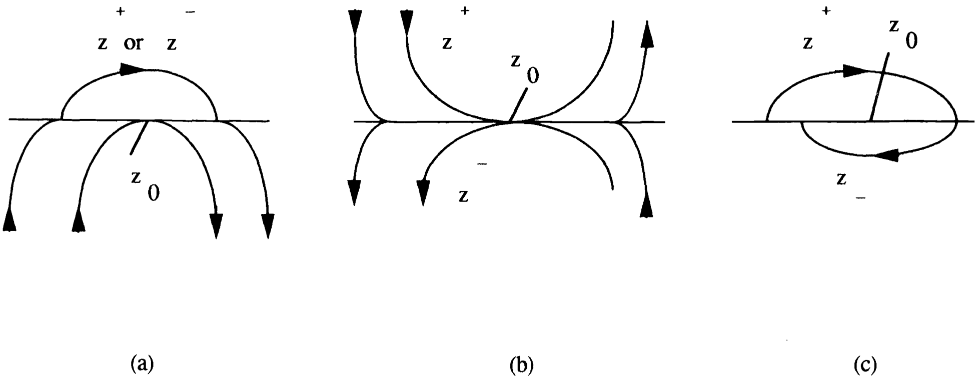


FIG. 3.

(i) *if* $\langle p_0, [X,Y](x_0)\rangle < 0$, *then* $x = \Gamma_+\Gamma_-$;

(ii) *if* $\langle p_0, [X,Y](x_0)\rangle > 0$, *then* $x = \Gamma_-\Gamma_+$,

where Γ_+ (*resp.*, Γ_-) *is an arc solution of the system corresponding to* $u \equiv 1$ (*resp.*, $u \equiv -1$) *and* $\Gamma_+\Gamma_-$ *represents the trajectory corresponding to the concatenation of* $u \equiv +1$ *with* $u \equiv -1$; *i.e., an arc* Γ_+ *is followed by an arc* Γ_-.

3.7.b. The fold case. Let $z_0 = (x_0, p_0) \in \Sigma'$, and let us assume that $Y(x_0) \neq 0$ and both λ^+ and λ^- are different from zero where

$$\lambda^\pm = \langle p_0, \mathrm{ad}^2X(Y)(x_0) \mp \mathrm{ad}^2Y(X)(x_0)\rangle.$$

Such a point is called a *fold*, and the behavior of regular extremals near z_0 has been classified in [18]. We have three distinct cases (Fig. 3), which are characterized by

a) $\lambda^+\lambda^- > 0$ (*parabolic case*),

b) $\lambda^+ > 0$ and $\lambda^- < 0$ (*hyperbolic case*),

c) $\lambda^+ < 0$ and $\lambda^- > 0$ (*elliptic case*), and from [2] and [18] we have the following lemma.

LEMMA 3.15. *The point* z_0 *is a switching point in the hyperbolic case and is not a switching point in the elliptic case. If* z_0 *is an ordinary point, the singular dynamic feedback* $\hat{u}$ *given by* (3.7) *and evaluated at* z_0 *belongs to* $]-1,+1[$ *in the hyperbolic and elliptic case, contrary to the parabolic case. Moreover the extremals near* z_0 *are of the forms* $\Gamma_+\Gamma_s\Gamma_-$, $\Gamma_+\Gamma_s\Gamma_+$, $\Gamma_+\Gamma_s\Gamma_-$, or $\Gamma_-\Gamma_s\Gamma_-$, *where* Γ_s *is a singular arc, in the hyperbolic case, and* $\Gamma_+\Gamma_-\Gamma_+$ *or* $\Gamma_-\Gamma_+\Gamma_-$ *in the parabolic case; in the elliptic case the only extremal passing through* z_0 *is* Γ_s, *and the other extremals are of the form* $\Gamma_+\Gamma_-\Gamma_+\Gamma_-\ldots$ (*no uniform bounds on the number of switchings*). *Moreover, let us assume that the singular extremal arc* *passing through* z_0 *satisfies assumptions* (H0) – (H3). *Then it is hyperbolic* (*hence fast*) *in case* b) *and elliptic* (*hence slow*) *in case* c).

Remark 3.16. The adjoint vector p has to be oriented in our terminology with the convention (3.4) of the maximum principle ($H \geq 0$).

4. Time-minimal synthesis for planar systems in the neighborhood of a terminal manifold of codimension one.

4.1. Problem statement. Consider a system in $\mathbf{R}^2$ of the form

$$\dot{v} = X(v) + uY(v), \qquad |u| \leq 1, \tag{4.1}$$

where X and Y are analytic vector fields, and let N be an analytic regular submanifold of $\mathbf{R}^2$ of codimension one. The set of admissible controls $\mathcal{U}$ is the set of measurable functions with values in $[-1, +1]$. We shall study the following *local* problem. Let $v_0 \in N$. Compute, in a sufficiently small open neighborhood U of v_0, the optimal synthesis for the time-minimal control problem with terminal manifold N and system (4.1) restricted to U. This problem is well posed because a standard theorem [19] proves the existence of an optimal solution. It is similar to the problem studied by Schättler and Sussmann in a series of articles (see, for example, [26] and [27]) when the terminal manifold is reduced to a point. (This problem is called the *point-to-point problem.*) The aim of this section is to give the tools to solve the problem and to begin a classification of optimal *syntheses* in terms of relations between the Taylor expansions of X, Y, and f at v_0, where f is the mapping whose zero set is locally N. A more complete classification is given in [4].

DEFINITION 4.1 AND NOTATION. *Consider the system* (4.1) *written as* (X, Y), *and let us denote by* (x, y) *the coordinates of* $v \in \boldsymbol{R}^2$. *A coordinate system* (U, v) *such that the restriction of* Y *to* U *is* $\frac{\partial}{\partial y}$ *and will be called* adapted. *The optimal control problem is said to be flat if* Y *is tangent to* N *everywhere. A normal to* N *at* v *is denoted by* $n(v)$. *We lift* N *by using the transversality condition into* $\hat{N} = \{(v,p) \in \boldsymbol{R}^2 \times \boldsymbol{R}^2;\ v \in N, \langle p, w\rangle = 0 \,\forall w \in T_vN\}$. *An extremal* (v, p, u) *defined on* $[T, 0]$, $T < 0$ *that satisfies the boundary conditions* $(v(0), p(0)) \in \hat{N}$ *will be called a* BC-extremal. *We shall denote by* K *the projection on the v-space of the set of switching points for BC-extremals. Let* $v_0 \in N$, *and let* W *be the set of* optimal *switching points for the time-minimal control problem for* (4.1) *restricted to a sufficiently small neighborhood of* v_0, *with* N *the terminal manifold. By convention, any piecewise analytic control is taken right-continuous. The optimal closed-loop function, if it exists, is denoted by* $v \longrightarrow u^*(v)$. *For the concepts of synthesis we follow* [28]. *A stratum of the switching curve* W *is of* first kind *if the optimal trajectories are tangent to the stratum and of second kind if they are transverse. Following* [24], *the* splitting line L *is the set of points where the optimal feedback is not unique.* (*It will form the* cut locus.)

Let $v_1 \in \boldsymbol{R}^2$, *and let us denote by* $v(t, v_1, u)$ *the trajectory of* (4.1), *when defined, associated with* $u \in \mathcal{U}$ *and starting from* v_1 *at time* $t = 0$. *Let us denote by* $A^+(v_1, t)$, $t > 0$, *the set of points* $\{v(t, v_1, u); u \in \mathcal{U})$ accessible *from* v_1 *in time* t, *and let* $A^-(v_1, t)$ *be the set of points* v_2 *such that* v_1 *is accessible from* v_2 *in time* t. *The* accessibility set *is* $A^+(v_1) = \cup_{t>0} A^+(v_1, t)$, *and let* $A^-(v_1) = \cup_{t>0} A^-(v_1, t)$. *At* v_0, *we shall denote by* $C(v_0)$ *the convex set* $\{X(v_0) + uY(v_0); |u| \leq 1\}$. *In our analysis, we shall assume that* $Y(v_0) \neq 0$ *and* $C(v_0)$ *lies entirely in one half space limited by* $T_{v_0}N$. *If* $X(v_0) + Y(v_0)$ *or* $X(v_0) - Y(v_0)$ *is tangent to* N, *we are in the* exceptional case. *In the* nonexceptional case, *near* v_0, $n(v)$ *will be oriented toward the half space containing* $C(v_0)$.

4.2. Generic case. Let us assume that both $X(v_0) \pm Y(v_0)$ are not tangent to N. Then with our convention we have $\langle n(v), X(v)\rangle > 0$ for $|v - v_0|$ small. Let (v, p, u) be a BC-extremal defined on $[T, 0]$. Since we are in the nonexceptional case, one can set $p(0) =$

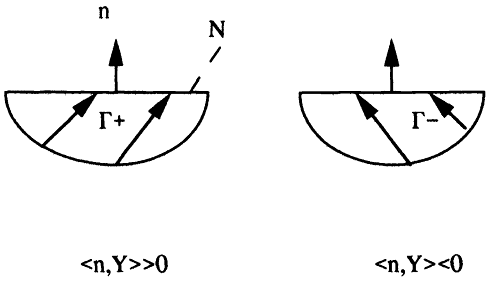


FIG. 4.

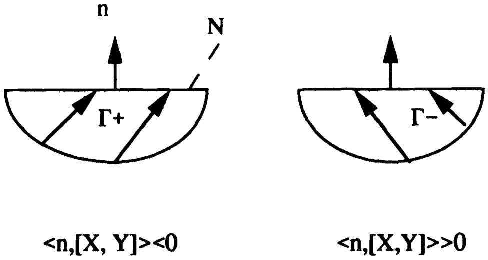


FIG. 5.

$n(v(0))$. Let us assume $\langle n(v_0), Y(v_0)\rangle \neq 0$. Then, by using the *transversality condition*, the optimal synthesis, in a sufficiently small open neighborhood of v_0, is given in Fig. 4, where Γ_+ (resp., Γ_-) are arcs corresponding to $u = 1$ (resp., -1).

4.3. Generic flat case. As before, we assume that both $X(v_0) \pm Y(v_0)$ are not tangent to N. If (v, p, u) is a BC- extremal defined on $[T, 0]$, one may set $p(0) = n(v(0))$. Since Y is tangent to N everywhere, we have $\langle n(v), Y(v)\rangle = 0$ for each $v \in N$. Hence, the transversality condition tells us nothing about the optimal synthesis. But since N lies in K (i.e., 0 is a switching time of (v, p)), then by Lemma 3.14, if $\langle n(v_0), [X, Y](v_0)\rangle \neq 0$ the arcs Γ_+ (resp., Γ_-) hitting the target N are BC-extremals if and only if $\langle n(v_0), [X, Y](v_0)\rangle < 0$ (resp., > 0), and the synthesis is seen in Fig. 5.

4.4. Generic switching point. If $\langle n(v_0), Y(v_0)\rangle = 0$, then $\hat{N}$ intersects the set $\Sigma = \langle p, Y\rangle = 0$ at $(v_0, n(v_0))$. To analyze this singularity one needs some preliminary lemmas.

LEMMA 4.2. *Let us assume $\langle n(v_0), [X, Y](v_0)\rangle \neq 0$. Then the arcs Γ_+ and Γ_- arriving at v_0 cannot be sets of input switching points.*

Proof. For instance, let us assume that $(v, p, u \equiv 1)$ is a BC-extremal on $[T, 0]$, with $v(0) = v_0$, and each point of v is an input switching point. Then, since $(v(0), p(0))$ is a normal switching point, from Lemma 3.14, there exist extremals $\Gamma = \Gamma_-\Gamma_+$, where Γ_+ are any subarcs of v. If Γ is a BC-extremal at v_0, the adjoint vector can be taken as $p(0)$. Now, since the image of $v \subset K$, we have $\langle p(t), Y(v(t))\rangle = 0$, $\forall t \in [T, 0]$. Hence the arc $v(.)$ is singular, which is absurd.

LEMMA 4.3. *Let Γ be an admissible trajectory arriving at v_0 and associated with a* constant *control u_0. Let us set $Z = X + u_0 Y$ and $\lambda(\delta, P) = \sum_{k\geq 0}((-1)^k\delta^k/k!)\,\mathrm{ad}^k Z(P)(v_0) - Z(v_0)$. Then if Γ is optimal for each $\delta \geq 0$ small and each vector field P of $\{X + uY;\ |u| \leq 1\}$, we must have $\langle n, \lambda(\delta, P)\rangle \leq 0$, where n is the unit normal to N at v_0, outwardly oriented with respect to Γ.*

Proof. We use a technique from the proof of PMP and its refinements [19]. We construct along a reference trajectory, an approximation of the accessibility set. Since the terminal manifold is of codimension one, this approximation *need not be convex* to decide about optimality. The necessary condition of the lemma is obtained as follows.

Let Γ be a reference trajectory defined on $[0,T]$ and with terminal point v_0. If V is a vector field, it is convenient to denote by $\{\exp tV\}$ the local one-parameter group generated by V. In particular the arc Γ starting from v_1 at $t=0$ is given by $\exp tZ(v_1)$ and $\exp TZ(v_1)=v_0$.

Now take δ, $\varepsilon>0$, sufficiently small and any vector field P of $\{X+uY; |u|\leq 1\}$, and consider for δ fixed the curve

$$\alpha(\varepsilon)=(\exp\delta Z)(\exp\varepsilon P)(\exp(T-\delta-\varepsilon)Z)(v_1).$$

By construction $\alpha(0)=v_0$, and $\alpha(\varepsilon)$ lies in the accessibility set $A^+(v_1,T)$. Now since $v_0=\exp TZ(v_1)$ we have

$$\alpha(\varepsilon)=(\exp\delta Z)(\exp\varepsilon P)(\exp(-\delta-\varepsilon)Z)(v_0),$$

and from the Baker–Campbell–Hausdorff formula we have

$$\alpha(\varepsilon)=\left(\exp\left(\varepsilon\left[\sum_{k\geq 0}\frac{(-1)^k\delta^k}{k!}\,\mathrm{ad}^k Z(P)-Z\right]+o(\varepsilon)\right)\right)(v_0).$$

Hence $d\alpha/d\varepsilon_{|\varepsilon=0}=\lambda(\delta,P)$. And clearly, if $\langle n,\lambda(\delta,P)\rangle>0$, the reference trajectory Γ is not optimal.

Assumption 4.4. From now on we shall assume that $\langle n(v_0),Y(v_0)\rangle=0$ and both $\langle n(v_0),X(v_0)\rangle$ and $\langle n(v_0),[X,Y](v_0)\rangle$ are nonzero.

4.4.1. Method of analysis. To evaluate the switching curve and the splitting line near v_0, it is convenient to use the following model.

First, one may set $v_0=(0,0)$, and as in [2], since X and Y are transverse at v_0, one may assume locally that $Y=\frac{\partial}{\partial y}$ and $t\to(t,0)$ is the trajectory corresponding to $u\equiv 0$. Hence (4.1) can be written

$$\begin{aligned}\dot{x}&=1+\textstyle\sum_{i=1}^{+\infty}a_i(x)y^i,\\ \dot{y}&=\textstyle\sum_{i=1}^{+\infty}b_i(x)y^i+u.\end{aligned}\tag{4.2}$$

Moreover if we change y into $-y$ and u into $-u$ if necessary, we can assume that $a=a_1(0)>0$, where $a=-\langle n(0),[X,Y](0)\rangle, n(0)=(1,0)$ being the unit normal to N at 0. The terminal manifold is given locally by $s\to(c(s),s)$, where $c(s)=ks^2+o(s^2)$ and k parametrizes the curvature of N in the adapted coordinate system. At 0, we choose $n(0)=(1,0)$, and for v small, using the convention $\langle n(v),X(v)\rangle>0$, one can set $n=(n_1,n_2), n_1=1$, and $n_2=-\frac{dc}{ds}=-2ks+o(s)$. Hence for s small we have that if $k<0$, then $n_2>0$ if $s>0$ and $n_2<0$ if $s<0$ and conversely if $k>0$. The Hamiltonian at a point (v,n) of $\hat{N}$ is $H(v,n,u)=\langle n,X(v)+uY(v)\rangle$, and for s small, its maximum over $|u|\leq 1$ is obtained as follows: If $k<0$, $s>0$, then $n_2>0$ and u maximizing H is $+1$. If $k<0$, $s<0$, then u maximizing H is -1; the converse is true if $k>0$. Hence we get the following important geometric behaviors. *If $k<0$, the arcs Γ_+ and Γ_- satisfying the transversality conditions, the normal to N being oriented as $n(v)$, can cut themselves, contrary to the case when $k>0$* (Fig. 6).

The adjoint system associated with (4.1), with $p=(p_1,p_2)$, is

$$\dot{p}_1=-p_1\sum_{i=1}^{+\infty}a_i'(x)y^i-p_2\sum_{i=1}^{+\infty}b_i'(x)y^i,\tag{4.3}$$

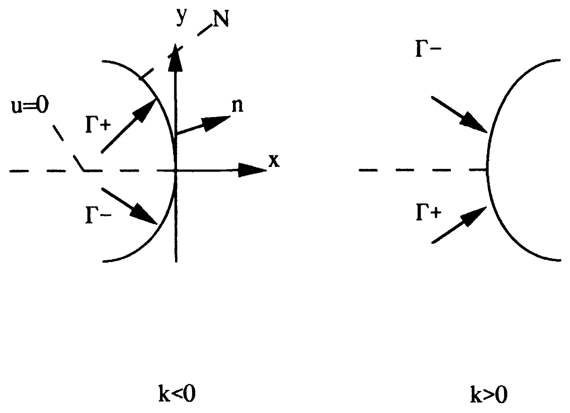


FIG. 6.

$$\dot{p}_2 = -p_1 \sum_{i=1}^{+\infty} a_i(x) y^{i-1} - p_2 \sum_{i=1}^{+\infty} b_i(x) y^{i-1},$$

where a_i' and b_i' are the derivatives with respect to x. *If u is a piecewise analytic control, every solution of* (4.2) *and* (4.3) *satisfying the boundary conditions can be evaluated for small t by analyticity.*

LEMMA 4.5. *Near* 0, *every optimal solution is of the form* $\Gamma_+\Gamma_-$.

Proof. From Lemma 3.14, we know that every BC-extremal is of form $\Gamma_+\Gamma_-$ or $\Gamma_-\Gamma_+$. In fact it follows from [27] that every optimal solution for the point-to- point optimal problem is of this form.

Since X and Y are linearly independent near 0, to compare the times along the solutions of the system we can introduce the one-form ω defined by $\omega(Y) = 0$ and $\omega(X) = 1$. If (X_1, X_2) are the components of X, we have $\omega = (1/X_1)\, dx$ and $d\omega = \frac{1}{X_1^2}\frac{\partial X_1}{\partial y} dx \wedge dy$, and by computing with (4.2), we see that the sign of $d\omega$ near 0 is the sign of $a > 0$. Let $\Gamma_1 = \Gamma_+\Gamma_-$ and $\Gamma_2 = \Gamma_-\Gamma_+$ be two arcs joining v_1 to v_2 near 0 with respective time duration t_1 and t_2. By using Stokes' theorem we have

$$\int_{\Gamma_1} \omega - \int_{\Gamma_2} \omega = t_1 - t_2 = \int_D d\omega,$$

where D is the closed domain limited by $\Gamma_1 \vee -\Gamma_2$. If the orientation is < 0 (resp., > 0), since by $d\omega > 0$ on D we have $t_2 > t_1$ (resp., $t_1 > t_2$). Therefore, optimal solutions for the point-to-point problem are of the form $\Gamma_+\Gamma_-$. Clearly, every solution for the optimal problem with N as terminal manifold has to be solution for the point to point problem.

LEMMA 4.6. *The arc* Γ_-^0 *arriving at* 0 *is not optimal.*

Proof. By computing we have $\langle n(0), [X, Y](0)\rangle = -a < 0$. Hence from Lemma 3.14 the arc Γ_-^0 is not a BC-extremal.

LEMMA 4.7. *Let us assume* $k \neq 0$; *then the switching points of BC-extremals* $\Gamma_+\Gamma_-$ *form an analytic curve* K *whose tangent space at* 0 *is* $\mathbf{R}(-2k/a, 1 + 2k/a)$.

Proof. We integrate (4.2) and (4.3) backward in time, with initial conditions given by the boundary conditions $v(0) \in N$ and $p(0) = n(v(0)) = (1, -2ks + o(s))$. We get $p_1(t) = 1 + 0(s,t)$, $p_2(t) = -2ks - at + o(s,t)$. The switching times w are given by solving

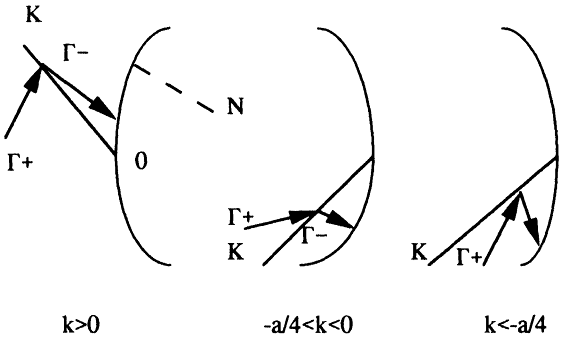


FIG. 7.

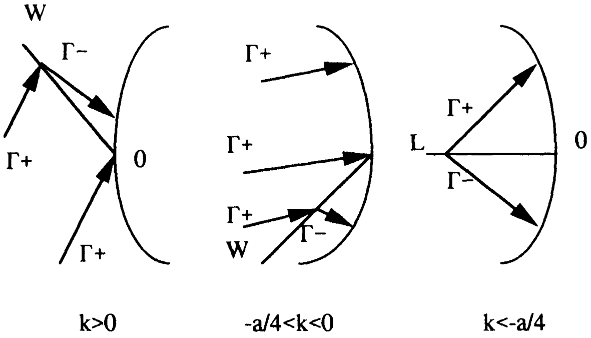


FIG. 8.

$p_2(t) = 0$, $t \leq 0$. We get $w = -2ks/a + o(s)$. If $k < 0$, we must have $s < 0$, and if $k > 0$, $s > 0$, and the BC-extremal Γ_- is switching at $(x(w), y(w)) = s(-2k/a, 1+2k/a)+o(s)$. The lemma is then proved.

Clearly the optimal synthesis depends on the fact that a BC- extremal $\Gamma_+\Gamma_-$ is *crossing* K or *reflecting* on K.

LEMMA 4.8. *A BC-extremal* $\Gamma_+\Gamma_-$ *is crossing* K *if* $k > 0$ *or* $-a/4 < k < 0$, *and it is reflecting on* K *if* $k < -a/4$.

Proof. At 0, the slope of the tangent to K is $-1 - a/2k$, and Γ_+ and Γ_- have (1,1) and $(-1, 1)$ as tangents. If $k > 0$, the slope of the tangent to K is less than -1. If $k < 0, -1 - a/2k > 1$ if and only if $-a/4 < k$. Hence the geometries are given in Fig. 7.

PROPOSITION 4.9. *The optimal syntheses are given by Fig.* 8. *In the first two cases, the switching curve* W *is an analytic curve which coincides with* K, *the slope of the tangent at* 0 *being* $-1 - a/2k$. *In the third case, there exists a splitting line* L *which is an analytic curve on which the optimal feedback can be* ± 1, *the slope of its tangent at* 0 *being* $-a/4k$.

Proof. In the first two cases, the situation is clear because from each point near 0, at the left of the target N, there is only one BC-extremal $\Gamma_+\Gamma_-$. In the third case, the situation is more complicated because more than one BC- extremal $\Gamma_+\Gamma_-$ is possible to reach the target. More precisely, let Γ_+^0 be the extremal arc arriving at 0 and A be the acute sector delimited by K and Γ_+^0. Clearly above K the optimal feedback law is $u = 1$, and below Γ_+^0 it has to be -1. Let us define the splitting line L near 0 as follows. L is the set of $v = (x, y)$, v small, $x < 0$, such that $\exists\, t \geq 0$ such that both $\exp t(X \pm Y)(v)$ intersects N (see Fig. 9).

By construction, near 0, L is an analytic curve which is located in the sector A. Clearly,

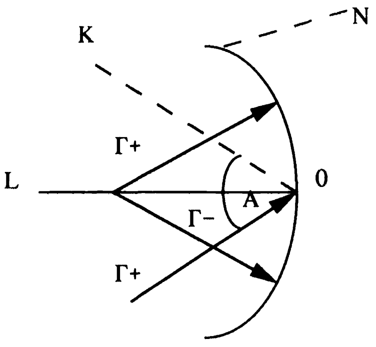


FIG. 9.

above L the optimal feedback is $+1$, and below it is -1.

In fact, everything can be evaluated by using (4.2) and (4.3). If $k + a/4 < 0$, it can be shown that the arc Γ_+^0 is not optimal. Moreover, the slope of the tangent to L at 0 is $-a/4k \in]0, 1[$.

From our analysis we deduce that $\dot{x} = 1 + ay, \dot{y} = u$ is the local model of the behaviors, and the linear approximation of K and L is given by the model.

4.5. Generic fold case. In this section, we shall analyze the situation encountered when a singular extremal satisfying the transversality condition meets the terminal manifold N. In fact, we shall only consider the *hyperbolic situation.* This situation is technically relevant because the analysis is carried on using an *evaluation of the accessibility set.* A complete analysis, which is lengthy, is given in [4].

Assumption 4.10. Let $(p, v_0) \in \hat{N}$, and let us assume $\langle p, Y(v_0)\rangle = \langle p, [X, Y](v_0)\rangle = 0$ and $\langle p, X(v_0)\rangle \neq 0$. Let $S = \{v \in \mathbf{R}^2;\ \det\,(Y(v), [X, Y](v)) = 0\}$. From §3.2, all singular extremals are contained in S, the singular control $\hat{u}$ being given by (3.7), and it is *admissible* if $\hat{u}(v) \in [-1, +1]$. We shall assume that v_0 is a *regular point* of S. Hence, since X and Y are not collinear at v_0, there exists a unique singular arc Γ_s which is a simple curve defined on $[T, 0]$, $T < 0$, with $\Gamma_s(0) = v_0$, which can be lifted into a unique extremal $(\Gamma_s, p, \hat{u})$, where p satisfies $p(0) = p_0$, unit vector transverse to N, and oriented such that $\langle p_0, X(v_0)\rangle > 0$. According to §3.7, (p_0, v_0) is a fold. Let us assume that this is an ordinary point. From §3.7, if $\hat{u}(v_0) \notin]-1, +1[$, it is parabolic, and if $\hat{u}(v_0) \in]-1, +1[$, it can be hyperbolic (hence fast) or elliptic (hence slow). We shall analyze only the hyperbolic case.

4.5.1. Model. We choose an adapted coordinate system such that $v_0 = 0$ and in which the singular arc Γ_s is identified with $t \longrightarrow (t, 0)$. Hence system (4.1) can be written

$$\begin{aligned} \dot{x} &= 1 + a(x)y^2 + o_x(y^2), \\ \dot{y} &= -\hat{u}_{|y=0} + yX_2(v) + u, \end{aligned} \tag{4.4}$$

where $a(0) = \langle n, \mathrm{ad}^2 Y(X)(0)\rangle \neq 0$, with $n(0) = (1, 0)$ unit normal to $N = (ks^2 + o(s^2), s)$. Observe that if $a(0) < 0$ (resp., > 0), the singular arc is hyperbolic (resp., elliptic).

Let $p = (p_1, p_2)$ be the adjoint variable with $p(0) = (1, -2ks + 0(s^2))$ orthogonal to N. By computing with (4.4) and $u = \varepsilon$, $\varepsilon = \pm 1$ we get

$$p_2(t) = -2ks - a(0)(\varepsilon - \hat{u}(0))t^2 + k_1 st + k_2 s^2 + o(s, t)^2, \tag{4.5}$$

where k_1, $k_2 \in \mathbf{R}$ are coefficients which are unimportant for our discussion. Hence, we get the following lemma.

LEMMA 4.11. *In the hyperbolic case, any BC- regular extremal which meets N at a point $v \neq 0$, v small, has no switching if $k \neq 0$.*

Now, let us evaluate the accessibility set near an hyperbolic singular trajectory.

PROPOSITION 4.12. *Let (v, p, u) be a hyperbolic singular extremal defined on $[0, T]$, with $\hat{u}$ given by* (3.7) *belonging to* $]-1, +1[$, *satisfying the assumptions* (H0) – (H3). *Then there exists a neighborhood U of v such that v is the time-optimal trajectory joining $v(0)$ to $v(T)$ among all solutions of* (4.1). *Moreover if U is sufficiently small, the accessibility set $A_U^+(v(0), T)$ near $v(T)$ is a closed convex set with nonempty interior whose boundary is a curve $s \mapsto d(s)$ with $d(0) = v(T)$, $d'(0) \in \mathbf{R}Y(v(T))$, C^2 but not in general C^3. Moreover in every adapted coordinate system its curvature is zero.*

Proof. From [2], since for a planar system a singular extremal satisfying (H0)–(H3) is without conjugate points, v is time minimal among all solutions of $\dot{v} = X + uY$ contained in a sufficiently small neighborhood U and joining $v(0)$ to $v(T)$ with $u \in \mathbf{R}$. Hence it has to be optimal if $|u| \leq 1$. Moreover from [2], [27], we can choose U such that every optimal trajectory starting from $v(0)$ is a singular arc Γ_s followed by Γ_+ or Γ_-. Hence, near $v(T)$, the boundary of $A_u^+(v(0), T)$ is parametrized by $s \mapsto d(s)$, where $s \geq 0$ and

$$d(s) = (\exp s(X \pm Y))(\exp (T - s)\hat{X})(v(0)),$$

with $\hat{X} = X + \hat{u}Y$. Since $v(T) = \exp T\hat{X}\,(v(0))$, we get

$$d(s) = (\exp s(X \pm Y))(\exp -s\hat{X})(v(T)),$$

and by using the Baker–Campbell–Hausdorff formula we have

$$d(s) = \exp \left[s(\pm 1 - \hat{u})Y + \frac{1}{2}s^2[X \pm Y, \hat{X}] + o(s^2) \right] (v(T)).$$

The curve $d(s)$ can be evaluated by using Chen's formula,

$$\exp\, s\,Z(v) = \sum_{n \geq 0} \frac{s^n Z^n}{n!}(\text{id})(v),$$

for s sufficiently small, where Z is any vector field acting by Lie derivative on the mappings and id is the identify mapping. If $Y = \frac{\partial}{\partial y}$, we have $Y^n(\text{id}) = 0$, where $n > 1$, and since along a singular trajectory Y and $[X, Y]$ are collinear we get

$$d(s) = v(T) + [s(\pm 1 - \hat{u}) + o(s)]Y(v(T)) + o(s^2).$$

This proves the assertion.

Higher-order dimensional expansions would tell us the nature of it singularity. For instance, if the system is given by $\dot{x} = 1 - y^2$, $\dot{y} = u$, and $v(0) = (0, 0)$, we get $d(s) = (T - \frac{s^3}{3}, \varepsilon s)$ with $\varepsilon = \pm 1$. Hence the boundary is the graph of $x = T - \frac{|y|y^2}{3}$.

PROPOSITION 4.13. *In the hyperbolic case, if $k \neq 0$, the optimal syntheses are given by Fig.* 10. *Moreover, in the flat case, the synthesis is given by the second case in Fig. 10.*

Proof. Let us assume $k \neq 0$. From Lemma 4.11, any BC-extremal which meets the target N at a point $v \neq v_0$ has no switching point. To decide if the arc Γ_s is optimal, we use the previous proposition. The system can be written as (4.4), and Γ_s identified with $t \rightarrow (t, 0)$. Let $v = (-T, 0)$, $T > 0$, be a point of Γ_s. For U small, the boundary of the accessibility set $A_U^+(v, +T)$ has zero curvature; hence we have two situations (Fig. 11).

In the first situation, N meets the interior of the accessibility set; hence Γ_s is not time optimal for our problem, contrary to the second situation or in the flat case. The syntheses follow. The analysis in the flat case is similar.

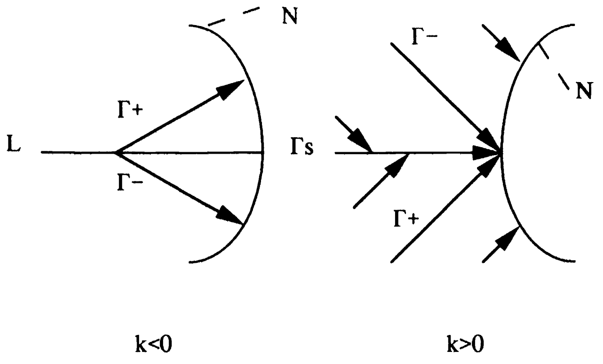


FIG. 10.

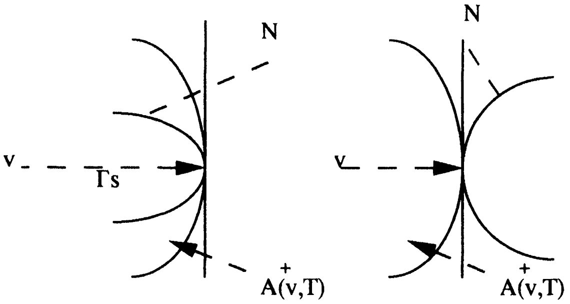


FIG. 11.

4.6. Generic exceptional case. *Assumption* 4.14 *and Normalizations*. Let $v_0 \in N$, and one may assume $v_0 = 0$. Suppose $X + Y$ tangent to N at 0. Moreover assume Y and $X - Y$ not tangent to N at 0. We can choose a coordinate system such that $Y = \frac{\partial}{\partial x}$ and N is identified to the curve $s \mapsto (0, s)$. Hence (4.1) can be written $\dot{x} = X_1 + u$, $\dot{y} = X_2$, with $X_1(0) = -1$ and $X_2(0) \neq 0$. We can suppose $X_2(0) > 0$. Moreover we assume $\frac{\partial X_1}{\partial y}(0) = a \neq 0$, which means that the contact of Γ_+ with N at 0 is one.

PROPOSITION 4.15. *Under the previous normalizations, the optimal synthesis is given by Fig.* 12.

Proof. First assume $a > 0$. The arc Γ_+^0 is a BC-extremal, $n = (1, 0)$ being the associated adjoint variable at 0. To prove that it is not optimal, we apply Lemma 4.3, the outward normal to N with respect to Γ_+^0 being $-n$, $Z = X + Y$, and $P = X - Y$. We get $\lambda(\delta, P) = -2Y(0) + o(1)$. Hence, for small δ, $\langle -n, \lambda(\delta, P)\rangle > 0$. This proves the assertion. Indeed, a simple computation shows the following. Assume that we are at distance ε from N in the domain $x > 0$. The time to reach the target N is of order $\sqrt{\varepsilon}$ along Γ_+^0 and of order ε along Γ_- because the contact of Γ_+^0 with N is one and Γ_- is transverse to N. In the domain $x < 0$, the optimal control is $u = +1$, the value function being not continuous.

When $a < 0$, the analysis is similar, but the target N is not accessible from the points in the sector $x < 0$ above Γ_+^0.

4.7. Generic flat exceptional case. The point $v_0 \in N$ is identified to 0, N to $s \mapsto (0, s)$; Y is assumed to be tangent to N, and X to be tangent to N at 0. Moreover we suppose Y, $X \pm Y$ not vanishing at 0 and $[X, Y]$ not tangent to N at 0. System (4.1) can be written $\dot{x} = ax + by + o(x, y)$, $\dot{y} = X_2 + u$, where $b = -\langle n, [X, Y](0)\rangle \neq 0$, $n = (1, 0)$ normal to N. Clearly, one may assume $a = 0$, $b = 1$, and $1 + X_2(0) > 0$.

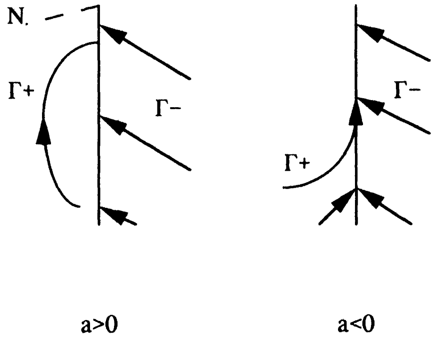


FIG. 12.

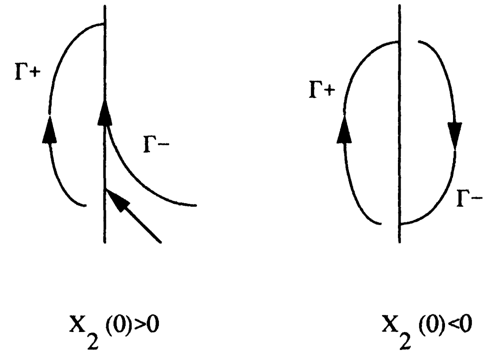


FIG. 13.

PROPOSITION 4.16. *Under the previous normalizations, the optimal synthesis is given by Fig.* 13.

Proof. According to §3.7 $z_0 = (0, n) \in \hat{N}$ is a normal switching point and hence every BC-extremal near z_0 is of the form $\Gamma_+\Gamma_-$ or $\Gamma_-\Gamma_+$. Since $\hat{N}$ is contained in $\langle p, Y\rangle = 0$, all the switching points are concentrated on N. Hence, near 0, every optimal trajectory is of an arc Γ_+ or Γ_-. Then, the synthesis follows from Lemmas 3.14 and 4.3.

5. Conjugate and focal point along a singular extremal.

5.1. Introduction. Consider a control system $\dot{x} = X + uY$ in $\mathbf{R}^n$, $|u| \leq 1$, and the time-optimal control problem for the *fixed end-point problem*. Let γ be an extremal defined on $[0, T]$. *One of the main problems is to compute the first point on γ where the extremal ceases to be minimal.* Such a point is called a *cut point*. In Riemannian geometry, a cut point can be a conjugate point which corresponds to a singularity of the exponential mapping or a point where *two isolated minimizing extremals* meet [15]. Hence, computing the cut locus splits into two problems. The second problem is clearly global, but the computation of conjugate points is *local* and accessible to the analysis when the correct topology has been identified (in fact the C^1-topology). The key tool for this analysis is *Jacobi's equation*. The behavior of the solutions of this equation depends on an invariant called the *sectional curvature*.

In time-optimal control the situation is much more complicated. We have to deal with regular and singular extremals, and the topology has to be chosen carefully. For *regular extremals* the problem has been well studied by Sussmann [27]. The computation of conjugate points along a *singular extremal* satisfying (H0)– (H3) is the object of [2]. The most interesting result is that computing a conjugate point can be done by using the equivalent of Jacobi's equation of the Riemannian case, although this equation is different if the reference extremal is hyperbolic or exceptional. This difference is due to the fact that in the hyperbolic case, an evaluation of the fixed-time accessibility set is sufficient to compute a conjugate point, contrary to the exceptional case, where its dependence with respect to the time has to be studied.

The object of this section is to give a *practical algorithm to compute conjugate and focal points* along a *singular extremal* for the batch reactor problem. For this reason and for the sake of simplicity, we consider only systems in $\mathbf{R}^3$, but our results can be extended to $\mathbf{R}^n$ by using Hamiltonian formalism [21]. A connection with the concept of *optimal synthesis* also is indicated, which is an attempt to *unify* the concept of conjugate point along a regular *and* a singular extremal.

DEFINITION 5.1. *Consider a control system* $\dot{x} = X + uY$, $x \in \mathbf{R}^n$, $u \in \mathbf{R}$, *and let* γ *be a singular extremal defined on* $[0,T]$. *The point* $y = \gamma(t_{1c}), t_{1c} \in [0,T[$ *is said to be the first conjugate point to* $x = \gamma(0)$ *if, for each* C^0*-sufficiently small neighborhood* U *of* γ, γ *is time optimal for all solutions of the system restricted to* U, *with the same initial and terminal conditions, on* $[0,t_{1c}[$ *and no more time optimal on* $[0,t]$ *if* $t > t_{1c}$.

5.2. Notation. From now on, we consider a system of the form

$$\dot{v} = X(v) + uY(v), \tag{5.1}$$

where $v = (x,y,z) \in \mathbf{R}^3$, X and Y being analytic vector fields. Let D =det $(Y,[X,Y],[Y,[X,Y]])$, $D' =$ det $(Y,[X,Y],[X,[X,Y]])$, and $D'' =$ det $(Y,[X,Y],X)$. Let us assume that D is not identically 0, and let us denote by $\hat{S}$ the vector field given on $\mathbf{R}^3\backslash\{D=0\}$ by $X - (D'/D)Y$.

LEMMA 5.2. *The singular trajectories satisfying* (H0)–(H3) *are contained in* $\mathbf{R}^3\backslash\{D=0\}$ *and are the nonperiodic solutions of the analytic differential equation* $\dot{v} = \hat{S}(v)$. *The sets* $D''=0$, $DD''>0$, *and* $DD''<0$ *are invariant sets for the solutions of this equation. The hyperbolic* (*resp., elliptic, exceptional*) *trajectories are the solutions contained in* $DD''>0$ (*resp.*, $DD''<0, D''=0$).

Proof. The proof follows from the results of §3.2. A singular extremal has to satisfy the constraints $\langle p, Y(v)\rangle = \langle p,[X,Y](v)\rangle = 0$, and the singular control is given by the dynamic feedback $\hat{u}(v,p)$ defined by (3.7). Now, on $D\backslash\{0\}$, Y and $[X,Y]$ are linearly independent, and since $p \in \mathbf{R}^3\backslash\{0\}$, the feedback $\hat{u}$ is independent of p and is clearly $\hat{u}(v) = -\frac{D'(v)}{D(v)}$.

From Definition 3.3, an exceptional singular trajectory corresponds to a zero energy level, $H=0$. This property is invariant for the singular flow, and for trajectories in $\mathbf{R}^3\backslash\{D=0\}$ it projects onto the set $D''=0$ which contains the exceptional extremals. The lemma is then proved.

5.3. First method of computing conjugate points.

5.3.1. Preliminaries. We briefly recall the method given in [2] to compute conjugate points along a reference singular extremal satisfying (H0)–(H3) in the *hyperbolic* and *exceptional* cases. (In the elliptic case, the trajectory is time *maximizing*, and according to our definition the conjugate point to $\gamma(0)$ is $\gamma(0)$ itself.) We use the fact that the optimality status of γ for (5.1) (no a priori bound is imposed on u) is left invariant under the action of the *feedback group* [1]. Hence, applying to the system a well-defined change of coordinates and a feedback law, we can transform the system near γ into the following systems.

In the *exceptional case*

$$(1+y)\frac{\partial}{\partial x} + u\frac{\partial}{\partial y} + a(t)y^2\frac{\partial}{\partial z} + R(v), \tag{5.2}$$

where $a > 0$ on $[0, T]$, and in the *hyperbolic case*

$$L(t, y, z)\frac{\partial}{\partial x} + z\frac{\partial}{\partial y} + u\frac{\partial}{\partial z} + R(v), \tag{5.3}$$

where L is the quadratic form $a(t)z^2 + 2b(t)yz + c(t)z^2$ and $a < 0$ on $[0, T]$. In both cases the reference trajectory γ is identified to $t \longrightarrow (t, 0, 0)$ and corresponds to the zero control. The remaining terms R are given in [2] and can be neglected to analyze the optimality of γ with respect to all C^0-neighboring trajectories, and R will be supposed to be zero. The corresponding systems are called *models*.

Now, observe that for the models *the input-state mapping* $u(.) \longrightarrow v(t, v_0, u)$, where $v(t, v_0, u)$ is the solution associated with u such that $v(0) = v_0$, *can be explicitly computed*, and the optimality of γ is analyzed as follows.

Fact 5.3. By definition, $\gamma : t \longrightarrow (t, 0, 0)$ is time optimal on $[0, T]$ if $\forall\, t \in]0, T[$, $(t, 0, 0)$ is not accessible from (0,0,0) in a time $t' < t$.

Hence, let $v(t) = (t, 0, 0) + \varphi(t)$, with $\varphi = (\varphi_1, \varphi_2, \varphi_3)$, be a solution of one of the models (5.2) or (5.3); then φ is a solution of

$$\dot\varphi_1 = \varphi_2, \quad \dot\varphi_2 = u, \quad \dot\varphi_3 = a(t)\varphi_2^2 \tag*{(5.2)$'$}$$

or

$$\dot\varphi_1 = L(t, \varphi_2, \varphi_3), \quad \dot\varphi_2 = \varphi_3, \quad \dot\varphi_3 = u. \tag*{(5.3)$'$}$$

Fact 5.4. In the exceptional case, $\gamma : t \longrightarrow (t, 0, 0)$ is time optimal. Indeed, let $0 < t' < t$, and if $v(t') = (t, 0, 0)$, v being the solution of (5.2), we get $\varphi_3(t') = \int_0^{t'} a(s)\varphi_2^2(s)\, ds = 0$. Since $a > 0$, this implies $\varphi_2 \equiv 0$ on $[0, t']$.

Fact 5.5. In the hyperbolic case, the condition $v(t') = (t, 0, 0)$ implies $\varphi_2(t') = \varphi_3(t') = 0$, and clearly γ is time optimal on $[0, T]$ if and only if the functional

$$J(t) = \int_0^t (a(s)\varphi_3^2 + 2b(s)\varphi_2\varphi_3 + c(s)\varphi_2^2)\, ds \tag{5.4}$$

satisfies $J(t) \le 0\, \forall\, t \in]0, T]$ when evaluated on the set of curves φ_2, φ_3 solutions of the equations $\dot\varphi_2 = \varphi_3$, $\dot\varphi_3 = u$ with boundary conditions $\varphi_2(0) = \varphi_3(0) = \varphi_2(t) = \varphi_3(t) = 0$.

Now, from [2], since $u \in \mathbf{R}$, the variable φ_3 can be taken as the control (see the concept of *reduced system* introduced in Definition 3.6), and we have to study the sign of J on the set $\mathcal{C}$ of nontrivial smooth curves φ_2, with $\dot\varphi_2 = \varphi_3$ (control) and satisfying boundary conditions $\varphi_2(0) = \varphi_2(t) = 0$. (The constraints on φ_3 have been relaxed.)

5.3.2. Notation. Let t_{1c} be the first time $0 < t \le T$ such that the maximum of $J(t)$ on $\mathcal{C}$ is zero.

According to classical calculus of variations we have the following lemma (see [9]).

LEMMA 5.6. *If $t < t_{1c}$, then $J(t) < 0$ on $\mathcal{C}$, and if $t > t_{1c}$, then $J(t)$ takes positive and negative values.*

In other words, t_{1c} is the time t such that γ is time optimal on $[0, t]$ if $t < t_{1c}$ and no more optimal if $t > t_{1c}$. Hence $\gamma(t_{1c})$ is the conjugate point to $\gamma(0)$. Now, from [9], the computation of t_{1c} is straightforward.

LEMMA 5.7. *The time t_{1c} is the first time t such that there exists a nontrivial solution φ_2 for Euler–Lagrange equation*

$$\frac{d}{dt}\left(\frac{\partial L}{\partial \dot\varphi_2}\right) - \frac{\partial L}{\partial \varphi_2} = 0, \quad \textit{with}\, \varphi_2(0) = \varphi_2(t) = 0 \tag{5.5}$$

and

$$L(t,\varphi_2,\dot\varphi_2) = a(t)\dot\varphi_2^2 + 2b(t)\varphi_2\dot\varphi_2 + c(t)\varphi_2^2.$$

Conclusion 5.8. Consider system (5.1), $u \in \mathbf{R}$, and let γ be a singular trajectory satisfying (H0)–(H3). Then we have the following:

(i) If γ is elliptic, then γ is not time optimal. (In fact, it corresponds to a slow displacement direction.)

(ii) If γ is exceptional, then γ is time minimal with respect to all solutions of (5.1) contained in a sufficiently small C^0-neighborhood of γ.

(iii) If γ is hyperbolic, then γ is time optimal with respect to all solutions contained in a sufficiently small C^0-neighborhood of γ *until the first conjugate point*. This point can be computed by integrating the linear differential equation (5.5), which is Jacobi's equation associated with γ. This equation can be found by constructing the model. From [2], this construction requires two nonlinear operations: the integrations of the reference trajectory, which has to be identified to $t \to (t,0,0)$, and of the vector field Y, identified to $\frac{\partial}{\partial z}$.

5.4. Conjugate points and the synthesis problem: Intrinsic computation. In the classical calculus of variations, the concept of conjugate point is deeply connected with the concept of extremal field (see [9]). This was used in [22] to define a concept of conjugate point along a reference hyperbolic trajectory in the time-optimal control problem. We briefly recall these results and their connection with our previous analysis.

5.4.1. Preliminaries. Consider a control system in $\mathbf{R}^3$, $\dot v = X + uY$, where $|u| \le M$ and $0 < M \le +\infty$. (Our analysis can be straightforwardly extended to the n-dimensional case.) Recall that $\hat S = X - (D'D)Y$ is the vector field whose nonperiodic trajectories are (H0)–(H3) singular extremals. Let γ be such a reference trajectory defined on $[0,T]$ and corresponding to a control taking its values in $]-M,+M[$, and let us assume that γ is hyperbolic. Let $V(t), t \in [0,T]$ be the solution of the *variational equation*

$$\delta\dot v(t) = \frac{\partial \hat S}{\partial v}(y(t))\delta v(t) \tag{5.6}$$

with $V(0) = Y(\gamma(0))$.

Let $\varepsilon, \varepsilon' = \pm 1$ and g be the mapping $(t_1,t_2,t_3,\varepsilon,\varepsilon') \to \exp\, t_3(X+\varepsilon' MY).\exp\, t_2\hat S.\exp\, t_1(X+\varepsilon MY)(\gamma(0))$, and let $\mathcal{F}$ be its image for $t_2 \in [0,T]$ and t_1, $t_3 \ge 0$, sufficiently small. If $\det\,(V(\gamma(t)), Y(\gamma(t)), \hat S(\gamma(t)))$ is never vanishing on $]0,T]$, then $\mathcal{F}$ is a field about the arc γ in the following sense. There exists a C^0-neighborhood of γ, U, such that every point of U is the image of only one $(t_1,t_2,t_3,\varepsilon,\varepsilon')$. If $M < +\infty$, Moyer proved in [22] that $\mathcal{F}$ is an *optimal field*, i.e., every arc of the field is time optimal with respect to C^0-neighboring trajectories, and from [2] this result is still valid if $M = +\infty$. *In other words, we have constructed the time-optimal synthesis in a neighborhood of the reference trajectory.*

Let us denote by t_{1c} the first $0 < t \le T$ such that $\det\,((V(\gamma(t)), Y(\gamma(t)), \hat S(\gamma(t)))$ vanishes. Next, we compare t'_{1c} with t_{1c} defined in (5.5).

LEMMA 5.9. *We have the following*:

(i) $V(t) \in \operatorname{Span}\{Y(\gamma(t)), [X,Y](\gamma(t))\}$,

(ii) $\det\,(V(t),\, Y(\gamma(t)),\, \hat S(\gamma(t))) = 0$ *for* $t \in]0,T]$ *if and only if* $v(t)$ *and* $Y(\gamma(t))$ *are collinear.*

Proof. By construction, $V(t)$ is the derivative at $\varepsilon = 0$ of the curve $\varepsilon \to \exp\, t\hat S.\exp\, \varepsilon Y(\gamma(0))$ which can be written as $\varepsilon \to \exp\, t\hat S.\exp\, \varepsilon Y.\exp -t\hat S(\gamma(t))$. Now, from the ad-formula, we have for small t

$$V(t) = \sum_{n \geq 0} (-1)^n \frac{t^n}{n!} \mathrm{ad}^n \, \hat{S}(Y)(\gamma(t)),$$

and since γ is a singular arc and X, $[X, Y]$ are linearly independent, we have

$$\mathrm{Span}\,\{\mathrm{ad}^n \, \hat{S}\,(Y);\; n \geq 0\}_{|\gamma} = \mathrm{Span}\,\{Y, [\hat{S}, Y]\}_{|\gamma}.$$

Hence, since Span $\{Y, [\hat{S}, Y]\}_{|\gamma}$ = Span $\{Y, [X, Y]\}_{|\gamma}$, we get $V(t) \in$ Span $\{Y(\gamma(t)), [X, Y]\,(\gamma(t))\}$ for small t and everywhere by analycity. This proves (i). Now, in the hyperbolic case det $(Y, [X, Y], X)$ never vanishes along γ, and then (i) implies (ii).

LEMMA 5.10. $t_{1c} = t'_{1c}$.

Proof. To compute t'_{1c}, one may assume system (5.1) written in the normal form of [2]:

$$X = (1 + Q_1)\frac{\partial}{\partial x} + (z + Q_2)\frac{\partial}{\partial y}, \qquad Y = \frac{\partial}{\partial z},$$

where $Q_i \in (\mathbf{R}[x])[y, z]$ and $\delta^\circ Q$ in $y, z \geq 2$, γ being identified to $t \to (t, 0, 0)$ and corresponding to the zero-control. By computing we get

$$\begin{aligned}
[X, Y] &= -\frac{\partial Q_1}{\partial z}\frac{\partial}{\partial x} - \left(1 + \frac{\partial Q_2}{\partial z}\right)\frac{\partial}{\partial y},\\
[Y, [X, Y]] &= -\frac{\partial^2 Q_1}{\partial z^2}\frac{\partial}{\partial x} - \frac{\partial^2 Q_2}{\partial z^2}\frac{\partial}{\partial y},\\
[X, [X, Y]] &= \chi_1 \frac{\partial}{\partial x} + \chi_2 \frac{\partial}{\partial y},
\end{aligned}$$

where

$$\begin{aligned}
\chi_1 &= \frac{\partial Q_1}{\partial x}\frac{\partial Q_1}{\partial z} + \frac{\partial Q_1}{\partial y}\left(1 + \frac{\partial Q_2}{\partial_z}\right) - \frac{\partial^2 Q_1}{\partial x \partial z}(1 + Q_1) - \frac{\partial^2 Q_1}{\partial y \partial z}(z + Q_2),\\
\chi_2 &= \frac{\partial Q_2}{\partial x}\frac{\partial Q_1}{\partial z} + \frac{\partial Q_2}{\partial y}\left(1 + \frac{\partial Q_2}{\partial z}\right) - \frac{\partial^2 Q_2}{\partial x \partial z}(1 + Q_1) - \frac{\partial^2 Q_2}{\partial y \partial z}(z + Q_2).
\end{aligned}$$

Then

$$\begin{aligned}
D = \det(Y, [X, Y], [Y, [X, Y]]) &= \frac{\partial Q_1}{\partial z}\frac{\partial^2 Q_2}{\partial z^2} - \left(1 + \frac{\partial Q_2}{\partial z}\right)\frac{\partial^2 Q_1}{\partial z^2},\\
D' = \det(Y, [X, Y], [X, [X, Y]]) &= \chi_1\left(1 + \frac{\partial Q_2}{\partial z}\right) - \chi_2 \frac{\partial Q_1}{\partial z}.
\end{aligned}$$

Since $\gamma : t \to (t, 0, 0)$, to compute the variational equation of $\hat{S} = X + \hat{u}Y$, $\hat{u} = -D'/D$ along γ, we need only consider the terms in D and D' which are at most linear in y and z. If we set $Q_1 = a(x)z^2 + 2b(x)yz + c(x)z^2 + o_x(y, z)^2$ and $Q_2 = \bar{a}(x)z^2 + 2\bar{b}(x)yz + \bar{c}(x)z^2 + o_x(y, z)^2$, the relevant terms in $\frac{\partial \hat{u}}{\partial v}$ are given by

$$\frac{cy - a'z - b'y}{[-a(1 + 2\bar{a}z + 2\bar{b}z) + \bar{a}(2az + 2bz)]}.$$

Observe that since the numerator is without a constant term, we can take $\bar{a} = \bar{b} = 0$, i.e., one can assume $Q_2 = 0$, and we have only to compute the Jacobian matrix of $\psi = a^{-1}(cy - a'z - b'y)$. Hence we have

$$\frac{\partial\psi}{\partial x} = 0, \quad \frac{\partial\psi}{\partial y} = \frac{c - b'}{a}, \quad \frac{\partial\psi}{\partial z} = -\frac{a'}{a},$$

and the variational equation along γ is

$$\dot{\delta x} = 0, \quad \dot{\delta y} = \delta z, \quad \dot{\delta z} = \frac{c - \dot{b}}{a}\delta y - \frac{\dot{a}}{a}\delta z. \tag{5.7}$$

On the other hand, the Euler–Lagrange equation corresponding to $L(t, \varphi_2, \dot{\varphi}_2) = a(t)\dot{\varphi}_2^2 + 2b(t)\varphi_2\dot{\varphi}_2 + c(t)\varphi_2^2$ is

$$a\,\ddot{\varphi}_2 + (\dot{b} - c)\varphi_2 + \dot{a}\,\dot{\varphi}_2 = 0, \tag{5.8}$$

which is equivalent to the two last equations of (5.7). Therefore $V(t)$ is collinear to $Y(\gamma(t))$ at time t'_{1c} if and only if there exists a nontrivial solution φ_2 of (5.8), with $\varphi_2(0) = \varphi_2(t'_{1c}) = 0$. Hence, we have proved $t_{1c} = t'_{1c}$.

5.4.2. Curvature. Consider the system

$$\dot{\delta y} = \delta z, \qquad \dot{\delta z} = \frac{c - \dot{b}}{a}\delta y - \frac{\dot{a}}{a}\delta z.$$

It can be written as the second-order differential equation

$$\ddot{\delta y} + \frac{\dot{a}}{a}\,\dot{\delta y} + \frac{\dot{b} - c}{a}\delta y = 0.$$

Every equation of the form

$$\ddot{\delta y} + A\,\dot{\delta y} + B\,\delta y = 0$$

can be transformed into

$$\ddot{Y} + KY = 0$$

if we set $\delta y = CY$, where $C = \exp \int_0^t -\frac{A(s)}{2}\,ds$ and K is given by

$$K = \ddot{C} + A\dot{C} + BC.$$

Computing with $A = \frac{\dot{a}}{a}$, $B = \frac{\dot{b} - c}{a}$, and $a < 0$, we get $C = 1/\sqrt{|a|}$.

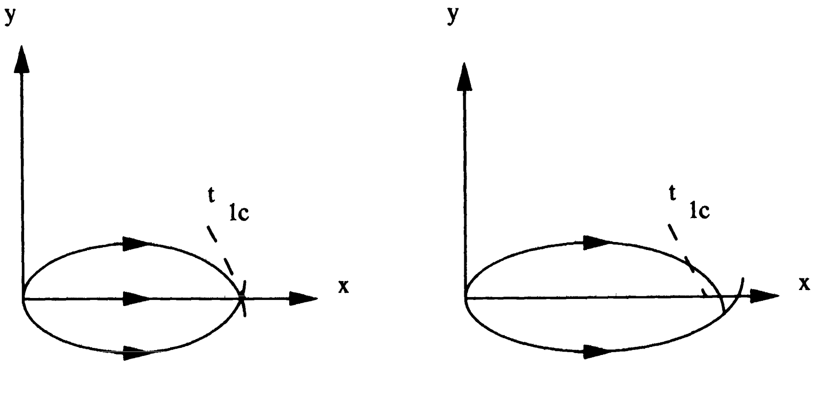
The mapping K defined on $D''D > 0$ corresponds to the concept of *curvature* in Riemannian geometry.

5.4.3. Geometric interpretation. First, let us assume that system (5.1) coincides with the model

$$L(t, y, z)\frac{\partial}{\partial x} + z\frac{\partial}{\partial y} + u\frac{\partial}{\partial z}$$

and the associated reduced system defined in Definition 3.6 is then

$$\dot{x} = 1 + L(t, y, z), \qquad \dot{y} = z, \tag{5.9}$$



(i) (ii)

FIG. 14.

where z is the control variable. By definition of t_{1c}, there exists a curve $\bar{y}$ such that $\dot{\bar{y}}(0) = 1$, $\bar{y}(0) = \bar{y}(t_{1c}) = 0$, $y(t) \neq 0$ on $]0, t_{1c}[$, and $\int_0^{t_{1c}} L(t, \bar{y}, \bar{y})\,dt = 0$. Hence the corresponding solution starting from (0, 0) satisfies $\bar{x}(t_{1c}) = t_{1c}$, and $(\bar{x}, \bar{y})$ intersects $(]0, t_{1c}], 0)$ only at $(t_{1c}, 0)$.

Let $\varepsilon \in \mathbf{R}$ and $\bar{x}_\varepsilon$ be the solutions of (5.9) starting from 0 and corresponding to the controls $z = \varepsilon \bar{y}$. From our previous analysis the family of curves $(\bar{x}_\varepsilon, \varepsilon \bar{y})$ intersects $(]0, t_{1c}], 0)$ only at $(t_{1c}, 0)$ (see Fig. 14(i)), and as in [9] one can show that for the reduced system associated with (5.1), $(t_{1c}, 0)$ is the limiting point of the intersections of the projection of the extremals on the reduced space with the projection of the reference extremal γ (Y can be identified to $\frac{\partial}{\partial z}$ and γ to $t \to (t, 0, 0)$, the reduced space being the (x, y)-space); see Fig. 14(ii).

5.5. Focal points. When we deal with optimal control problems where the terminal manifold is not necessarily a point, the concept of conjugate point has to be generalized to the concept of a *focal point*. To make the application to the batch reactor problem, throughout this section we shall assume that the terminal manifold N is of codimension one and that we are in the flat case, Y tangent everywhere to N.

5.5.1. Generalities. We consider system (5.1) in $\mathbf{R}^3$, with $|u| \leq 1$. The terminal manifold N is assumed to be a regular submanifold of codimension one, its tangent space being given by Span $\{Y, W\}$, where W is a vector field. Let γ be a reference singular (H0)–(H3)-extremal, *hyperbolic*, defined on $[T, 0]$, $T < 0$, and corresponding to an admissible control taking its values in $]-1, +1[$. Let us assume that $\gamma(0) \in N$ and γ satisfies the transversality condition, which can be expressed as follows: $\gamma(0) \in L = \{v \in N;\ \det\,(Y, W, [X, Y])(v) = 0\}$.

We will assume that L is a simple curve. *Now, to define the concept of a focal point along* γ, *we must know the time-optimal synthesis function near* $\gamma(0)$. Hence, we shall assume that γ is time optimal on $[t, 0]$ for the optimal problem with terminal condition on N for t small and the optimal synthesis given in a neighborhood of $\gamma(0)$ by Fig. 15; i.e., the surface formed by the singular arcs arriving at L divides the space into two domains, one in which the optimal feedback is $+1$, and one in which it is -1, the optimal feedback in the surface being the singular control $\hat{u}$. (For a method characterizing such a synthesis, simply see §4.)

DEFINITION 5.11. *Let $V'(t)$ be the solution of the variational equation*

$$\dot{\delta v}(t) = \frac{\partial \hat{S}}{\partial v}(\gamma(t))\delta v(t),$$

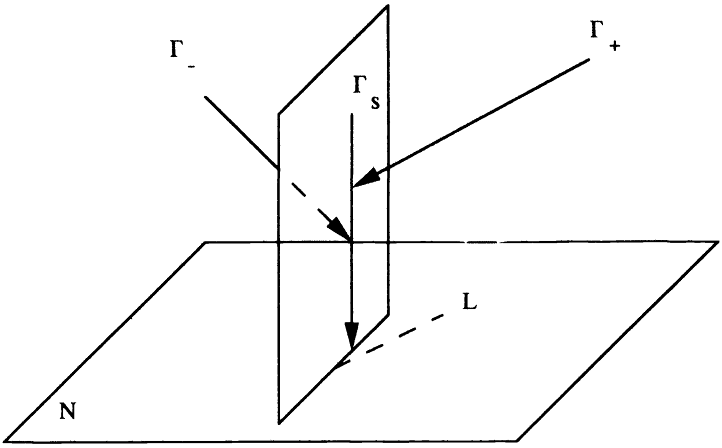


FIG. 15.

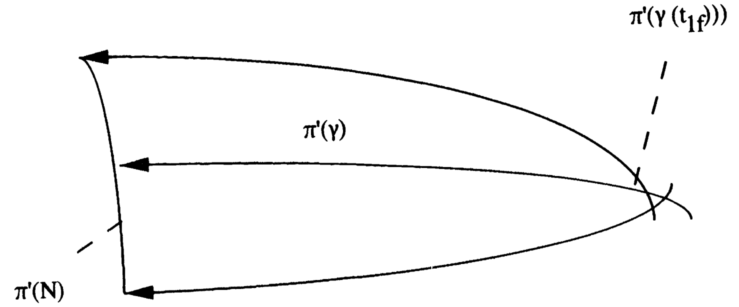


FIG. 16.

with $V'(0) = Z$ *unit tangent vector to* L *at* $\gamma(0)$. *The point* $\gamma(t_{1f})$ *with* $T \le t_{1f} < 0$ *will be called the* (*first*) *focal point about* γ *if* t_{1f} *is the first time* $t < 0$ *such that* det $(V'(t), Y(\gamma(t)), \hat{S}(\gamma(t))) = 0$.

LEMMA 5.12. $V'(t) \in \text{Span}\,\{Y(\gamma(t)), [X,Y](\gamma(t))\}$.

Proof. The vector Z can be written $\lambda_1 Y(\gamma(0)) + \lambda_2[\hat{S},Y](\gamma(0))$, with $\lambda_1, \lambda_2 \in \mathbf{R}$. By definition

$$V'(t) = \frac{d}{d\varepsilon}_{|\varepsilon=0}[\exp t\hat{S}.\exp \varepsilon Z(\gamma(0))], \qquad t < 0$$

and from the ad-formula, the second member of the previous equation belongs to Span $\{\text{ad}^k\hat{S}(Y)(\gamma(t)); k \in \mathbf{N}\}$. This space coincides with Span $\{Y(\gamma(t)), [X,Y](\gamma(t))\}$.

5.5.2. Geometric interpretation. Let $Y = \frac{\partial}{\partial z}$, and let us consider the reduced system. If π' is the projection $(x,y,z) \to (x,y)$, since Y is tangent to the terminal manifold we have the interpretation for the concept of focal points in terms of the behaviors of the projected singular extremals in Fig. 16.

6. Application to the time-optimal control for batch reactors. Now we will compute the optimal synthesis for the problem $\mathcal{P}$ defined in §2 for a sequence of two reactions $X \to Y \to Z$ and where the terminal condition belongs to $N = \{(x,y); y/x = k\}$, k being given, where x and y are the respective (normalized) concentrations of species X and Y. System

(2.3) can be written

$$\frac{dx}{dt} = -vx, \quad \frac{dy}{dt} = vx - \beta v^{\alpha} y, \quad \frac{dv}{dt} = h(v)u, \tag{6.1}$$

where $v = A_1 e^{-E_1/RT}$, $0 < v < A_1$, $h(v) = \frac{R}{E_1} v \ln^2(v/A_1)$, $\alpha = E_2/E_1$, $\beta = A_2/A_1^{\alpha}$, and $w = (x, y, v) \in P$, the physical space defined by $0 < x \leq 1$, $y \geq 0$, $x + y \leq 1$, $0 < v < A_1$, and $y/x < k$.

6.1. Computations. The first step in our analysis is to compute the singular extremals. Using 3.2, we get the following lemma.

LEMMA 6.1. *The singular extremals of order* 2 *for system* (2.3), *written as* $((x, v), (p, \lambda), u) \in \mathbf{R}^n \times \mathbf{R} \times \mathbf{R}^n \times \mathbf{R} \times \mathbf{R}$, *are the solutions of*

$$\begin{aligned} \frac{dx}{dt} &= K(v)x, \quad \frac{dp}{dt} = -pM(v), \quad \frac{dv}{dt} = h(v)\hat{u}(x,p), \\ \hat{u}(x,p) &= \frac{p[M'(v), M(v)]x}{pM''(v)xh(v)}, \qquad \lambda = 0, \end{aligned} \tag{6.2}$$

contained in $\Sigma' = \{(x,p); pM'(v)x = 0\}$. (*Here p is a row vector, M' and M'' are the first and second derivative of M with respect to v, and* $[M', M] = MM' - M'M$.)

PROPOSITION 6.2. *Consider system* (6.1). *If* $\alpha \neq 1$, *all the singular extremals* $w(.)$ *in the physical space p are hyperbolic and the solutions of*

$$\frac{dx}{dt} = -vx, \quad \frac{dy}{dt} = vx - \beta v^{\alpha} y, \quad \frac{dv}{dt} = h(v)\hat{u} \tag{6.3}$$

restricted to P, where the singular control is $\hat{u} = -\dfrac{v^2 x}{h(v)y}$.

Proof. For a system in $\mathbf{R}^3$, $\hat{u}$ is given by $-D'/D$ and the hyperbolic trajectories are contained in $DD'' > 0$, where D, D', and D'' are defined in §5.2. By computing in our case, we get $D = h^4 \alpha(\alpha - 1)\beta v^{\alpha-2} xy$, $D' = h^3 \beta(\alpha - 1)v^{\alpha} x^2$, and $D'' = h^2 \beta v^{\alpha} xy(\alpha - 1)$. This proves the proposition.

6.2. Projected system. System (5.3) is left invariant by the transformations $(x, y, v) \to (\lambda x, \lambda y, v)$, $\lambda \in \mathbf{R} \backslash \{0\}$. Hence it can be projected onto $\mathbf{P}^1 \times \mathbf{R}$, where $\mathbf{P}^1$ = one-dimensional projective space. More precisely, since x never vanishes in the physical space, in the coordinates $(x, z = y/x, v)$ it becomes

$$\frac{dx}{dt} = -vx, \quad \frac{dz}{dt} = v - \beta v^{\alpha} z + vz, \quad \frac{dv}{dt} = h(v)u. \tag{6.4}$$

The system

$$\frac{dz}{dt} = v - \beta v^{\alpha} z + vz, \qquad \frac{dv}{dt} = h(v)u \tag{6.5}$$

is called the *projected system.*

Now, we can project differential equation (6.3) onto

$$\frac{dz}{dt} = v - \beta v^{\alpha} z + vz, \qquad \frac{dv}{dt} = -\frac{v^2}{\alpha z}. \tag{6.6}$$

But from §3.5, not every solution of this equation corresponds to a singular extremal of the projected system. Indeed, let (X, Y) be a system on $\mathbf{R}^2$. The singular extremals are contained in the set

$$S = \{x;\ \det(Y(x), [X, Y](x)) = 0\},$$

the singular control being given by $\hat{u}$. By computing for the planar system (6.5) we get the following lemma.

LEMMA 6.3. *The singular extremals for system* (6.5) *are contained in* $S = \{(z,v);\ z(\alpha\beta v^{\alpha-1}-1) = 1\}$, *and the singular control is given by* $\hat{u} = -(v^2/h(v)\alpha z)$. *The differential equation describing the evolution on a singular arc is*

$$\frac{dv}{dt} = \frac{v^2}{\alpha}(1-\alpha\beta v^{\alpha-1}). \tag{6.7}$$

6.3. Conjugate points. From now on we shall assume $\alpha > 1$, which is the physical interesting situation. All the singular trajectories satisfy (H0)–(H3) and are hyperbolic, and we have to compute the first conjugate point along a reference singular trajectory γ, using the algorithm in §5.4. The variational equation associated with (6.3) is

$$\begin{aligned} \dot{\delta x} &= -v\delta x - x\delta v, \\ \dot{\delta z} &= v(1-\beta v^{\alpha-1})\delta z + \psi(z,v)\delta v, \\ \dot{\delta v} &= \frac{v^2}{\alpha z^2}\delta z - \frac{2v}{\alpha z}\delta v, \end{aligned} \tag{6.8}$$

where $\psi(z,v) = 1 + z(1-\alpha\beta v^{\alpha-1})$ and $\psi = 0$ is the set S. Let $Q = S \times \mathbf{R}$ and $G = \{(x,z,v);\ \dot{z} = 0\}$.

To compute the conjugate points, we must integrate (6.8) with the initial condition $\delta x(0) = \delta z(0) = 0$ and $\delta v(0) = 1$. We have to distinguish two cases.

LEMMA 6.4. *A singular trajectory in* $P \cap Q$ *is without conjugate points.*

Proof. We integrate (6.8) with the initial condition (0,0,1) and $\psi = 0$. Hence we have

$$\delta z = 0, \qquad \delta v = \exp \int_0^t -\frac{2v}{\alpha z}\, ds,$$

and clearly $\delta x(t) < 0$ for all $t > 0$. Therefore the condition $\delta x(t_{1c}) = 0$ cannot be satisfied.

6.3.1. Singular trajectories not contained in $P \cap Q$. Let us write (6.3) as $\dot{w} = X + uY$, and let us denote by π the projection $(x,z,v) \to (z,v)$, X^π and Y^π being the vector fields $d\pi(X)$ and $d\pi(Y)$, respectively. Let γ be a singular trajectory in $P \backslash Q$ and $V(t)$ be the solution of the variational equation along γ, with $V(0) = Y(\gamma(0))$. From Lemma 5.9, $V(t)$ can be written $\lambda_1(t)Y(\gamma(t)) + \lambda_2(t)[X,Y](\gamma(t))$, and its projection on the space (z,v) is

$$d\pi(V(t)) = \lambda_1(t)Y^\pi(\pi(\gamma(t)) + \lambda_2(t)[X^\pi, Y^\pi](\pi(\gamma(t))).$$

Now, t_{1c} is the conjugate time if it is the first t such that $\lambda_2(t) = 0$. Since on $P \backslash Q$, $V(t)$ is collinear to $Y(\gamma(t))$ if and only if $d\pi(V(t))$ is collinear to $Y^\pi(\gamma(t))$, we have proved the following lemma.

LEMMA 6.5. *The time* t_{1c} *is the conjugate time if and only if* t_{1c} *is the first t such that the solution of*

$$\begin{aligned} \dot{\delta z} &= v(1-\beta v^{\alpha-1})\delta z + \psi(z,v)\delta v, \\ \dot{\delta v} &= \frac{v^2}{\alpha z^2}\delta z - \frac{2v}{\alpha z}\delta v \end{aligned} \tag{6.9}$$

passing through (0,1) *at* $t = 0$ *is such that* $\delta z(t_{1c}) = 0$.

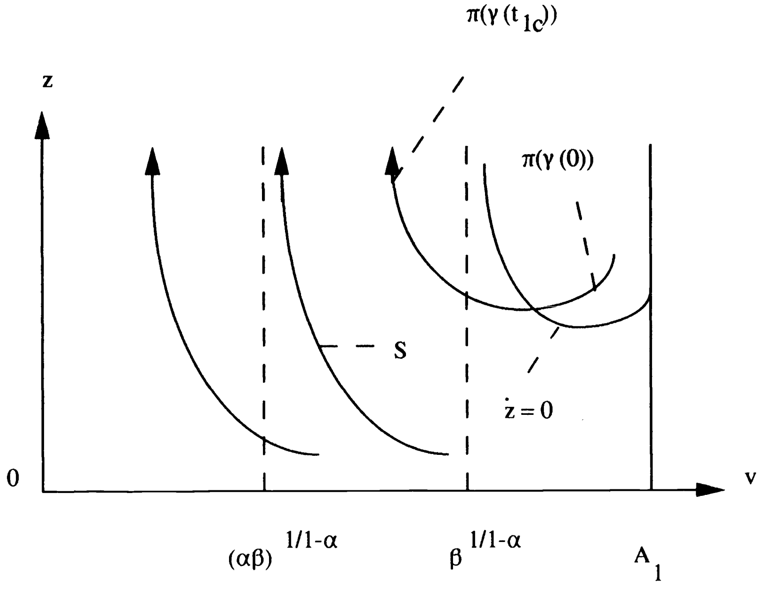


FIG. 17.

6.3.2. Computations. By setting $J = \delta z/\psi(z,v)v$, equation (6.9) can be written in the canonical form $\ddot{J} + (K o \gamma)J = 0$, where the curvature is

$$K = \frac{(\alpha - 1)\beta v^{\alpha+1}}{\alpha z}.$$

From Sturm's theorem [15], we must find L such that $0 < L < K$ to guarantee the existence of a conjugate time on $[0, \pi/\sqrt{L}]$. Since along a singular trajectory γ in P, $z(t) \longrightarrow +\infty$ when $t \longrightarrow +\infty$, such a lower bound $L > 0$ doesn't exist in our case and we have to use numerical simulations to compute conjugate points. If $\beta^{1/(1-\alpha)} < A_1$, they show the existence of conjugate point for singular arcs such that $\gamma(0) \in \{\dot{z} < 0\}$ (see Fig. 17).

6.4. Optimal synthesis.

6.4.1. Preliminaries. From §3.5, to solve problem $\mathcal{P}$ it is sufficient to solve the projected problem $\mathcal{P}'$: minimize the time duration to reach the target $N = \{(z,v);\ z = k,\ k \text{ fixed}\}$ for the solution of

$$\frac{dz}{dt} = v - \beta v^{\alpha} z + vz, \qquad \frac{dv}{dt} = -\frac{v^2}{\alpha z},$$

where $u \in [u_-, u_+]$. Moreover we shall assume $u_- < 0 < u_+$, $\alpha > 1$, $A_1 > \beta^{1/(1-\alpha)}$, and $-v_1^2/h(v_1)\alpha k \in]u_-, u_+[$, where v_1 is defined as the v-coordinate of $P_1 = (v_1, k)$ intersection of S with N. (The other cases can be easily deduced from our analysis.)

6.4.2. Singular arc. A singular trajectory belongs to the set $S = \{(z,v);\ z(\alpha\beta v^{\alpha-1} - 1) = 1\}$, and the singular control is defined by $\hat{u} = -v^2/h(v)\alpha z$ and has to belong to $[u_-, u_+]$ to be admissible. By computing, we get $\hat{u} < 0$ and $\frac{\partial \hat{u}}{\partial v} < 0$. Moreover v decreases along a singular arc and when $v \longrightarrow A_1^-$, $\hat{u} \longrightarrow -\infty$. Since at $P_1 = S \cap N$, $\hat{u}$ is admissible by hypothesis we have Fig. 18, where P_2 is the unique point such that $\hat{u} = u_-$ and Γ_s denotes the maximal admissible singular arc.

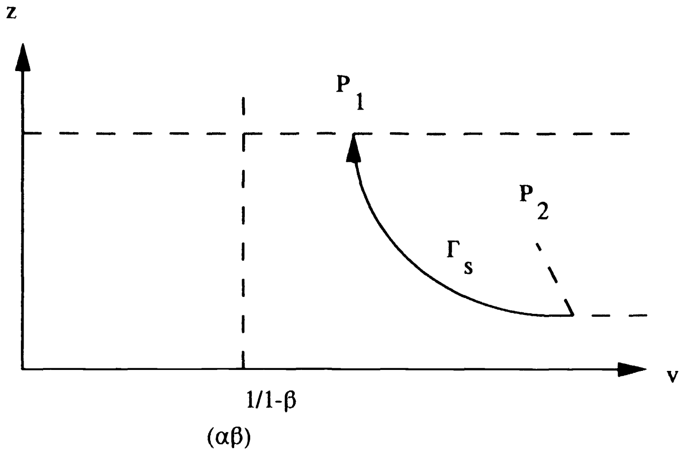


FIG. 18.

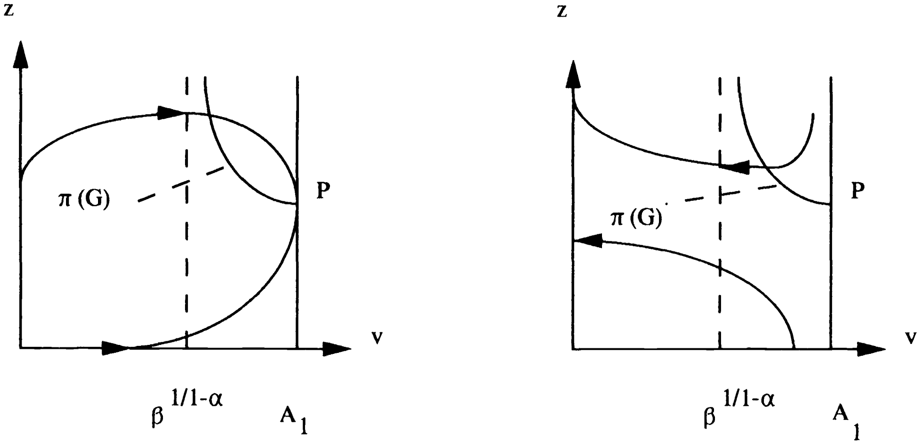


FIG. 19.

6.4.3. Regular arcs. Now, we have to analyze the behaviors of trajectories corresponding to the constant control $u = u_-$ or u_+. First, observe that $h(v) > 0$ if $0 < v < A_1$. Hence, the sign of $\dot{v}$ is given by u_- or u_+. Now, $\dot{z}$ vanishes for $z = (\beta v^{\alpha-1} - 1)^{-1}$, whose graph is denoted by $\pi(G)$. The singular points of the vector fields associated with u_- and u_+ in $0 \leq v \leq A_1$ are the points of the line $v = 0$ and the point $P = ((\beta A_1^{\alpha-1} - 1), A_1)$, which belongs by assumption to $z > 0$. A straightforward analysis gives the phase portraits in Fig. 19.

6.4.4. Switching function. One major problem in optimal control is *estimating the number of switchings* of an optimal control. In our case, it is possible by using the following analysis. The adjoint equations associated with the system, with $p = (p_1, p_2)$, are

$$\frac{dp_1}{dt} = p_1 v(\beta v^{\alpha-1} - 1), \qquad \frac{dp_2}{dt} = p_1 v(\alpha\beta v^{\alpha-1} - 1 - z) - p_2 \frac{dh(v)}{dv} u.$$

Let (γ, p, u) be a smooth extremal defined on $[0, T]$, p being nonzero by assumption. The switching function $\Phi(t) = p_2(t)h(v(t))$ evaluated along this extremal satisfies, from (3.9),

$$\dot{\Phi}(t) = p_1 h(v)[z(\alpha\beta v^{\alpha-1} - 1) - 1],$$
$$\ddot{\Phi}(t) = p_1 h(v)[\beta v^{\alpha}(\alpha - 1) + u\alpha(\alpha - 1)\beta v^{\alpha-2} z h(v)] + u\frac{dh}{dv}(v)\dot{\Phi}.$$

The set $\dot{\Phi}(t) = 0$ plays an important role when the switching points are computed. Observe that if p_1 is nonzero, its projection on the state space is the curve S. We have the following lemma.

LEMMA 6.6. *Let us assume* $\Phi(0) = \Phi(T) = 0$. *Then the extremal* γ *meets* S *for a* $t \in]0, T[$.

Proof. By definition p never vanishes and $p_2(0) = 0$. Now, the sign of p_1 is constant on $[0, T]$ and p_1 never vanishes. Since $\Phi(0) = \Phi(T)$, then there exists $t \in]0, T[$ such that $\dot{\Phi}(t) = 0$. Since p_1 never vanishes, at t we have $z(\alpha\beta v^{\alpha-1} - 1) = 1$.

LEMMA 6.7. *Let us assume that* (γ, p, u) *is such that* $u = u_+, p_1 > 0$ *and* $\Phi(T) = 0$. *Then* $\Phi(0)$ *is nonzero.*

Proof. Let $\varphi(t) = \dot{\Phi}(t)/h(v(t))$ evaluated along the given extremal. We have

$$\dot{\varphi}(t) = p_1[\beta v^{\alpha}(\alpha - 1) + u_+\alpha(\alpha - 1)\beta v^{\alpha-2} z h(v)].$$

Hence, sign $\dot{\varphi}$ = sign $p_1 > 0$ and φ is a strictly increasing function. Now since (γ, p, u_+) is an extremal, if T is a switching time, from §3.6 we must have $\dot{\Phi}(T) \le 0$. Hence $\varphi(0) < \varphi(T) \le 0$. Let us assume $\Phi(0) = 0$; then again we must have $\dot{\Phi}(0) \ge 0$. This contradicts $\varphi(0) = \dot{\Phi}(0)/h(v(0)) < 0$.

6.4.5. Synthesis. Now, we can show the optimal synthesis in Fig. 20.

First, observe that by [19, Thm. 4, p. 259], there exists an optimal controller in the family of all measurable mappings with values in $[u_-, u_+]$, provided that the system is restrained to the compact set $0 \le z \le k$, $0 \le v \le A_1$, and the target is the compact set $z = k$ and $0 \le v \le A_1$.

To compute the synthesis we proceed as follows. First, we use the local classification in §4 to obtain the synthesis function near the target. We are in the flat case. The two singularities of the analysis are at P_1, where an optimal singular arc meets the target, and the point P_3, where both are Γ_+ and Γ_- are tangent to the terminal manifold, the respective synthesis being given by Propositions 4.13 and 4.16.

Now, from the analysis in §4, at the terminal point, the adjoint variable can be taken as $p = (1, 0)$. Since we are in the flat case, all the points of N are virtual switching points. Hence from Lemma 6.7 an extremal arc Γ_+ hitting the target is not allowed to switch at the initial point, and from Lemma 6.6 an extremal arc Γ_- hitting the target has to meet S to switch at the initial time. The point $P_2 = (v_2, *)$ is the point on S such that an arc Γ_- is tangent to S, and if $v > v_2$, then $\hat{u} < u_-$. This will cause the existence of optimal laws with two switchings, one on the singular arc Γ_s and the other on a curve C. As in Lemma 4.2, one can show that this curve cannot be an arc Γ_- and is in fact contained strictly in the acute domain limited by S and the arc Γ_- passing through P_2.

The optimal synthesis for the original problem $\mathcal{P}$ is obtained by adding the x-variable. Observe that it is without a conjugate point, since a singular arc in $\mathbf{R}^2$ is without a conjugate point, and this is in accordance with the analysis of §6.3.

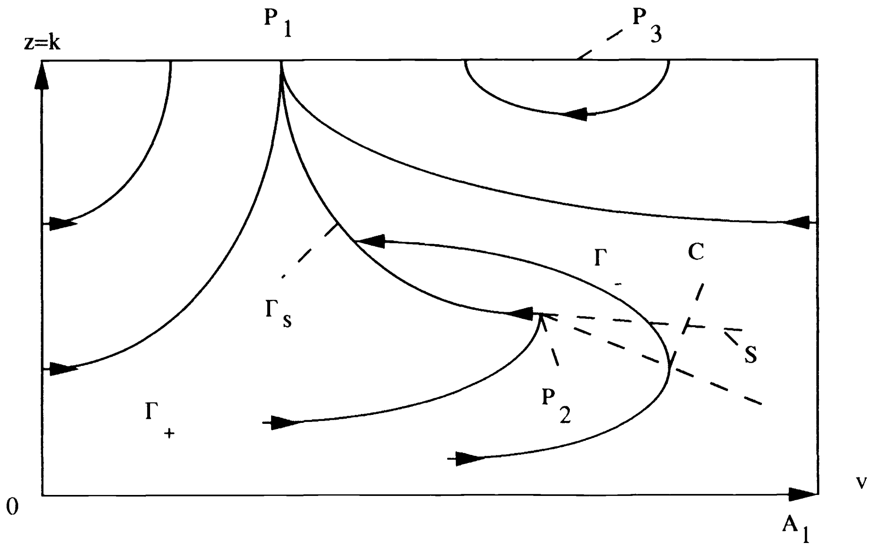


FIG. 20.

REFERENCES

[1] B. BONNARD, *Feedback equivalence for nonlinear systems and the time optimal control problem*, SIAM J. Control Optim., 29 (1991), pp. 1300–1321.

[2] B. BONNARD AND I. KUPKA, *Théorie des singularités de l'application entrée/sortie et optimalité des trajectoires singulières dans le problème du temps minimal*, Forum Math., 5 (1993) pp. 111–159.

[3] B. BONNARD, J. P. GAUTHIER, AND J. DE MORANT, *Geometric time optimal control for batch reactors*, part I, in Analysis of Controlled Dynamical Systems, B. Bonnard, B. Bride, J. P. Gauthier, I. Kupka, eds., Birkhäuser, Boston, 1991; part II, in Proc. 30th IEEE/CDC Conf., Brighton, 1991.

[4] B. BONNARD AND M. PELLETIER, *Time Minimal Synthesis for Planar Systems in the Neighborhood of a Terminal Manifold of Codimension One*, Laboratoire de Topologie de Dijon, 1992, preprint.

[5] I. EKELAND, *Discontinuité des champs hamiltoniens et existence de solutions optimales en calcul des variations*, Pub. IHES, No. 47, 1977, pp. 1–32.

[6] J. EVANGELISTA AND S. KATZ, *Best temperature schedules in batch reactors*, Indust. Engn. Chem., 60 (1968), pp. 24–33.

[7] M. FEINBERG, *Chemical reaction network structure and stability of complex isothermal reactions*, Chem. Engn. Sci., 42 (1987), pp. 2229–2268.

[8] B. FRÉMAUX, *Eléments de cinétique et de catalyse, Technique et documentation*, Lavoisier, Paris, 1989.

[9] I. M. GELFAND AND S. V. FOMIN, *Calculus of Variations*, Prentice-Hall, Englewood Cliffs, NJ, 1963.

[10] H. HERMES, *Local controllability and sufficient conditions in singular problems*, J. Differential Equations, 20 (1976), pp. 213–232.

[11] J. HICKS, A. MOHAN, AND RAY, *The optimal control of polymerization reactors*, Canad. J. Chem. Engn., 47 (1969), pp. 590–597.

[12] C. G. HILL, *An Introduction to Chemical Engineering Kinetics and Reactor Design*, John Wiley, New York, 1977.

[13] M. KARAPETIANZ, *Initiation à la théorie des phénomènes chimiques*, Mir, Moscow, 1978.

[14] W. KLINGENBERG, *Riemannian Geometry*, De Gruyter studies in Mathematics 1, De Gruyter, Berlin, 1982.

[15] S. KOBAYASHI, *On conjugate and cut loci*, in Studies in Global Geometry, Vol. 4, S. S. Chern, ed., Prentice-Hall, Englewood Cliffs, NJ, 1967.

[16] N. C. T. KOPPERT, *The Numerical Optimisation of Chemical Processes*, Technical Report, Delft University of technology, 1991.

[17] A. J. KRENER, *The higher-order maximal principle and its applications to singular extremals*, SIAM J. Control Optim., 15 (1977), pp. 256–293.

[18] I. KUPKA, *Geometric theory of extremals in optimal control problems. I. The fold and Maxwell cases*, Trans. Amer. Math. Soc., 299 (1973), pp. 225–243.

[19] E. B. L EE AND L. MARKUS, *Foundations of Optimal Control Theory*, John Wiley, New York, 1967.

[20] O. LEVENSPIEL, *Chemical Reaction Engineering*, John Wiley, New York, 1972.

[21] J. DE MORANT, *Contrôle en temps minimal des réacteurs chimiques discontinus*, Ph.D. thesis, Université de Rouen, Rouen, France, 1992.

[22] H. G. MOYER, *Sufficient conditions for a strong minimum in singular problems*, SIAM J. Control Optim., 11 (1973), pp. 620–636.

[23] A. MOUSTAFA, *Simulateur, régulateur et protocole de communication PC/VAX pour la mise en oeuvre d'une commande optimale d'un réacteur batch—Etude et implémentation sur site réel*, Masters thesis, Lab. D'Automatique de Grenoble, Grenoble, France, 1991.

[24] H. POINCARÉ, *Sur les lignes géodésiques des surfaces convexes*, Trans. Amer. Math. Soc., 6 (1905), pp. 237–274.

[25] L. PONTRYAGIN, V. BOLTYANSKI, R. GAMKRELIDZE, AND E. MISCHENKO, *Théorie mathématique des processus optimaux*, Mir, Moscow, 1974.

[26] H. SCHÄTTLER, *The local structure of time-optimal trajectories in dimension* 3 *under generic conditions*, SIAM J. Control Optim., 26 (1988), 899–918.

[27] H. J. SUSSMANN, *The structure of time-optimal trajectories for single-imput systems in the plane: The C^∞ nonsingular case*, SIAM J. Control Optim., 25 (1977), pp. 433–465.

[28] ———, *Regular synthesis for time-optimal control for single-imput real analytic systems in the plane*, SIAM J. Control Optim., 25 (1987), pp. 1145–1162.

[29] ———, *Envelopes, conjugate points and optimal bang-bang extremals*, in Proceedings of the Conference of Nonlinear Systems, Paris 1985, Fliess M. and Hazewinkel M., eds., Reidel, Dordrecht, 1986.

[30] H. J. SUSSMANN AND G. TANG, *Shortest Paths for the Reed–Shepp Car: A Worked Out Example of the Use of Geometric Techniques in Nonlinear Optimal Control*, Report Sycon 91-10, Rutgers University, New Brunswick, NJ, 1991.

[31] V. S. VARADARAJAN, *Lie Groups, Lie Algebras and Their Representations*, Prentice-Hall, Englewood Cliffs, NJ, 1974.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 33, No. 5, pp. 1312–1325, September 1995

© 1995 Society for Industrial and Applied Mathematics
002

NEUMANN BOUNDARY VALUE PROBLEMS FOR SECOND-ORDER ORDINARY DIFFERENTIAL EQUATIONS ACROSS RESONANCE*

WANG HUAIZHONG† AND LI YONG†

Abstract. There have been some applications of optimal control theory to boundary value problems for ordinary differential equations. Among previous works, the best lengths of intervals on which the boundary value problem admits a solution are estimated by Pontryagin's maximum principle. Hence such approaches are local and the presented conditions are actually not across points of resonance as in the Lazer–Leach condition. Here we consider the existence–uniqueness problem in a class of Neumann boundary value problems for second-order ordinary differential equations probably across several points of resonance. By the optimal control theory method and a careful analysis, we obtain some global optimality results about the existence and uniqueness of solutions for boundary value problems.

Key words. optimal control, Neumann boundary value problems, existence and uniqueness, across points of resonance

AMS subject classifications. 34B, 49B, 49K

1. Introduction. As is well known, the Neumann condition is one of the basic boundary conditions appearing in mathematical physics (for example, equilibrium problems concerning beams, columns, or strings; fluid flow problems; and heat transfer problems). This kind of boundary value problems (BVP) has the following standard form:

$$y'' + f(t, y, y') = 0, \quad y'(a) = A, \quad y'(b) = B, \tag{p}$$

where $A,\ B \in R^n$, $f : [a, b] \times R^n \times R^n \to R^n$. The Neumann BVP (p) differs from a corresponding initial value problem or Dirichlet boundary value problem in that a Lipschitz condition for f does not guarantee the existence nor the uniqueness of solutions even for the simple $f \equiv f(t)$ unless $\int_a^b f(s)\, ds = B - A$. Furthermore, Green's function of the problem $y'' = 0$, $y'(a) = 0$, $y'(b) = 0$ does not exist. This shows that one cannot utilize Green's function in reformulating (p) as an integral equation, unlike the Dirichlet boundary value problem. Therefore, it is more difficult to find general conditions for existence and uniqueness. Under suitable monotonicity conditions or nonresonance conditions, some nice existence or uniqueness theorems or methods for (p) have been presented (see, for example, [1]–[9], [16]–[18], [23], [24], and [26]).

In this paper we are concerned with the following Neumann BVP:

$$y'' + f(t, y) = 0, \tag{1.1}$$

$$y'(0) = A, \qquad y'(1) = B, \tag{1.2}$$

where $A,\ B \in R$ and $f : [0, 1] \times R \to R$ is continuous and satisfies some additional conditions.

In general, the set $\{n^2\pi^2\}_0^\infty$ is called the set of points of resonance for BVP (1.1)–(1.2). If we assume that $f(t, y)$ is continuously differentiable with respect to y, then applying previous works to BVP (1.1)–(1.2), the results that we can obtain are only the following:

(i) If there is an $a > 0$ such that

$$-f'_y(t, y) \left(f'_y = \frac{\partial f}{\partial y} \right) \ge \alpha \quad \text{on } [0, 1] \times R,$$

then BVP (1.1)–(1.2) has a unique solution.

* Received by the editors March 10, 1993; accepted for publication (in revised form) April 17, 1994.

† Department of Mathematics, Jilin University, Changchun 130023, People's Republic of China.

(ii) If there are $\lambda, \mu > 0$ such that

$$N^2\pi^2 < \lambda \leq f'_y(t,y) \leq \mu < (N+1)^2\pi^2 \quad \text{on } [0,1] \times R, \tag{1.3}$$

where N is some nonnegative integer, then BVP (1.1) – (1.2) has a unique solution.

The above condition (1.3) is the usual Lazer–Leach-type condition (see [17]). However, if for some $B > \pi^2$

$$0 \leq f'_y(x,y) \leq B \quad \text{on}, [0,1] \times R, \tag{1.4}$$

then when does BVP (1.1) – (1.2) have a unique solution? The condition (1.4) is usually called one crossing points of resonance.

In this paper, by the optimal control theory method, we prove two optimality results on the existence and uniqueness of BVP (1.1) – (1.2). Our main result is the following theorem.

THEOREM A. *Assume that the following conditions are fulfilled*:

(i) $f(t,y)$ *and* $f_y(t,y)$ *are continuous on* $[0,1] \times R$.

(ii) *For some* $B \geq \pi^2$,

$$0 \leq f_y(t,y) \leq \beta(t) \leq B \quad \textit{on}\, [0,1] \times R,$$

where $\beta \in L[0,1]$ *and satisfies* $\int_0^1 \beta(t)\, dt < B\alpha$. *Here* α *is the minimal positive root of the equation*

$$\cos\left(\sqrt{B}\frac{x}{2}\right) = \frac{1}{2}\sqrt{B}(1-x)\, \sin\left(\sqrt{B}\frac{x}{2}\right).$$

(iii) $\int_0^1 f(t,0)\, \mathrm{d}t = 0$ *and*

$$\text{meas}\, \{t \in [0,1],\ f_y(t,x(t)) > 0\} > 0,$$

for each $x \in C([0,1],R)$.

Then BVP (1.1) – (1.2) *has a unique solution.*

Here the optimality means that given $B \geq \pi^2$, there exists a $u \in C([0,1],R)$ with

$$0 \leq u(t) \leq B \quad \text{for}\, t \in [0,1], \qquad \int_0^t u(t)\, \mathrm{d}t > B\alpha$$

such that the BVP

$$y'' + u(t)y = 0, \qquad y'(0) = y'(1) = 0$$

has nontrivial solutions.

It should be pointed out that Mawhin, Ward, and Willem [19] have presented a sufficient and necessary condition of solvability for the Neumann problem relative to semilinear elliptic equations across resonance by means of the variational method.

Some interesting applications of the control theory method to several BVPs for ordinary differential equations have been presented by the authors of [8], [11]–[15], [20]–[22], and [25]. However, all of these papers do not deal with the Neumann BVPs, and the conditions of the theorems established in the papers are all nonresonant.

The plan of this paper is the following: §2 deals with an optimal control problem for a linear BVP. There, using Pontryagin's maximum principle, we prove the existence of the optimal control for the problem. In particular, we find an explicit expression of the optimal control, which is vital to our discussion. Finally in §3, applying the results of §2 and Schauder's fixed-point theorem, we give the proofs of Theorem A and some other results.

2. Linear problem. Consider the following linear BVP:

$$y'' + u(t)y = 0, \qquad y'(0) = y'(1) = 0, \tag{2.1}$$

where $u \in L[0, 1]$ and satisfies $0 \leq u(t) \leq B$, where $B \geq \pi^2$.

Choose a suitable admissible set Ω as follows:

$$\Omega = \{u \in L[0, 1] : 0 \leq u(t) \leq B,\ \text{meas}\,\{t \in [0, 1] : u(t) > 0\} > 0 \\ \text{and for } u = u(t),\ \text{BVP (2.1) has nontrivial solutions}\}.$$

We are to seek a function $u^* \in \Omega$ such that $u^*(t)$ minimizes the functional $J[u]$ defined by

$$J[u] = \int_0^1 u(t)\,\mathrm{d}t, \qquad u \in \Omega;$$

that is,

$$J[u^*] = \min_{\Omega} J[u]. \tag{2.2}$$

Thus the problem (2.1) – (2.2) is an optimal control problem.

We need the following lemmas.

LEMMA 1. *Problem* (2.1) – (2.2) *has a solution*; *that is*, *the optimal control function exists*.

Proof. Since $u(t) \equiv \pi^2 \in \Omega$, Ω is not empty. Obviously,

$$\mu_0 = \inf_{\Omega} J[u] \in [0, B].$$

Hence there exists a minimizing sequence $\{u_n(t)\} \subset \Omega$ such that

$$J[u_n] \to \mu_0 (n \to \infty).$$

For each $u = u_n(t)$, BVP (2.1) has a solution $y_n(t)$ with

$$\|y_n\| = \max_{[0,1]} |y_n(t)| + \max_{[0,1]} |y_n'(t)| = 1$$

by the definition of Ω. Then $\{y_n''(t)\}$ is uniformly essentially bounded on [0,1]. These imply that $\{y_n(t)\}$ and $\{y_n'(t)\}$ are uniformly bounded and equicontinuous on [0,1]. By the Arzela–Ascoli theorem, passing to a subsequence if necessary, we may assume

$$y_n \to y_0, \quad y_n' \to z_0 \quad (n \to \infty)$$

uniformly on [0,1] for suitable $y_0, z_0 \in C([0, 1], R)$. Since $\{u_n\} \subset L^2[0, 1]$ and is uniformly bounded on [0,1], we may assume that $u_n(t)$ is weakly convergent to $u^* \in L^2[0, 1]$ as $n \to \infty$. By the Hahn–Banach theorem, we see $u^* \in [0, B]$. Consequently,

$$y_0(t) = y_0(0) + \int_0^t z_0(s)\,\mathrm{d}s,$$
$$z_0(t) = z_0(0) + \int_0^t [-u^*(s)y_0(s)]\,\mathrm{d}s$$

for $t \in [0, 1]$. Obviously,

$$y_0'(0) = y_0'(1) = 0, \qquad \|y_0\| = 1. \tag{2.3}$$

These imply that $y_0(t)$ is a solution of BVP (2.1) for $u = u^*(t)$. Note that for each n,

$$\operatorname{meas}\{t \in [0,1] : u_n(t) > 0\} > 0.$$

Therefore, for each n, $y_n(t)$ has a zero in [0,1], which shows that $y_0(t)$ has a zero in [0,1]. From this, (2.3), and the uniqueness for initial value problems it follows that

$$\operatorname{meas}\{t \in [0,1] : u^*(t) > 0\} > 0.$$

To summarize, we have that $u^* \in \Omega$, which completes the proof of the lemma.

LEMMA 2. *Let $u_1 \in \Omega$. If the BVP*

$$y'' + u_1(t)y = 0, \qquad y'(0) = y'(1) = 0$$

has a nontrivial solution $y_1(t)$ such that for some $t_1 \in (0,1)$, $y_1'(t_1) = 0$, then there exists a $u_2 \in \Omega$ such that the BVP

$$y'' + u_2(t)y = 0, \qquad y'(0) = y'(1) = 0$$

has a nontrivial solution $y_2(t)$ with

$$\begin{aligned} |y_2'(t)| &> 0 \quad \text{on } (0,1) \text{ and} \\ \int_0^1 u_2(t)\,\mathrm{d}t &< \int_0^1 u_1(t)\,\mathrm{d}t. \end{aligned} \tag{2.4}$$

Proof. Since $u_1 \in \Omega$, we have

$$\operatorname{meas}\{t \in [0,1] : u_1(t) > 0\} > 0.$$

Hence it is easy to prove that $y_1(t)$ has a zero t_2 in (0,1). By the uniqueness of initial value problems, we get $y_1'(t_2) \neq 0$. Note that $t_1 \in (0,1)$ and $y_1'(t_1) = 0$. Therefore, there exists $[a,b] \subset [0,1]$ such that

$$0 < b - a < 1, \quad t_2 \in (a,b), \quad y_1'(a) = y_1'(b) = 0,$$

and

$$|y_1'(t)| > 0 \quad \text{on } (a,b).$$

Set $t = (b-a)s + a$. Then $y_2(s) = y_1((b-a)s + a)$ is a nontrivial solution of the BVP

$$\begin{aligned} &\frac{\mathrm{d}^2y}{\mathrm{d}s^2} + (b-a)^2 u_1((b-a)s + a)y = 0, \\ &y'(0) = y'(1) = 0. \end{aligned}$$

Take $u_2(s) = (b-a)^s u_1((b-a)s + a)$. Then

$$\int_0^1 u_2(t)\,\mathrm{d}t = (b-a)\int_a^b u_1(t)\,\mathrm{d}t < \int_0^1 u_1(t)\,\mathrm{d}t,$$

which completes the proof.

By Lemma 2, we get the following lemma.

LEMMA 3. *If $u^*(t)$ is an optimal control of problem* (2.1)–(2.2), *then for $u = u^*(t)$ every nontrivial solution $y^*(t)$ of BVP* (2.1) *satisfies*

$$y^{*'}(t) \neq 0 \quad \text{on } (0,1).$$

The main result of this section is the following.

THEOREM 1. *Let $B > \pi^2$. Then the problem* (2.1) – (2.2) *has a unique optimal control $u^* \in \Omega$. In addition, $u^*(t)$ has the following form*:

$$u^*(t) = \begin{cases} B, & 0 \le t \le \frac{1}{2}\alpha, \\ 0, & \frac{1}{2}\alpha < t < 1 - \frac{1}{2}\alpha, \\ B, & 1 - \frac{1}{2}\alpha \le t \le 1, \end{cases}$$

where $\alpha = \alpha(B)$ is the minimal positive root of the equation

$$\cos\left(\sqrt{B}\,\frac{x}{2}\right) = \frac{1}{2}\sqrt{B}(1-x)\,\sin\left(\sqrt{B}\,\frac{x}{2}\right) \tag{2.5}$$

and

$$\mu_0 = J[u^*] = B\alpha(B).$$

Proof. By Lemma 1, problem (2.1) – (2.2) has an optimal control $u^* \in \Omega$. Set $y_1 = y$, $y' = y_2$, and $u = u^*$. Then (2.1) turns into the system

$$y_1' = y_2, \qquad y_2' = -u^* y_1, \qquad y_2(0) = y_2(1) = 0.$$

According to Pontryagin's maximum principle, the Hamilton function of the system reads as follows:

$$H = -u^* + \lambda_1 y_2 + \lambda_2(-u^* y_1),$$

where $\lambda_1(t)$ and $\lambda_2(t)$ satisfy

$$\begin{gathered} \lambda_1' = n^* \lambda_2, \qquad \lambda_2' = -\lambda_1, \\ \lambda_2'(0) - \lambda_1(0) = 0, \qquad \lambda_2'(1) = \lambda_1(1) = 0. \end{gathered} \tag{2.6}$$

Hence $u^*(t)$ satisfies the following:

$$u^*(t) = \begin{cases} 0, & 1 + \lambda_2(t) y_1(t) > 0, \\ B, & 1 + \lambda_2(t) y_1(t) < 0. \end{cases} \tag{2.7}$$

By (2.6), $\lambda_2(t)$ and $y_1(t)$ are solutions of BVP (2.1) for $u = u^*$. Therefore, they are linearly dependent. Thus we can assume that for some constant c,

$$\lambda_2(t) = c y_1(t) \quad \text{on } [0, 1],$$

which implies

$$1 + \lambda_2(t) y_1(t) = 1 + c y_1^2(t) = 1 + c y^2(t).$$

Since $u^*(t)$ is an optimal control, it follows from Lemma 3 that

$$y_1'(t) \ne 0 \quad \text{on } (0, 1). \tag{2.8}$$

Hence $y_1(t)$ is increasing or decreasing on (0,1). Since $u^*(t) \ge 0$ and $y_1'(0) = y_1'(1) = 0$, $y_1(t)$ has a unique zero in (0,1). By (2.7), we get $u^*(t) = 0$ on a interval near the zero, and consequently,

$$y_1(t) = lt + m \tag{2.9}$$

on the interval satisfying $u^*(t) = 0$, where l and m are suitable constants. From

$$\text{meas}\,\{t \in [0,1] : u^*(t) > 0\} > 0,$$

the monotonicity of $y_1(t)$, and (2.7), (2.9), it follows that there exist $t_1, t_2 \in [0,1]$ such that

$$u^*(t) = \begin{cases} B, & 0 \le t \le t_1, \\ 0, & t_1 \le t \le t_2, \\ B, & t_2 \le t \le 1. \end{cases}$$

Noting (2.9) and

$$y_1'(0) = y_1'(1) = 0,$$

we get

$$0 < t_1 < t_2 < 1.$$

Hence

$$y_1(t) = \begin{cases} k_1 \cos(\sqrt{B}t), & 0 \le t \le t_1, \\ lt + m, & t_1 < t < t_2, \\ k_2 \cos(\sqrt{B}(1-t)), & t_2 \le t \le 1, \end{cases} \tag{2.10}$$

where k_1, k_2, l, and m are suitable constants and $k_1, k_2 \neq 0$. From the continuous differentiability of $y_1(t)$ on [0,1] and (2.10), we have

$$k_1 \cos(\sqrt{B}t_1) = lt_1 + m, \tag{2.11a}$$

$$-k_1\sqrt{B} \sin(\sqrt{B}t_1) = l, \tag{2.11b}$$

$$lt_2 + m = k_2 \cos(\sqrt{B}(1-t_2)), \tag{2.11c}$$

$$l = k_2\sqrt{B} \sin(\sqrt{B}(1-t_2)). \tag{2.11d}$$

By (2.7), we see

$$1 + c[y_1(t_1)]^2 = 1 + c[y_1(t_2)]^2 = 0,$$

which shows

$$y_1(t_1) = -y_1(t_2) = \sqrt{-\frac{1}{c}}\,\text{sgn}\,(y_1(0)).$$

See Fig. 1.
Thus,

$$lt_1 + m = -(lt_2 + m);$$

that is,

$$l(t_1 + t_2) + 2m = 0. \tag{2.12}$$

Note

$$H(y_1(t), y_2(t), \lambda_1(t), \lambda_2(t), u^*(t)) = \begin{cases} -B(1 + ck_1^2), & 0 \le t \le t_1, \\ -cl^2, & t_1 \le t \le t_2, \\ -B(1 + ck_2^2), & t_2 \le t \le 1, \end{cases}$$

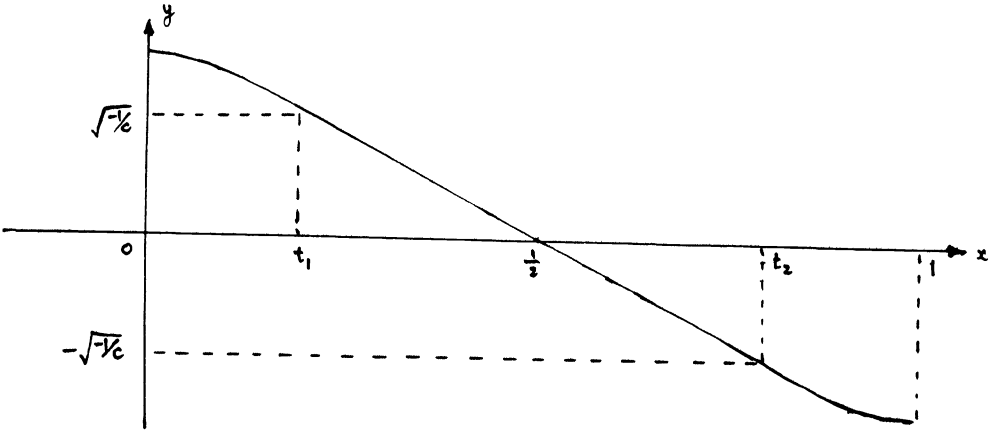


FIG. 1. $\mathbf{y}_1(0) > 0$.

by (2.6), (2.10), and

$$H = -u^* + \lambda_1 y_2 + \lambda_2[-u^* y_1] = \begin{cases} -B - \lambda_2' y_2 - Bcy_1^2, & 0 \le t \le t_1, \\ -\lambda_2' y_2, & t_1 \le t \le t_2, \\ -B - \lambda_2' y_2 - Bcy_1^2, & t_2 \le t \le 1. \end{cases}$$

By Pontryagin's maximum principle,

$$H(y_1, y_2, \lambda_1, \lambda_2, u^*) \equiv \text{const},$$

and hence $k_1^2 = k_2^2$. The monotonicity of $y_1(t)$ and (2.10) yield $k_1 = -k_2$. Since $y_1(t_1) = -y(t_2)$, we have

$$\cos(\sqrt{B} t_1) = \cos(\sqrt{B}(1 - t_2)).$$

By the monotonicity of $y_1(t)$, the arguments $\sqrt{B} t_1$ and $\sqrt{B}(1 - t_2)$ are smaller than $\frac{\pi}{2}$. Thus we get

$$t_1 = 1 - t_2. \tag{2.13}$$

Thus, we need only to study (2.11a), (2.11b), and (2.12) with unknown variables k_1, l, and m. Because $k_1 \neq 0$, the following equality holds certainly:

$$\begin{vmatrix} \cos(\sqrt{B} t_1) & -t_1 & -1 \\ -\sqrt{B} \sin(\sqrt{B} t_1) & -1 & 0 \\ 0 & t_1 + t_2 & 2 \end{vmatrix} = 0.$$

Simplifying it, we get

$$(t_2 - t_1)\sqrt{B} \sin(\sqrt{B} t_1) = 2\cos(\sqrt{B} t_1). \tag{2.14}$$

Set $t_1 = \frac{1}{2}\alpha$. Then $t_2 - t_1 = 1 - \alpha$. Hence, (2.14) turns into

$$\cos\left(\sqrt{B}\frac{\alpha}{2}\right) = \frac{1}{2}(1 - \alpha)\sqrt{B} \sin\left(\sqrt{B}\frac{\alpha}{2}\right). \tag{2.15}$$

Since $u^*(t)$ is an optimal control, α is the minimal positive root of (2.15). Furthermore,

$$\begin{aligned}\int_0^1 u^*(t)\,\mathrm{d}t &= B(t_1 + (1-t_2)) \\ &= B\alpha = \mu_0.\end{aligned}$$

The proof of the theorem is complete.

LEMMA 4. *Let $B > \pi^2$. Then $\alpha(B)$ and $B\alpha(B)$ are strictly decreasing, where $\alpha(B)$ is the minimal positive root of* (2.5).

Proof. By Theorem 1, for $B > \pi^2$, $\alpha(B)$ is well defined and satisfies (2.15). Set

$$F(B,\alpha) = \cos\left(\sqrt{B}\,\frac{\alpha}{2}\right) - \frac{1}{2}(1-\alpha)\sqrt{B}\,\sin\left(\sqrt{B}\,\frac{\alpha}{2}\right).$$

Then if (2.15) holds,

$$\begin{aligned}\frac{\partial F}{\partial \alpha} &= -\frac{1}{4}(1-\alpha)B\cos\left(\sqrt{B}\,\frac{\alpha}{2}\right) \\ &= -\frac{1}{8}(1-\alpha)^2 B\sqrt{B}\,\sin\left(\sqrt{B}\,\frac{\alpha}{2}\right), \\ \frac{\partial F}{\partial B} &= -\frac{1}{4\sqrt{B}}\sin\left(\sqrt{B}\,\frac{\alpha}{2}\right) - \frac{1}{8}\alpha(1-\alpha)\cos\left(\sqrt{B}\,\frac{\alpha}{2}\right) \\ &= -\left[\frac{1}{4\sqrt{B}} + \frac{1}{16}\alpha(1-\alpha)^2\sqrt{B}\right]\sin\left(\sqrt{B}\,\frac{\alpha}{2}\right).\end{aligned}$$

Hence,

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}B}\alpha &= -\frac{4+\alpha(1-\alpha)^2 B}{2(1-\alpha)^2 B^2} < 0, \\ \frac{\mathrm{d}}{\mathrm{d}B}(B\alpha(B)) &= \alpha(B) + B\frac{\mathrm{d}}{\mathrm{d}\alpha}\alpha(B) = -\frac{4-\alpha(1-\alpha)^2 B}{2(1-\alpha)^2 B}.\end{aligned}$$

It suffices to prove $4 - \alpha(1-\alpha)^2 B > 0$. Note that

$$B\alpha \le \pi^2, \qquad \alpha \le 1.$$

Therefore, by (2.15), we get

$$\frac{4}{\alpha(1-\alpha)B} = \frac{\mathrm{tg}\left(\sqrt{B}\,\frac{\alpha}{2}\right)}{\sqrt{B}\,\frac{\alpha}{2}} > 1,$$

since $0 < \frac{\sqrt{B\alpha}}{2} \le \frac{\pi}{2}$. This completes the proof.

LEMMA 5. $B\alpha(B) \searrow 4$ (*as* $B \to \infty$).

Proof. Equation (2.15) yields

$$\mathrm{tg}\left(\sqrt{B}\,\frac{\alpha}{2}\right) = \frac{1}{\frac{1}{2}(1-\alpha)\sqrt{B}}. \tag{2.16}$$

Since

$$B\alpha \le \pi^2 \quad \text{for all } B > \pi^2,$$

we have

$$\alpha(B) \to 0 \qquad (B \to \infty),$$

which implies (by (2.16))

$$\operatorname{tg}\left(\sqrt{B}\,\frac{\alpha}{2}\right) \to 0 \qquad (B \to \infty).$$

Because

$$B\frac{\alpha^2}{4} = (B\alpha)\,\frac{\alpha}{4} \le \frac{\pi^2}{4}\alpha \to 0 \qquad (B \to \infty),$$

multiplying (2.16) by $2\sqrt{B}$ yields

$$B\alpha \cdot \frac{\operatorname{tg}\left(\sqrt{B}\,\frac{\alpha}{2}\right)}{\frac{\alpha}{2}\sqrt{B}} = \frac{4}{1-\alpha}.$$

Letting $B \to \infty$ in this equality, we get the desired conclusion.

As applications, we have the following theorems.

THEOREM 2. *Let $B > \pi^2$, and let $b, f : [0, 1] \to R$ be continuous functions such that*

$$0 \le b(t) \le B, \qquad b(t) \not\equiv 0 \quad \text{on } [0,1], \qquad \int_0^1 b(t)\,\mathrm{d}t < B\alpha(B),$$

where $\alpha(B)$ is the minimal positive root of (2.5). *Then for each $A, B \in R$, the BVP*

$$y'' + b(t)y = f(t), \qquad y'(0) = A, \qquad y'(1) = B \tag{2.17}$$

has a unique solution.

Proof. By Theorem 1, it is obvious that the BVP

$$y'' + b(t)y = 0, \qquad y'(0) = 0, \qquad y'(1) = 0$$

has at most one solution. Since the equation is linear, the uniqueness implies the existence. The proof is complete.

THEOREM 3. *Let $b \in L^\infty[0, 2\pi]$ such that*

$$b(t) \ge 0 \quad \text{on } [0,1],$$

$$\operatorname{meas}\{t \in [0,1] : b(t) > 0\} > 0,$$

and

$$\int_0^1 b(t)\,\mathrm{d}t \le 4.$$

Then for each $f \in L[0,1]$, the BVP (2.17) *has a unique solution.*

Proof. The proof follows from Theorem 1 and Lemmas 4 and 5.

Remark 1. From Theorem 3 we can obtain the following conclusion.

If $b \ge 0$ on $[0, 1]$, $b(t) > 0$ on a set of positive measure, and $b \in L^\infty[0, 1]$, then in order that there exist a nontrivial solution of the Neumann problem

$$y'' + b(y)y = 0 \quad \text{a.e.}, \qquad y'(0) = y'(1) = 0,$$

it is necessary that

$$\int_0^1 b(t)\,\mathrm{d}t > 4. \tag{2.18}$$

For the Dirichlet problem

$$y'' + b(t) = 0 \quad \text{a.e.}, \qquad y(0) = y(1) = 0,$$

there is also the same conclusion, which is the classic Hartman – Wintner criterion [10].

3. Nonlinear equations. In this section, we shall give some applications of Theorems 2 and 3. First we prove Theorem A. Without loss of generality, let $A = B = 0$. We first prove uniqueness. Let $y_1(t)$ and $y_2(t)$ be any two solutions of BVP (1.1) – (1.2). Then $y = y_1(t) - y_2(t)$ is a solution of the BVP

$$\begin{aligned} y'' &+ \int_0^1 f_y(t, y_2 + \theta y)\,\mathrm{d}\theta y = 0, \\ y'(0) &= y'(1) = 0. \end{aligned} \tag{3.1}$$

From condition (ii) it follows that

$$0 \le \int_0^1 f_y(t, y_2 + \theta y)\,\mathrm{d}\theta \le \beta(t) \le B \quad \text{on } \{0, 1],$$

$$\text{meas}\,\{t \in [0, 1] : \int_0^1 f_y(t, y_2 + \theta y)\,\mathrm{d}\theta > 0\} > 0.$$

Hence by Theorem 2, $y(t) \equiv 0$ on [0, 1].

Now we prove existence. Rewrite (1.1) as follows:

$$y'' + b(t, y)y = -f(t, 0),$$

where $b(t, y) = \int_0^1 f_y(t, \theta y)\,\mathrm{d}\theta$. Set

$$X = \{y \in C^1([0, 1], R) : y'(0) = y'(1) = 0\}$$

with the norm $\|\cdot\|$ defined by

$$\|y\| = \max_{[0,1]} |y(t)| + \max_{[0,1]} |y'(t)| \quad \text{for each } y \in X.$$

Define an operator $T : X \to X$ by

$$Tx = y_x,$$

where $y_x(t)$ is a solution of the BVP

$$y'' + b(t, x)y = -f(t, 0), \qquad y'(0) = y'(1) = 0. \tag{3.2}$$

By virtue of Theorem 2, BVP (3.2) has a unique solution, and hence T is well defined on X.

We claim that there is $M > 0$ such that

$$\|Tx\| \le M \quad \text{for all } x \in X.$$

In fact, if not, there would exist a sequence $\{x_n(t)\} \subset X$ such that $\|y_{x_n}\| \to \infty\,(n \to \infty)$. Since

$$0 \le b(t, x_n) \le \beta(t) \quad \text{on } [0,1],$$

passing to a subsequence if necessary, we may assume that $b(t, x_n)$ is weakly convergent to $b_0 \in L^2[0,1]$. Because the set

$$S = \{u \in L^2[0,1] : 0 \le u(t) \le \beta(t) \quad \text{on } [0,1]\}$$

is bounded convex closed in $L^2[0,1]$, by the Hahn–Banach theorem, we see $b_0 \in S$. By (3.2) and the Arzela–Ascoli theorem, passing a subsequence, we may assume that

$$y_n = y_{x_n}/\|y_{x_n}\| \to y_0, y'_{z_n}/\|y_{x_n}\| \to z_0 \quad \text{in } C([0,1], R).$$

Note that for each n,

$$\begin{aligned} y_n(t) &= y_n(0) + \int_0^t y'_n(s)\,\mathrm{d}s, \\ y'_n(t) &= -\int_0^t b(s, x_n) y_n(s)\,\mathrm{d}s - \frac{1}{\|y_{x_n}\|}\int_0^t f(s,0)\,\mathrm{d}s, \end{aligned} \tag{3.3}$$

for all $t \in [0,1]$. Letting $n \to \infty$ in (3.3) by Lebesgue's dominant convergence theorem, we get

$$\begin{aligned} y_0(t) &= y_0(0) + \int_0^t z_0(s)\,\mathrm{d}s, \\ z_0(t) &= -\int_0^t b_0(s) y_0(s)\,\mathrm{d}s \end{aligned}$$

for all $t \in [0,1]$.

Note that $\|y_0\| = 1$. These show that $y_0(t)$ is a nontrivial solution of the BVP

$$y'' + b_0(t)y = 0, \qquad y'(0) = y'(1) = 0.$$

Since for each n

$$\text{meas}\,\{t \in [0,1] : b(t, x_n) > 0\} > 0,$$

$$\int_0^1 b(t, x_n(t)) y_{x_n}(t)\,\mathrm{d}t = -\int_0^1 f(t,0)\,\mathrm{d}t = 0 \quad \text{(by (iii))},$$

$y_{x_n}(t)$ has a zero in [0,1], and hence so does $y_0(t)$. This, combined with $\|y_0\| = 1$, implies that

$$\text{meas}\,\{t \in [0,1] : b_0(t) > 0\} > 0, \tag{3.4}$$

but by (ii) and (3.4), the BVP

$$y'' + b_0(t)y = 0, \qquad y'(0) = y'(1) = 0$$

has only the zero solution, which contradicts $\|y_0\| = 1$. Hence the claim holds.

Given any $\{x_n\} \subset X$ such that $x_n \to x_0 \in X$, by definition we know that

$$y''_{x_n} + b(t, x_n) y_{x_n} = -f(t,0), \qquad y'_{x_n}(0) = y'_{x_n}(1) = 0,$$

and

$$y''_{x_0} + b(t, x_0) y_{x_0} = -f(t,0), \qquad y'_{x_0}(0) = y'_{x_0}(1) = 0.$$

Hence, setting $w_n = y_{x_n} - y_{x_0}$, we have

$$w''_n + b(t, x_n) y_{x_n} - b(t, x_0) y_{x_0} = 0, \qquad w'_n(0) = w'_n(1) = 0. \tag{3.5}$$

Thus

$$w''_n + [b(t, x_n) - b(t, x_0)] y_{x_n} + b(t, x_0) w_n = 0 \quad \text{on } [0,1]. \tag{3.6}$$

We claim that

$$w_n \to 0 \quad \text{in} \, C^1([0,1], R). \tag{3.7}$$

If not, then there would be an $m > 0$ such that

$$\overline{\lim_{n\to\infty}} \|w_n\| \geq m.$$

Applying the Arzela–Ascoli theorem and passing to a subsequence if necessary, by the above claim, we may assume that

$$w_n \to w_0 (n \to \infty) \quad \text{in} \, X. \tag{3.8}$$

Note that

$$\frac{w''_n}{\|w_n\|} + b(t, x_0) \frac{w_n}{\|w_n\|} + [b(t, x_n) - b(t, x_0)] \frac{y_{x_n}}{\|w_n\|} = 0 \quad \text{on } [0,1] \tag{3.9}$$

and that

$$\begin{aligned} \frac{w_n}{\|w_n\|} - \frac{w_0}{\|w_0\|} = \frac{w_n - w_0}{\|w_n\|} + \frac{w_0(\|w_0\| - \|w_n\|)}{\|w_n\|\|w_0\|} \to 0 \\ (n \to \infty \, \text{by} \, (3.8)) \, \text{in} \, X. \end{aligned} \tag{3.10}$$

From (3.9), (3.10) it follows that

$$\begin{aligned} \frac{w_0(t)}{\|w_0\|} &= \frac{w_0(0)}{\|w_0\|} + \int_0^t \frac{w'_0(s)}{\|w_0\|} \, ds, \\ \frac{w'_0(t)}{\|w_0\|} &= -\int_0^t b(s, x_0) \frac{w_0(s)}{\|w_0\|} \, ds \quad \text{on } [0,1]. \end{aligned}$$

This shows that $\frac{w_0(t)}{\|w_0\|}$ is a nontrivial solution of the BVP

$$y'' + b(t, x_0) y = 0, \qquad y'(0) = y'(1) = 0. \tag{3.11}$$

On the other hand, by Theorem 2, the BVP (3.11) has only a zero solution, which leads to a contradiction. Hence, the claim is proved.

From two claims and the Arzela–Ascoli theorem, it follows that $T : X \rightarrow X$ is completely continuous and TX is bounded. According to Schauder's fixed-point theorem, T has a fixed point $y(t)$ in X; that is, BVP (3.1) – (3.2) has a solution $y(t)$. This completes the proof of the theorem.

Remark 2. We can remove (iii) by assuming the following condition:

(iii)′ There exists $\eta \in L^1[0, 1]$ such that

$$0 \leq \eta(t) \leq f_y(t, y) \quad \text{on } [0, 1] \times R$$

and

$$\text{meas}\, \{t \in [0, 1] : \eta(t) > 0\} > 0.$$

This is also available for the following theorem.

Using Theorem 3, we can prove the following theorem.

THEOREM B. *Assume that f satisfies conditions* (i) *and* (iii) *and*

$$0 \leq f_y(t, y) \leq \beta(t) \quad \textit{on } [0, 1] \times R,$$

where $\beta \in L^1[0, 1]$ and satisfies $\int_0^1 \beta(t)\, dt \leq 4$. Then BVP (1.1) – (1.2) *has a unique solution.*

Acknowledgment. The authors thank the referees for their valuable suggestions.

REFERENCES

[1] G. Anichini and G. Conti, *Existence of solutions of a boundary value problem through the solution map of a linearized type problem*, Rend. Sem. Mat. Univ. Politec. Torino, 48 (1990), pp. 149–159.

[2] L. H. Erbe, *Existence of solutions to boundary value problems for ordinary differential equations*, Nonlinear Anal., 6 (1982), pp. 1155–1162.

[3] R. E. Gaines and J. Mawhin, *Ordinary differential equations with nonlinear boundary conditions*, J. Differential Equations, 26 (1977), pp. 200–222.

[4] H. Gingold, *Uniqueness of solutions of boundary value problems of systems of ordinary differential equations*, Pacific J. Math., 75 (1987), pp. 107–136.

[5] ———, *Uniqueness criteria for second order nonlinear boundary value problems*, J. Math. Anal. Appl., 73 (1980), pp. 392–410.

[6] A. Granas, R. B. Guenther, and J. W. Lee, *On a theorem of S. Bernstein*, Pacific J. Math., 74 (1978), pp. 67–82.

[7] ———, *Some general existence principles in the Carathéodory theory of nonlinear differential systems*, J. Math. Pures appl. (9), 70 (1991), pp. 153–196.

[8] D. Hankerson and J. Henderson, *Optimality for boundary value problems for Lipschitz equations*, J. Differential Equations, 77 (1989), pp. 392–404.

[9] G. A. Harris, *On multiple solutions of a nonlinear Neumann problem*, J. Differential Equations, 95 (1992), pp. 75–104.

[10] P. Hartman and A. Wintner, *On an oscillation criterion of Liapounoff*, Amer. J. Math., 73 (1951), pp. 885–890.

[11] J. Henderson, *Best interval lengths for boundary value problems for third order Lipschitz equations*, SIAM J. Math. Anal., 18 (1987), pp. 293–305.

[12] ———, *Boundary value problems for nth order Lipschitz equations*, J. Math. Anal. Appl., 134 (1988), pp. 196–210.

[13] J. Henderson and R. McGwier, JR., *Uniqueness, existence and optimality for fourth order Lipschitz equations*, J. Differential Equations, 67 (1987), pp. 414–440.

[14] L. Jackson, *Existence and uniqueness of solutions of boundary value problems for Lipschitz equations*, J. Differential Equations, 32 (1979), pp. 76–90.

[15] L. Jackson, *Boundary value problems for Lipschitz equations*, in Differential Equations, S. Ahmad, M. Keener, and A. Lazer, eds, Academic Press, New York, 1980, pp. 31–50.

[16] R. Kannan and J. Locker, *On a class of nonliner boundary value problems*, J. Differential Equations, 26 (1977), pp. 1–8.

[17] A. C. Lazer and D. E. Leach, *On a nonlinear two-point boundary value problem*, J. Math. Anal. Appl., 26 (1969), pp. 20–27.

[18] J. Mawhin, *Topological Degree Methods in Nonlinear Boundary Value Problems*, CBMS Regional Conference Series in Mathematics 40, American Mathematical Society, Providence, RI, 1979.

[19] J. Mawhin, J. R. Ward, and M. Willem, *Variational methods and semi-linear elliptic equations*, Arch. Rational Mech. Anal. 95 (1986), pp. 269–277.

[20] Yu. Melentsova, *A best possible estimate of the nonoscillation interval for a linear differential equation with coefficients bounded in Lr*, Differentsial'nye Uravneniya, 13 (1977), pp. 1776–1786 (in Russian); Differential Equations, 13 (1977), pp. 1236–1244 (in English).

[21] Yu. Melentsova and G. Mil'shtein, *An optimal estimate of the interval on which a multipoint boundary value problem possesses a solution*, Differentsial'nye Uraveniya, 10 (1974), pp. 1630–1641 (in Russian); Differential Equations, 10 (1974), pp. 1257–1265 (in English).

[22] ———, *Optimal estimation of the nonoscillation interval for linear differential equations with bounded coefficients*, Differentsial'nye Uraveniya 17 (1981), pp. 2160–2175, 2299. (in Russian); Differential Equations 17 (1981), pp. 1368–1379 (in English).

[23] W. V. Petryshyn, *Solvability of various boundary value problems for the equation* $x'' = f(t, x, x', x'') - y$, Pacific J. Math., 122 (1986), pp. 169–195.

[24] J. Saranen and S. Seikkala, *Solution of a nonlinear two-point boundary value problem with Neumann-type boundary data*, J. Math. Anal. Appl., 135 (1988), pp. 691–701.

[25] J. Troch, *On the interval of disconjugacy of linear autonomous differential equations*, SIAM J. Math. Anal., 12 (1981), pp. 78–89.

[26] G. Vidossich, *A general existence theorem for boundary value problems for ordinary differential equations*, Nonlinear Anal., 15 (1990), pp. 897–914.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 33, No. 5, pp. 1326 1356, September 1995

© 1995 Society for Industrial and Applied Mathematics
003

EQUIVALENT CONDITIONS FOR THE SOLVABILITY OF THE NONSTANDARD LQ-PROBLEM FOR PRITCHARD–SALAMON SYSTEMS*

BERT VAN KEULEN†

Abstract. Equivalent conditions are presented for the solvability of the infinite-horizon LQ-problem with stability for a class of infinite-dimensional systems with unbounded input and output operators. In particular, these equivalent conditions are given in terms of a Riccati equation and a frequency domain inequality just as in the finite-dimensional case.

Key words. infinite-dimensional systems, linear quadratic control, unbounded input and output operators, Pritchard–Salamon systems

AMS subject classifications. 93C25, 49J27

1. Introduction.

The linear quadratic control problem (LQ-problem) has received a lot of attention in the area of mathematical systems theory. Many interesting system theoretical problems can be formulated in this framework, for instance, questions arising in optimal control, identification theory, and robust stability theory (Kalman–Yakubovich–Popov Lemma, Bounded Real Lemma and so on; see, e.g., the expository paper by Willems [22]). Moreover, the recent results in the area of H_∞-control owe much to the LQ-theory.

The first solutions for the LQ-problem were obtained for the case that the cost criterion is positive definite (sometimes called the *standard* case):

$$J(x_0, u(\cdot)) = \int_0^\infty (\|Cx(t)\|_Y^2 + \|R^{1/2}u(t)\|_U^2)\, dt.$$

However, for many applications a more general formulation is needed. In [22], the LQ-problem is treated for finite-dimensional systems in its most general form; that is, the cost criterion is not necessarily positive definite. In particular, we refer to [22, Thm. 5], which shows the equivalence of the solvability of the LQ-problem with stability to the existence of a solution to a *nonstandard* Riccati equation and a frequency domain inequality.

The system theoretical questions mentioned above are, of course, also interesting for infinite-dimensional systems, and the corresponding LQ-problems can be formulated in a similar way. In fact, much attention has been devoted to the standard LQ-problem for infinite-dimensional systems. In [4], [1] (and references therein) the case is treated where the input and output operators are bounded, and in [10], [15], [11] (and references therein) the unbounded case is considered. The nonstandard problem, however, has received less attention. In [25] the bounded input/output case is treated assuming that the generator of the semigroup of the system is bounded, and in [21] and [12], [13] this result is extended to unbounded generators. These authors have obtained a complete generalization of the aforementioned result in [22] for their classes of systems. It is the purpose of this paper to do the same for a class of unbounded input/output systems (the Pritchard–Salamon class introduced in [15]).

* Received by the editors April 13, 1992; accepted for publication (in revised form) February 22, 1994.

† Mathematics Institute, P.O. Box 800, 9700 AV Groningen, The Netherlands (bertvk@rug.nl).

We shall consider nonstandard cost functions of the form

$$J(x_0,u(\cdot)) = \int_0^\infty (\|C_1x(t)\|_{Y_1}^2 - \|C_2x(t)\|_{Y_2}^2 + \langle Lx(t),u(t)\rangle_U + \|R^{1/2}u(t)\|_U^2)\,dt,$$

where C_1, C_2, and L are unbounded output operators that are "admissible" in the sense of Pritchard and Salamon in [15]. In the paper [15] Pritchard and Salamon present a general framework to model a large class of systems with unbounded control and observation operators (it extends the class of bounded input/output semigroup control systems). Their class of systems includes many delay systems: neutral systems with output delays [16] (see also §4) and retarded systems with delays in input and output [15], [19]. Furthermore, the Pritchard–Salamon class includes a class of parabolic PDE systems with unbounded control and observation [15] (see also §4). In [15] it is shown that certain hyperbolic PDE systems also fall into the Pritchard–Salamon framework, but generally these are only exponentially stabilizable if there is some internal damping (in this paper exponential stabilizability will be a standing assumption). In [2] it is shown that the Euler–Bernoulli beam with Kelvin–Voigt damping (see also §4) fits into the Pritchard–Salamon class, but usually one does not call this a hyperbolic system.

It should be noted that the *standard LQ*-problem can be treated for many PDE systems that do not fit into the Pritchard–Salamon class (see [10], [11] and references therein). However, usually one has to distinguish several cases (essentially the parabolic and the hyperbolic cases). In the Pritchard–Salamon framework it is possible to treat delay systems and PDE systems at the same time, at the expense of the allowed amount of input/output unboundedness for PDE systems. It can be argued that the strength of the Pritchard–Salamon class lies in the fact that it contains so many delay systems; a weak point is that not so many PDE systems fit in. In any case, this paper contains the *first generalization of the nonstandard LQ-problem to infinite-dimensional systems with unbounded inputs and outputs*; all the above-mentioned examples can now be accommodated.

In §2 we give some preliminary results, some of which are known (stated with references) and some of which are new. In particular, we refer to [3] where many system theoretic results were obtained for the Pritchard–Salamon class, including some that made this paper possible. After the introduction of the Pritchard–Salamon class, some perturbation theory is given and we show how preliminary (unbounded) feedbacks can be handled. Then we derive an interesting formula that shows how the "unbounded part" of a Pritchard–Salamon system (i.e., unbounded with respect to the larger state space) can be expressed as "something bounded." Section 2 is concluded with some frequency domain results for Pritchard–Salamon systems and a quotation of the main result in [13], which is the generalization of [22, Theorem 5] to the class of infinite-dimensional systems with bounded input and output operators. Our main result, which is a generalization of this result to the Pritchard–Salamon class, is given in §3. In the proof of the main result we apply some preliminary (unbounded) feedback in order to take care of the "cross term" in the LQ-cost criterion. Furthermore, we use the formula for the "unbounded part" given in §2 so that the LQ-result for the bounded input/output case in [13] can be used. Finally, in §4 we shall treat several examples of the theory presented in this paper and give some conclusions.

2. Preliminary results.

We consider the same class of systems as in [15], [16], [19], and [3]. Let W and V be real Hilbert spaces satisfying

$$W \hookrightarrow V, \tag{2.1}$$

where by $\hookrightarrow$ we mean that $W \subset V$ and the canonical injection $W \to V$, $x \mapsto x$ is continuous and W is dense in V. We consider C_0-semigroups $S(\cdot)$ on V that restrict to C_0-semigroups on W (we note that the growth bounds of $S(\cdot)$ on W and V need not be the same, as shown in [3]). The infinitesimal generators of $S(\cdot)$ on V and W will be denoted by A^V and A^W, respectively. We note that A^W is the part of A^V in W so that $D(A^W) = \{x \in D(A^V) \cap W \mid A^V x \in W\}$ and $A^W x = A^V x$ for all $x \in D(A^W)$ (see [14, §4.5]). Using $W \hookrightarrow V$, it is not difficult to show that in fact

$$D(A^W) \hookrightarrow D(A^V).$$

Hilbert adjoints of linear operators are denoted by $*$.

The following definition can be found in [15] and [3]. It was first stated in [15], but we use the formulation of [3].

DEFINITION 2.1. *Let U and Y be Hilbert spaces.*

1. *An operator $B \in \mathcal{L}(U, V)$ is called an admissible input operator if there exist $t_1 > 0$ and $\alpha > 0$ such that*

$$\int_0^{t_1} S(t_1 - s)Bu(s)\,ds \in W \tag{2.2}$$

and

$$\left\| \int_0^{t_1} S(t_1 - s)Bu(s)\,ds \right\|_W \leq \alpha \|u(\cdot)\|_{L_2(0,t_1;U)} \tag{2.3}$$

for all $u(\cdot) \in L_2(0, t_1; U)$.

2. *An operator $C \in \mathcal{L}(W, Y)$ is called an admissible output operator for $S(\cdot)$ if there exist $t_1 > 0$ and $\alpha > 0$ such that*

$$\|CS(\cdot)x\|_{L_2(0,t_1;Y)} \leq \alpha \|x\|_V \text{ for all } x \in W. \tag{2.4}$$

3. *Let $B \in \mathcal{L}(U, V)$ and $C \in \mathcal{L}(W, Y)$ be admissible input and output operators, respectively. The system $\Sigma(S(\cdot), B, C, D)$ given by*

$$\begin{aligned} x(t) &= S(t)x_0 + \int_0^t S(t - s)Bu(s)\,ds, \\ y(t) &= Cx(t) + Du(t), \end{aligned} \tag{2.5}$$

where $x_0 \in V, t \geq 0$, and $u(\cdot) \in L_2^{\text{loc}}(0, \infty; U)$ is called a Pritchard–Salamon system.

Remark 2.2. As is mentioned in [3], if the statements (2.3) and (2.4) hold for some t_1 and α, then they hold for any $t_1 > 0$ and some $\alpha > 0$, α depending on t_1.

If $x_0 \in W$, then $x(t)$ defined by (2.5) is continuous with respect to the topology on W and the controllability map $\mathcal{C}$ from $L_2(0, T; U)$ to W given by

$$\mathcal{C}u(\cdot) = \int_0^T S(T - \tau)Bu(\tau)\,d\tau$$

satisfies $\mathcal{C} \in \mathcal{L}(L_2(0, T; U), W)$.

Furthermore, (2.4) implies that the linear map from W to $L_2(0, T; Y)$, $x \mapsto CS(\cdot)x$ has a unique bounded extension from V to $L_2(0, T; Y)$, which will be denoted by $x \mapsto \overline{CS}(\cdot)x$ for $x \in V$. Hence, the ouput $y(\cdot) \in L_2(0,T;Y)$ in (2.5) should be interpreted as

$$y(\cdot) = \overline{CS}(\cdot)x_0 + C\int_0^{\cdot} S(\cdot - s)Bu(s)\,ds.$$

Finally, we note that if $W = V$, then $B \in \mathcal{L}(U,V)$ and $C \in \mathcal{L}(W,Y)$ are automatically admissible (this is the so-called "bounded input/output case").

In the following lemma we state some perturbation results that follow from [3]. In the last part of the lemma we shall use the additional assumption that

$$D(A^V) \hookrightarrow W. \tag{2.6}$$

Here $D(A^V)$ is the Hilbert space with the inner product given by

$$\langle x,y\rangle_{D(A^V)} := \langle x,y\rangle_V + \langle A^V x, A^V y\rangle_V. \tag{2.7}$$

We note that this assumption is not very restrictive, as explained in [15].

LEMMA 2.3. *Let $\Sigma(S(\cdot),B,C,D)$ be a Pritchard–Salamon system of the form* (2.5), *and let $F \in \mathcal{L}(W,U)$ be an admissible output operator for this system. Then there exists a unique C_0-semigroup $S_{BF}(\cdot)$ on V that restricts to a C_0-semigroup on W such that $\Sigma(S_{BF}(\cdot),B,C,\ D)$ is a Pritchard–Salamon system and*

$$S_{BF}(t)x = S(t)x + \int_0^t S(t-s)B\overline{FS}_{BF}(s)x\,ds \text{ for all } x \in V. \tag{2.8}$$

Furthermore,

$$S_{BF}(t)x = S(t)x + \int_0^t S_{BF}(t-s)B\overline{FS}(s)x\,ds \text{ for all } x \in V \tag{2.9}$$

and

$$D(A^V_{BF}) = D(A^V) \quad \text{with equivalent graph norms,}$$

$$D(A^W_{BF}) = \{x \in D(A^V) \cap W) \mid A^V_{BF}x \in W\},$$

$$A^V_{BF}x = A^V x + \overline{BF(\sigma I - A^W)^{-1}}(\sigma I - A^V)x \text{ for all } x \in D(A^V),$$

where σ is any number with real part larger than the growth bounds of $S(\cdot)$ and $S_{BF}(\cdot)$ on W and V. If, in addition, assumption (2.6) *is satisfied, there holds*

$$A^V_{BF}x = (A^V + BF)x \text{ for all } x \in D(A^V).$$

Proof. Almost all of the results in this lemma can be found in [3]. Here we prove that the graph norms of $D(A^V)$ and $D(A^V_{BF})$ are equivalent and we derive the expressions for A^V_{BF}.

As explained in [3], we can apply the Laplace transform to (2.8) and (2.9) to obtain for all $x \in V$

$$(\sigma I - A^V_{BF})^{-1}x = (\sigma I - A^V)^{-1}x + (\sigma I - A^V)^{-1}\overline{BF(\sigma I - A^W_{BF})^{-1}}x$$

and

$$(\sigma I - A^V)^{-1}x = (\sigma I - A^V_{BF})^{-1}x - (\sigma I - A^V_{BF})^{-1}B\overline{F(\sigma I - A^W)^{-1}}x,$$

where σ is any number with real part larger than the growth bounds of $S(\cdot)$ and $S_{BF}(\cdot)$ on W and V. Defining $T_1 := I + B\overline{F(\sigma I - A^W_{BF})^{-1}} \in \mathcal{L}(V)$ and $T_2 := I - B\overline{F(\sigma I - A^W)^{-1}} \in \mathcal{L}(V)$, we can reformulate the above equations as $(\sigma I - A^V)x = T_1(\sigma I - A^V_{BF})x$ and $(\sigma I - A^V_{BF})x = T_2(\sigma I - A^V)x$, for all $x \in D(A^V) = D(A^V_{BF})$. Since $T_1, T_2 \in \mathcal{L}(V)$, it follows that the graph norms of $D(A^V)$ and $D(A^V_{BF})$ are equivalent. Manipulation of the last equation shows that $A^V_{BF}x = A^Vx + B\overline{F(\sigma I - A^W)^{-1}}(\sigma I - A^V)x$ for all $x \in D(A^V)$. Finally, we prove that if (2.6) is satisfied, it follows that for all $x \in D(A^V)$

$$\overline{F(\sigma I - A^W)^{-1}}(\sigma I - A^V)x = Fx \tag{2.10}$$

(note that this expression is valid for $x \in D(A^W)$). Since $W \hookrightarrow V$, we have $D(A^W) \hookrightarrow D(A^V)$; and because (2.10) is satisfied for all $x \in D(A^W)$, the result follows by taking a sequence $x_n \in D(A^W)$ converging to x in the topology of $D(A^V)$. □

Remark 2.4. We note that if $F \in \mathcal{L}(V, U)$, then F is an admissible output operator. We call the pair (A^V, B) *exponentially stabilizable on* V if there exists some $F \in \mathcal{L}(V, U)$ such that $S_{BF}(\cdot)$ is exponentially stable on V (this corresponds with the usual definition of exponential stabilizability on the Hilbert space V). Using the perturbation results of Lemma 2.3 and [6, Cor. 1], it is not difficult to show (see [3]) that the following two conditions are equivalent:

1. there exists an admissible output operator $F_1 \in \mathcal{L}(W, U)$ such that $S_{BF_1}(\cdot)$ is exponentially stable on V;
2. (A^V, B) is exponentially stabilizable on V.

Finally, we mention that if assumption (2.6) holds, condition (2.3) is implied by (2.2) (see [23]).

Using the perturbation results in Lemma 2.3, we can now make sense of a preliminary feedback "$u = Fx + v$" in system (2.5), where F is an admissible output operator. We note that if F is bounded (i.e., $F \in \mathcal{L}(V, U)$), this type of result can be found in [14].

LEMMA 2.5. *Let $\Sigma(S(\cdot), B, C, D)$ be a Pritchard-Salamon system of the form (2.5), and let $F \in \mathcal{L}(W, U)$ be an admissible output operator for this system.*

1. *Suppose that $v(\cdot) \in L_2^{\text{loc}}(0, \infty; U)$, and define*

$$x_F(t) := S_{BF}(t)x_0 + \int_0^t S_{BF}(t - s)Bv(s)\, ds \tag{2.11}$$

and

$$u(t) := Fx_F(t) + v(t). \tag{2.12}$$

Then $u(\cdot) \in L_2^{\text{loc}}(0, \infty; U)$ and $x(t)$ given by

$$x(t) = S(t)x_0 + \int_0^t S(t - s)Bu(s)\, ds \tag{2.13}$$

satisfies

$$x(t) = x_F(t) \textit{ for all } t \geq 0 \textit{ and } Cx(t) = Cx_F(t) \textit{ for a.e. } t \geq 0. \tag{2.14}$$

2. *Suppose that* $u(\cdot) \in L_2^{\text{loc}}(0,\infty;U)$, *and define* $x(\cdot)$ *by* (2.13) *and* $v(\cdot)$ *by*

$$v(t) := -Fx(t) + u(t). \tag{2.15}$$

Then $v(\cdot) \in L_2^{\text{loc}}(0,\infty;U)$ *and* $x_F(t)$ *given by* (2.11) *satisfies*

$$x(t) = x_F(t) \text{ for all } t \geq 0 \text{ and } Cx(t) = Cx_F(t) \text{ for a.e. } t \geq 0. \tag{2.16}$$

3. *Let* $v(\cdot) \in L_2^{\text{loc}}(0,\infty;U)$, *and define* $x_F(\cdot)$ *as in* (2.11) *and* $u(\cdot)$ *as in* (2.12) *i.e.,* $u(t) = Fx_F(t) + v(t)$. *Then* $v(\cdot)$ *satisfies* (2.15), *i.e.,* $v(t) = -Fx(t) + u(t)$. *Conversely, let* $u(\cdot) \in L_2^{\text{loc}}(0,\infty;U)$ *and define* $x(\cdot)$ *as in* (2.13) *and* $v(\cdot)$ *as in* (2.15). *Then* $u(\cdot)$ *satisfies* (2.12).

Proof. Proof of 1. The expression $Fx_F(\cdot)$ in $u(\cdot) := Fx_F(\cdot) + v(\cdot)$ should be interpreted as explained in Remark 2.2:

$$u(\cdot) = \overline{FS}_{BF}(\cdot)x_0 + F\int_0^{\cdot} S_{BF}(\cdot - s)Bv(s)\,ds + v(\cdot) \in L_2^{\text{loc}}(0,\infty;U),$$

where we note that F is an admissible output operator for $S_{BF}(\cdot)$ (see Lemma 2.3). Substituting $u(\cdot)$ in (2.13) gives

$$\begin{aligned} x(t) = S(t)x_0 &+ \int_0^t S(t-s)B\overline{FS}_{BF}(s)x\,ds + \int_0^t S(t-s)Bv(s)\,ds \\ &+ \int_0^t S(t-s)BF\left[\int_0^s S_{BF}(s-\tau)Bv(\tau)\,d\tau\right]ds. \end{aligned} \tag{2.17}$$

It follows from Lemma 2.3 that the two terms of $x_F(t)$ in (2.11) can be reformulated as

$$S_{BF}(t)x_0 = S(t)x_0 + \int_0^t S_{BF}(t-s)B\overline{FS}(s)x\,ds$$

and

$$\int_0^t S_{BF}(t-s)Bv(s)\,ds = \int_0^t S(t-s)Bv(s)\,ds$$

$$+\int_0^t\left[\int_0^{t-s} S(t-s-\sigma)B\overline{FS}_{BF}(\sigma)Bv(s)\,d\sigma\right]ds.$$

Comparing this with (2.17), we see that to prove that $x(t) = x_F(t)$, we only have to show that

$$\begin{aligned} &\int_0^t S(t-s)BF\left[\int_0^s S_{BF}(s-\tau)Bv(\tau)\,d\tau\right]ds \\ &\quad = \int_0^t\left[\int_0^{t-s} S(t-s-\sigma)B\overline{FS}_{BF}(\sigma)Bv(s)\,d\sigma\right]ds. \end{aligned} \tag{2.18}$$

This means that we have to somehow "get F inside the integral" (the other operators cause no problems, since they are bounded). It follows from [3] that if $v(\cdot)$ is a *step function*, we have

$$F\int_0^s S_{BF}(s-\tau)Bv(\tau)\,d\tau = \int_0^s \overline{FS}_{BF}(s-\tau)Bv(\tau)\,d\tau. \tag{2.19}$$

We note that in [15] a similar result was proved under the extra assumption that (2.6) is satisfied. Using (2.19) and Fubini's Theorem to interchange the integrals, it is straightforward to show that (2.18) is satisfied if $v(\cdot)$ is a step function. The general case can be obtained by introducing sequences $v_n(\cdot)$ on arbitrary intervals $[0, T]$ that converge to $v(\cdot)$ on these intervals in the $L_2(0, T; U)$-norm. Of course the admissibility of B and C should be used. We omit the details and refer to [15, Lem. 2.5] for a similar result that was proved using somewhat different arguments.

Finally, we sketch the proof of the second part of (2.14). This seems to follow trivially from the first part of (2.14). However, $Cx(t) = Cx_F(t)$ should be interpreted in the sense of Remark 2.2, i.e., in the L_2-sense. It is easy to see that $Cx(t) = Cx_F(t)$ holds if $x_0 \in W$. The general result can be obtained by introducing a sequence $x_{0n} \in W$ converging to x_0 as $n \to \infty$ and using the admissibility of B and C. We omit the details.

Proof of 2. Again, using the interpretation of Remark 2.2, we have

$$v(t) = -\overline{FS}(t)x_0 - F\int_0^t S(t-s)Bu(s)\,ds + u(t).$$

The rest of the proof of 2 is similar to the proof of 1 and, therefore, is deleted.

Proof of 3. This follows immediately from 1 and 2. □

In the following lemma we give a Lyapunov type result for systems of the form (2.5) where $S(\cdot)$ is exponentially stable on V. This result (which is also more or less given in [5]) could be considered as a special case of the results in [15], but for reasons of completeness we give a short proof.

LEMMA 2.6. *Let $S(\cdot)$ be a C_0-semigroup on W and V, and suppose that $S(\cdot)$ is exponentially stable on V. Suppose that assumption* (2.6) *holds. Let $C \in \mathcal{L}(W, Y)$ be an admissible output operator for $S(\cdot)$. Then the operator $\mathcal{C}^\infty$ defined on W by*

$$\mathcal{C}^\infty x := CS(\cdot)x \text{ for } x \in W \tag{2.20}$$

has a unique bounded extension on V (denoted by the same symbol) such that

$$\mathcal{C}^\infty \in \mathcal{L}(V, L_2(0, \infty; Y)). \tag{2.21}$$

Furthermore, the nonnegative definite operator $X \in \mathcal{L}(V)$ defined by

$$X := (\mathcal{C}^\infty)^*\mathcal{C}^\infty \tag{2.22}$$

satisfies

$$\langle A^V x, Xy\rangle_V + \langle x, XA^V y\rangle_V + \langle Cx, Cy\rangle_Y = 0 \tag{2.23}$$

for all $x, y \in D(A^V)$.

Proof. It follows from [15, Rem. 3.5(i)], that there exists some $c > 0$ such that for all $x \in W$,

$$\|CS(\cdot)x\|_{L_2(0,\infty;Y)} \le c\|x\|_V. \tag{2.24}$$

Therefore, $\mathcal{C}^\infty \in \mathcal{L}(V, L_2(0, \infty; Y))$ (cf. also Remark 2.2).

Now let $x, y \in D(A^W)$ and define the function

$$f(t) := \langle CS(t)x, CS(t)y\rangle_Y.$$

Then $f(\cdot)$ is defined for all t and it is differentiable; there holds

$$\dot{f}(t) = \langle CS(t)A^W x, CS(t)y\rangle_Y + \langle CS(t)x, CS(t)A^W y\rangle_Y. \tag{2.25}$$

For all $x \in D(A^W)$ we have

$$\|CS(t)x\|_Y \leq \|C\|\|S(t)x\|_W \leq \text{const}\|S(t)x\|_{D(A^V)},$$

where we have used assumption (2.6). Furthermore, for all $x \in D(A^W)$

$$\|S(t)x\|_{D(A^V)} \to 0 \text{ as } t \to \infty,$$

because $S(\cdot)$ is exponentially stable on V and therefore also on $D(A^V)$ (see, e.g., [19]). Therefore, we conclude that $f(t) \to 0$ as $t \to \infty$. Now integrating (2.25) from 0 to T gives

$$f(T) - f(0) = \int_0^T (\langle CS(t)A^W x, CS(t)y\rangle_Y + \langle CS(t)x, CS(t)A^W y\rangle_Y)\, dt.$$

Since $CS(\cdot)x \in L_2(0,\infty;Y)$ for all $x \in W$ (see (2.24)), we can take the limit as $T \to \infty$ to obtain

$$\int_0^\infty (\langle CS(t)A^W x, CS(t)y\rangle_Y + \langle CS(t)x, CS(t)A^W y\rangle_Y)\, dt = -\langle Cx, Cy\rangle_Y.$$

Since $x, y \in D(A^W)$ were arbitrary, we can use (2.20) and the fact that $A^W x = A^V x$ for all $x \in D(A^W)$ to conclude that for all $x, y \in D(A^W)$

$$\langle \mathcal{C}^\infty A^V x, \mathcal{C}^\infty y\rangle_{L_2(0,\infty;Y)} + \langle \mathcal{C}^\infty x, \mathcal{C}^\infty A^V y\rangle_{L_2(0,\infty;Y)} = -\langle Cx, Cy\rangle_Y. \tag{2.26}$$

Using the fact that $D(A^W) \hookrightarrow D(A^V)$ and the assumption that $D(A^V) \hookrightarrow W$, we can extend (2.26) to all $x, y \in D(A^V)$ by choosing sequences $x_n, y_n \in D(A^W)$ that converge to $x, y \in D(A^V)$. Finally, it is clear then that X defined by (2.22) satisfies (2.23). □

We are interested in LQ-problems with stability. We shall try to find controllers that stabilize the state $x(\cdot)$ in (2.5) in the sense that $x(\cdot) \in L_2(0,\infty;V)$. In the next lemma we shall show that, under certain conditions, in this case $y(\cdot) \in L_2(0,\infty;Y)$. This is not obvious in the general case because the growth constant of $S(\cdot)$ on W need not equal the growth constant on V (see [3]) and so we cannot conclude that $x(\cdot) \in L_2(0,\infty;W)$ in general. One required condition is that the pair (A^V, B) must be exponentially stabilizable on V. In this case, it is easy to see that the set U_{adm} defined by

$$U_{\text{adm}} := \{u(\cdot) \in L_2(0,\infty;U) \text{ such that } x(\cdot) \text{ given by}$$
$$(2.5) \text{ satisfies } x(\cdot) \in L_2(0,\infty;V)\}$$

is nonempty.

LEMMA 2.7. *Let $\Sigma(S(\cdot),B,C,D)$ be a Pritchard–Salamon system of the form* (2.5), *and suppose that (A^V,B) is exponentially stabilizable on V. Furthermore, suppose that assumption* (2.6) *is satisfied. For all $u(\cdot) \in U_{\text{adm}}$ it follows that $y(\cdot)$ given by* (2.5) *is in $L_2(0,\infty;Y)$ and if $x_0 = 0$ there holds*

$$\|y(\cdot)\|_{L_2(0,\infty;Y)} \leq \text{const}(\|u(\cdot)\|_{L_2(0,\infty;U)} + \|x(\cdot)\|_{L_2(0,\infty;V)}). \tag{2.27}$$

Proof. If C were an element of $\mathcal{L}(V,Y)$, the fact that $x(\cdot) \in L_2(0,\infty;V)$ would immediately imply the lemma. However, we only know that $C \in \mathcal{L}(W,Y)$ is admissible and so we have to do some work to establish the result.

Since (A^V, B) is exponentially stabilizable on V, there exists an $F \in \mathcal{L}(V,U)$ such that $S_{BF}(\cdot)$ is stable on V. Furthermore, it follows from assumption (2.6) and Lemma 2.3 that $D(A^V_{BF}) \hookrightarrow W$. Using Lemma 2.6 it follows that there exists a nonnegative definite $X_F \in \mathcal{L}(V)$ such that for all $x, y \in D(A^V) = D(A^V_{BF})$

$$\langle A^V_{BF}x, X_F y\rangle_V + \langle x, X_F A^V_{BF} y\rangle_V + \langle Cx, Cy\rangle_Y = 0.$$

Because of (2.6), Lemma 2.3 implies that $A^V_{BF} = A^V + BF$ on its domain and so the above formula can be expressed as

$$\langle A^V x, X_F y\rangle_V + \langle x, X_F A^V y\rangle_V = -\langle Cx, Cy\rangle_Y$$

$$- \langle (X_F BF + F^* B^* X_F)x, y\rangle_V \text{ for all } x, y \in D(A^V) = D(A^V_{BF}). \tag{2.28}$$

Now consider the system (2.5) and suppose that $u(\cdot) \in U_{\text{adm}}$. Let $u_n(\cdot) \in C^1(0,\infty;U)$ be such that $u_n(\cdot) \to u(\cdot)$ (in the L_2-norm) and $x_{0n} \in D(A^V)$ be such that $x_{0n} \to x_0$ (in the norm on V). Let $x_n(\cdot)$ be given by

$$x_n(t) = S(t)x_{0n} + \int_0^t S(t-s)Bu_n(s)\,ds.$$

It follows from known results (see, e.g., [8, App. A.3, A.6]) that $x_n(\cdot)$ is continuously differentiable, $x_n(t) \in D(A^V)$ for all $t \geq 0$,

$$\dot{x}_n(t) = A^V x_n(t) + Bu_n(t) \text{ for all } t \geq 0, \tag{2.29}$$

and for all $T > 0$

$$\|x_n(\cdot) - x(\cdot)\|_{L_2(0,T;V)} \to 0 \text{ and } \|x_n(T) - x(T)\|_V \to 0 \text{ as } n \to \infty. \tag{2.30}$$

Furthermore,

$$\|Cx_n(\cdot) - Cx(\cdot)\|_{L_2(0,T;Y)} \leq \|\overline{CS}(\cdot)(x_{0n} - x_0)\|_{L_2(0,T;Y)}$$

$$+ \left\| C \int_0^{\cdot} S(\cdot - s)B(u_n(s) - u(s))\,ds \right\|_{L_2(0,T;Y)}.$$

Hence, the fact that B and C are both admissible implies that

$$\|Cx_n(\cdot) - Cx(\cdot)\|_{L_2(0,T;Y)} \to 0 \text{ as } n \to \infty. \tag{2.31}$$

Now using (2.28) and (2.29), it is straightforward to show that

$$\frac{d}{dt}\langle x_n(t), X_F x_n(t)\rangle_V = -\langle Cx_n(t), Cx_n(t)\rangle_Y$$

$$- \langle (X_F BF + F^* B^* X_F)x_n(t), x_n(t)\rangle_V + 2\langle Bu_n(t), X_F x_n(t)\rangle_V. \tag{2.32}$$

Integrating (2.32) from 0 to T gives

$$\langle x_n(T), X_F x_n(T)\rangle_V - \langle x_{0n}, X_F x_{0n}\rangle_V + \int_0^T \|Cx_n(t)\|_Y^2\, dt$$

$$= \int_0^T (\langle -(X_F BF + F^* B^* X_F)x_n(t), x_n(t)\rangle_V + 2\langle Bu_n(t), X_F x_n(t)\rangle_V)\, dt.$$

Now we wish to take the limit as $n \to \infty$. Since $X_F \in \mathcal{L}(V), B \in \mathcal{L}(U,V)$, and $F \in \mathcal{L}(V,U)$, the only difficult part is the term $\langle Cx_n(t), Cx_n(t)\rangle_Y$, but this was already dealt with in (2.31). So using (2.30)–(2.31) we obtain

$$\langle x(T), X_F x(T)\rangle_V - \langle x_0, X_F x_0\rangle_V + \int_0^T \|Cx(t)\|_Y^2\, dt$$

$$= \int_0^T (\langle -(X_F BF + F^* B^* X_F)x(t), x(t)\rangle_V + 2\langle Bu(t), X_F x(t)\rangle_V)\, dt.$$

Since X_F is nonnegative definite, $u(\cdot) \in L_2(0,\infty;U)$, and $x(\cdot) \in L_2(0,\infty;V)$, the last equation implies that $Cx(\cdot) \in L_2(0,\infty;Y)$ and therefore $y(\cdot) \in L_2(0,\infty;Y)$. Furthermore, we have that $\|x(T)\|_V \to 0$ as $T \to \infty$ (see, e.g., [8]) and so, using the fact that $u(\cdot) \in L_2(0,\infty;U)$ and $x(\cdot) \in L_2(0,\infty;V)$, we may let $T \to \infty$ to obtain

$$\int_0^\infty \|Cx(t)\|_Y^2\, dt = \langle x_0, X_F x_0\rangle_V$$

$$+ \int_0^\infty (\langle -(X_F BF + F^* B^* X_F)x(t), x(t)\rangle_V + 2\langle Bu(t), X_F x(t)\rangle_V)\, dt. \tag{2.33}$$

Finally, it is clear that if $x_0 = 0$, then (2.27) follows from (2.33). □

Remark 2.8. The fact that $y(\cdot) \in L_2(0,\infty;Y)$ in Lemma 2.7 also follows from very general results in [24]. However, the direct proof that is given here provides us with formula (2.33), which will play an important role in our treatment of the *LQ*-problem. In fact, it shows how we can transform the "unbounded part" with $Cx(\cdot)$ into "something bounded."

In this paper we shall also be interested in frequency domain results. In the following, the complexification of a real Hilbert space H is denoted by the same symbol.

Suppose that H is some Hilbert space, and let $x(\cdot) \in L_2(0,\infty;H)$. The Fourier transform of $x(\cdot)$ denoted by $\hat{x}(\cdot)$ is defined by

$$\hat{x}(i\omega) = \underset{T\to\infty}{\text{l.i.m.}} \int_0^T \exp(-i\omega t)x(t)\, dt,$$

where l.i.m. stands for limit in (quadratic) mean. It is well known (Plancherel's Theorem; see, e.g., [18, §4.8]) that now the function determined by $\omega \mapsto \hat{x}(i\omega)$ is an element of $L_2(\mathbb{R}, H)$ and

$$\|x(\cdot)\|_{L_2(0,\infty;H)}^2 = (1/2\pi)\|\hat{x}(\cdot)\|_{L_2(\mathbb{R};H)}^2.$$

Furthermore, we define the Laplace transform $\hat{x}(\cdot)$ of $x(\cdot)$ by

$$\hat{x}(s) = \int_0^\infty \exp(-st)x(t)dt,\ \mathrm{Re}(s) > 0,$$

and we recall that $\hat{x}(s)$ converges to $\hat{x}(i\omega)$ a.e. on the imaginary axis (see again [18]).

In the next two lemmas we calculate a frequency domain representation of a Pritchard–Salamon system. If $S(\cdot)$ is exponentially stable on V, we find an expression for its transfer function on $\mathbb{C}^+$ and prove that it is in $H^\infty(\mathbb{C}^+)$. A particular difficulty here is that $S(\cdot)$ need not be stable on W and that $\widehat{Cx}(i\omega) = C\hat{x}(i\omega)$ is not immediate because in general $C \notin \mathcal{L}(V, Y)$. Some of the arguments in the proofs that are given here can also be found in [24].

LEMMA 2.9. *Let $\Sigma(S(\cdot), B, C, D)$ be a Pritchard–Salamon system of the form* (2.5)*, and suppose that $S(\cdot)$ is exponentially stable on V. Furthermore, suppose that assumption* (2.6) *is satisfied.*

Then $G(s)$ defined by

$$G(s) := C(sI - A^V)^{-1}B + D \tag{2.34}$$

is well defined for all $\mathrm{Re}(s) \geq 0$ *and*

$$G(s) \in H^\infty(\mathbb{C}^+, \mathcal{L}(U, Y)). \tag{2.35}$$

Furthermore, if $u(\cdot) \in L_2(0, \infty; U)$, then $u(\cdot) \in U_{\mathrm{adm}}$; and if $x_0 = 0$ there holds

$$\hat{y}(s) = G(s)\hat{u}(s) \text{ for all } s \in \mathbb{C}^+ \tag{2.36}$$

and

$$\hat{y}(i\omega) = G(i\omega)\hat{u}(i\omega) \text{ for a.e. } \omega \in \mathbb{R}. \tag{2.37}$$

Proof. First of all we show that $G(s)$ is well defined for all s with $\mathrm{Re}(s) \geq 0$ and holomorphic on $\mathrm{Re}(s) \geq 0$ w.r.t. the topology of $\mathcal{L}(U, Y)$. Using the fact that $S(\cdot)$ is exponentially stable on V, the resolvent identity for $(sI - A^V)^{-1}$ gives

$$(sI - A^V)^{-1} = (\alpha I - A^V)^{-1} + (\alpha - s)(\alpha I - A^V)^{-1}(sI - A^V)^{-1} \tag{2.38}$$

for $\alpha, s \in \mathrm{Re}(s) \geq 0$. We know that $(sI - A^V)^{-1}$ is holomorphic on $\mathrm{Re}(s) \geq 0$ (w.r.t. the topology of $\mathcal{L}(V)$) and that $(\alpha I - A^V)^{-1} \in \mathcal{L}(V, W)$ for $\mathrm{Re}(\alpha) \geq 0$ (use the fact that $(\alpha I - A^V)^{-1} \in \mathcal{L}(V, D(A^V))$ and $D(A^V) \hookrightarrow W$). Hence we can conclude from (2.38) that $(sI - A^V)^{-1} \in \mathcal{L}(V, W)$ on $\mathrm{Re}(s) \geq 0$ and $(sI - A^V)^{-1}$ is holomorphic on $\mathrm{Re}(s) \geq 0$ w.r.t. the topology of $\mathcal{L}(V, W)$. Therefore, $G(s) = C(sI - A^V)^{-1}B + D$ is well defined ($(sI - A^V)^{-1}B$ maps into W) and holomorphic on $\mathrm{Re}(s) \geq 0$ w.r.t. the topology of $\mathcal{L}(U, Y)$.

Next we show (2.35) and (2.36): consider (2.5) with $x_0 = 0$. Since $S(\cdot)$ is exponentially stable on V, we know that for all $u(\cdot) \in L_2(0, \infty; U)$ we have $x(\cdot) \in L_2(0, \infty; V)$ and

$$\|x(\cdot)\|_{L_2(0,\infty;V)} \leq \text{const}\, \|u(\cdot)\|_{L_2(0,\infty;U)} \tag{2.39}$$

(this is well known; see, e.g., [8]). Hence, it follows from Lemma 2.7 that there exists some $c > 0$ such that for all $u(\cdot) \in L_2(0, \infty; U)$ there holds

$$\|y(\cdot)\|_{L_2(0,\infty;Y)} \leq c\, \|u(\cdot)\|_{L_2(0,\infty;U)}, \tag{2.40}$$

so the map from $u(\cdot) \in L_2(0,\infty;U)$ to $y(\cdot) \in L_2(0,\infty;Y)$ is linear and bounded. It is easy to see that this map is also shift invariant and so it follows from a well-known result (see, e.g., [7]) that there exists a (transfer) function $\tilde{G} : \mathbb{C}^+ \rightarrow \mathcal{L}(U,Y)$ such that

$$\hat{y}(s) = \tilde{G}(s)\hat{u}(s) \text{ for all } s \in \mathbb{C}^+ \tag{2.41}$$

and

$$\tilde{G}(\cdot) \in H^\infty(\mathbb{C}^+, \mathcal{L}(U,Y)). \tag{2.42}$$

The next step is to show that $\tilde{G}(s) = C(sI - A^V)^{-1}B + D$ on $\mathbb{C}^+$. We know that for $\text{Re}(s)$ large enough there holds

$$\tilde{G}(s) = C(sI - A^V)^{-1}B + D \tag{2.43}$$

(see, e.g., [3]). We have seen above that $G(s) = C(sI - A^V)^{-1}B + D$ is holomorphic on $\mathbb{C}^+$ w.r.t. the topology of $\mathcal{L}(U,Y)$. Since also $\tilde{G}(s)$ is holomorphic on $\mathbb{C}^+$ w.r.t. the topology of $\mathcal{L}(U,Y)$, we conclude (2.35) and (2.36).

Finally we show (2.37). Because $\hat{u}(\cdot) \in H_2(\mathbb{C}^+,U)$ and $\hat{y}(\cdot) \in H_2(\mathbb{C}^+,Y)$, it follows from Fatou's Theorem (see, e.g., [18, §4.6]) that for a.e. $\omega \in \mathbb{R}$,

$$\lim_{\epsilon \downarrow 0} \|\hat{u}(i\omega + \epsilon) - \hat{u}(i\omega)\|_U = 0 \text{ and } \lim_{\epsilon \downarrow 0} \|\hat{y}(i\omega + \epsilon) - \hat{y}(i\omega)\|_Y = 0. \tag{2.44}$$

(2.37) now follows from (2.36), (2.44), and the fact that $G(s)$ is holomorphic on $\text{Re}(s) \geq 0$ w.r.t. the topology on $\mathcal{L}(U,Y)$. □

Now we derive a frequency domain representation of (2.5), using Lemma 2.9.

LEMMA 2.10. *Let $\Sigma(S(\cdot),B,C,D)$ be a Pritchard–Salamon system of the form* (2.5)*, and suppose that (A^V,B) is exponentially stabilizable on V. Furthermore, suppose that assumption* (2.6) *is satisfied. If $u(\cdot) \in U_{\text{adm}}$ and $x_0 = 0$, then*

$$\hat{x}(i\omega) \in D(A^V) \textit{ for a.e. } \omega \in \mathbb{R}, \tag{2.45}$$

$$i\omega\hat{x}(i\omega) = A^V\hat{x}(i\omega) + B\hat{u}(i\omega) \textit{ for a.e. } \omega \in \mathbb{R}, \tag{2.46}$$

and

$$\hat{y}(i\omega) = C\hat{x}(i\omega) + D\hat{u}(i\omega) \textit{ for a.e. } \omega \in \mathbb{R}. \tag{2.47}$$

Proof. Since $u(\cdot) \in U_{\text{adm}}$ and (A^V,B) is exponentially stabilizable on V, we can use the perturbation results of Lemma 2.3 to infer that

$$x(t) = \int_0^t S_{BF}(t-s)B(u(s) - Fx(s))\,ds,$$

where $F \in \mathcal{L}(V,U)$ is such that $S_{BF}(\cdot)$ is exponentially stable on V.

We know that $x(\cdot) \in L_2(0,\infty;V)$ and

$$\hat{x}(i\omega) = (i\omega I - A^V_{BF})^{-1}B(\hat{u}(i\omega) - F\hat{x}(i\omega)) \text{ for a.e. } \omega \in \mathbb{R} \tag{2.48}$$

(see also Lemma 2.9). Now (2.45) and (2.46) follow from (2.48) and Lemma 2.3.

In Lemma 2.7 we saw that $y(\cdot) \in L_2(0,\infty;Y)$. We have $y(t) = Cx(t) + Du(t)$ and so Lemma 2.9 tells us that

$$\text{(2.49)} \quad \hat{y}(i\omega) = C(i\omega I - A^V_{BF})^{-1}B(\hat{u}(i\omega) - F\hat{x}(i\omega)) + D\hat{u}(i\omega) \text{ for a.e. } \omega \in \mathbb{R}.$$

The combination of (2.48) and (2.49) gives (2.47). □

The next lemma is a crucial result for the theory of optimal control. The result follows immediately from [11, Thm. 1.1].

LEMMA 2.11. *Let H be a real Hilbert space, let $T \in \mathcal{L}(H)$ be coercive, and let y be an arbitrary element of H. Then there exists a unique $x^* \in H$ such that*

$$\text{(2.50)} \qquad \langle Tx, x\rangle_H + \langle x, y\rangle_H \geq \langle Tx^*, x^*\rangle_H + \langle x^*, y\rangle_H \textit{ for all } x \in H.$$

The last lemma of this section deals with the optimal control problem for infinite-dimensional systems with bounded input and output operators. It is a complete generalization of the finite-dimensional result in [22]. The lemma is proved in full detail in [12], [13]. In the next section, we shall generalize this result to systems in the Pritchard–Salamon class.

LEMMA 2.12. *Let A be the infinitesimal generator of a semigroup $S(\cdot)$ on a real Hilbert space H; let $B \in \mathcal{L}(U,H)$, where U is another Hilbert space; and suppose that (A,B) is exponentially stabilizable. Furthermore, let $Q = Q^* \in \mathcal{L}(H)$, $L \in \mathcal{L}(H,U)$, and $R = R^* \in \mathcal{L}(U)$, with R coercive. For all $(x,u) \in H \times U$ we define the quadratic form*

$$\text{(2.51)} \qquad \mathcal{F}(x,u) := \langle Qx, x\rangle_H + 2\langle Lx, u\rangle_U + \langle Ru, u\rangle_U.$$

Consider the system

$$\text{(2.52)} \qquad x(t) = S(t)x_0 + \int_0^t S(t-s)Bu(s)\,ds, \; x_0 \in H, \; t \geq 0.$$

Define

$$U_{\text{adm}} := \{u(\cdot) \in L_2(0,\infty;U) \textit{ such that } x(\cdot) \textit{ given by } (2.52)$$

$$\text{(2.53)} \qquad \textit{satisfies } x(\cdot) \in L_2(0,\infty;H)\}.$$

For all $u(\cdot) \in U_{\text{adm}}$ define the cost functional

$$\text{(2.54)} \qquad J(x_0, u(\cdot)) := \int_0^\infty \mathcal{F}(x(t), u(t))\,dt.$$

Now the following are equivalent:

1. *For all $x_0 \in H$ there exists a unique $\bar{u}(\cdot) \in L_2(0,\infty;U)$ such that*

$$\text{(2.55)} \qquad \inf_{u(\cdot)\in U_{\text{adm}}} J(x_0,u(\cdot)) = \min_{u(\cdot)\in U_{\text{adm}}} J(x_0,u(\cdot)) = J(x_0,\bar{u}(\cdot)).$$

2. *There exists a selfadjoint $X \in \mathcal{L}(H)$ such that for all $x, y \in D(A)$*

$$\langle Ax, Xy\rangle_H + \langle Xx, Ay\rangle_H$$

$$\text{(2.56)} \qquad - \langle (B^*X + L)^*R^{-1}(B^*X + L)x, y\rangle_H + \langle Qx, y\rangle_H = 0$$

and $A - BR^{-1}(B^*X + L)$ is the infinitesimal generator of an exponentially stable semigroup.

3. *There exists an* $\epsilon > 0$ *such that for all* $(\omega, x, u) \in \mathbb{R} \times D(A) \times U$ *that satisfy* $i\omega x = Ax + Bu$ *there holds*

$$\mathcal{F}(x,u) \geq \epsilon(\|x\|_H^2 + \|u\|_U^2). \tag{2.57}$$

Furthermore, if one of these conditions holds, the minimizing $\bar{u}(\cdot)$ *in* (2.55) *can be given in feedback form:*

$$\bar{u}(\cdot) = -R^{-1}(B^*X + L)x(\cdot), \tag{2.58}$$

and

$$\inf_{u(\cdot)\in U_{\text{adm}}} J(x_0, u(\cdot)) = J(x_0, \bar{u}(\cdot)) = \langle x_0, Xx_0\rangle_H. \tag{2.59}$$

Finally, if the X *in* 2 *exists, it is unique.*

3. Problem formulation and main result. The purpose of this section is to find a generalization of Lemma 2.12 to the class of systems as presented in §2. So let W and V be two Hilbert spaces that satisfy (2.1), and suppose that we have a C_0-semigroup $S(\cdot)$ on both spaces. Furthermore, let U, Y_1, and Y_2 be Hilbert spaces (as before U will play the role of the input space, Y_1 and Y_2 will be output spaces). Let $B \in \mathcal{L}(U, V)$ be an admissible input operator and $C_1 \in \mathcal{L}(W, Y_1), C_2 \in \mathcal{L}(W, Y_2)$, and $L \in \mathcal{L}(W, U)$ be admissible output operators.

For all $(x, u) \in W \times U$ we define the quadratic form

$$\mathcal{F}(x,u) := \langle C_1x, C_1x\rangle_{Y_1} - \langle C_2x, C_2x\rangle_{Y_2} + 2\langle Lx, u\rangle_U + \langle Ru, u\rangle_U, \tag{3.1}$$

where $R = R^* \in \mathcal{L}(U)$ is coercive. In the quadratic form $\langle Qx, x\rangle_H$ in (2.51), Q could have been expressed as $Q = C_1^*C_1 - C_2^*C_2$ for some $C_1 \in \mathcal{L}(H, Y)$ and $C_2 \in \mathcal{L}(H, Y)$ (use $Q = \text{const}I - (\text{const}I - Q)$ with const large enough). Therefore, (3.1) represents an appropriate generalization of (2.51). It may not be the most general quadratic criterion for Pritchard–Salamon systems, but it does contain the interesting applications (see Remark 3.3).

Our system is given by

$$x(t) = S(t)x_0 + \int_0^t S(t-s)Bu(s)\,ds, \tag{3.2}$$

where $x_0 \in V$ and $t \geq 0$; and we assume that

$$(A^V, B) \text{ is exponentially stabilizable on } V. \tag{3.3}$$

We define the class of admissible inputs as

$$U_{\text{adm}} := \{u(\cdot) \in L_2(0,\infty; U) \text{ such that } x(\cdot) \text{ given by (3.2)}$$

$$\text{satisfies } x(\cdot) \in L_2(0,\infty; V)\}. \tag{3.4}$$

From Lemma 2.7 it follows that for all $u(\cdot) \in U_{\text{adm}}$, the cost functional $J(x_0, u(\cdot))$ defined by

$$J(x_0, u(\cdot)) := \int_0^\infty \mathcal{F}(x(t), u(t))\,dt \tag{3.5}$$

is finite. Note that $\mathcal{F}(x(t), u(t))$ should not be interpreted pointwise but as explained in Remark 2.2:

$$\int_0^\infty \mathcal{F}(x(t), u(t))\,dt = \|C_1 x(\cdot)\|^2_{L_2(0,\infty;Y_1)} - \|C_2 x(\cdot)\|^2_{L_2(0,\infty;Y_2)}$$

$$+\, 2\langle Lx(\cdot), u(\cdot)\rangle_{L_2(0,\infty;U)} + \langle Ru(\cdot), u(\cdot)\rangle_{L_2(0,\infty;U)}. \tag{3.6}$$

The following result is a complete generalization of Lemma 2.12.

THEOREM 3.1. *Suppose that we have a Pritchard–Salamon system of the form* (3.2) *that satisfies the assumptions* (2.6) *and* (3.3). *Furthermore, let $\mathcal{F}$ be defined as in* (3.1). *The following three conditions are equivalent:*

1. *For all $x_0 \in V$ there exists a unique $\bar{u}(\cdot) \in L_2(0,\infty;U)$ such that*

$$\inf_{u(\cdot)\in U_{\mathrm{adm}}} J(x_0, u(\cdot)) = \min_{u(\cdot)\in U_{\mathrm{adm}}} J(x_0, u(\cdot)) = J(x_0, \bar{u}(\cdot)). \tag{3.7}$$

2. *There exists a selfadjoint $X \in \mathcal{L}(V)$ such that for all $x, y \in D(A^V)$*

$$\langle (A^V - BR^{-1}L)x, Xy\rangle_V + \langle Xx, (A^V - BR^{-1}L)y\rangle_V$$

$$-\langle XBR^{-1}B^*Xx, y\rangle_V - \langle R^{-1}Lx, Ly\rangle_U$$

$$+\, \langle C_1 x, C_1 y\rangle_{Y_1} - \langle C_2 x, C_2 y\rangle_{Y_2} = 0 \tag{3.8}$$

and $A^V - BR^{-1}(B^*X + L)$ is the generator of a C_0-semigroup, exponentially stable on V.

3. *There exists an $\epsilon > 0$ such that for all $(\omega, x, u) \in \mathbb{R} \times D(A^V) \times U$ that satisfy $i\omega x = A^V x + Bu$ there holds*

$$\mathcal{F}(x, u) \geq \epsilon(\|x\|^2_V + \|u\|^2_U). \tag{3.9}$$

Furthermore, if one of these conditions holds, the minimizing $\bar{u}(\cdot)$ in (3.7) *can be given in feedback form:*

$$\bar{u}(\cdot) = -R^{-1}(B^*X + L)x(\cdot), \tag{3.10}$$

and

$$\inf_{u(\cdot)\in U_{\mathrm{adm}}} J(x_0, u(\cdot)) = J(x_0, \bar{u}(\cdot)) = \langle x_0, Xx_0\rangle_V. \tag{3.11}$$

Finally, if the X in (3.8) *exists, it is unique.*

Remark 3.2. In [15] the authors assume the existence of a third Hilbert space H such that $W \hookrightarrow H \hookrightarrow V$ and $S(\cdot)$ is also a C_0-semigroup on H. In applications H is often the actual state-space so that $B \in \mathcal{L}(U, V)$ and $C \in \mathcal{L}(W, Y)$ are both unbounded with respect to this state-space (see [15]). The role of H is not important in the derivation of Theorem 3.1 because we consider the LQ-problem with initial conditions in V and with stability on V, just as in [15]. Identifying H with its dual we can use H as a pivot space, in order to obtain a "strong form" of the Riccati equation, also just as in [15]:

Using $H = H'$ as a pivot space we have $D(A^V) \hookrightarrow W \hookrightarrow H \hookrightarrow V$ and $V' \hookrightarrow H \hookrightarrow W' \hookrightarrow (D(A^V))'$. We let $i : V \to V'$ denote the canonical isometric isomorphism, and the duality pairing of V is denoted by $\langle \cdot, \cdot \rangle_{V,V'}$. Furthermore, we identify $U = U'$, $Y_1 = Y_1'$, and $Y_2 = Y_2'$. Then $X \in \mathcal{L}(V)$ satisfies (3.8) if and only if $\tilde{X} = iX \in \mathcal{L}(V, V')$ satisfies the equation

$$\langle (A^V - BR^{-1}L)x, \tilde{X}y \rangle_{V,V'} + \langle x, \tilde{X}(A^V - BR^{-1}L)y \rangle_{V,V'}$$

$$- \langle x, \tilde{X}BR^{-1}B'\tilde{X}y \rangle_{V,V'} - \langle R^{-1}Lx, Ly \rangle_U$$

$$+ \langle C_1 x, C_1 y \rangle_{Y_1} - \langle C_2 x, C_2 y \rangle_{Y_2} = 0 \text{ for all } x, y \in D(A^V). \tag{3.12}$$

Equation (3.12) can in turn be reformulated as

$$\Big((A^V - BR^{-1}L)'\tilde{X} + \tilde{X}(A^V - BR^{-1}L)$$

$$- \tilde{X}BR^{-1}B'\tilde{X} - L'R^{-1}L + C_1'C_1 - C_2'C_2 \Big) x = 0 \tag{3.13}$$

in $(D(A^V))'$ for all $x \in D(A^V)$. Here $(A^V - BR^{-1}L)' \in \mathcal{L}(V', (D(A^V))')$ can be considered as the dual of the bounded linear operator $(A^V - BR^{-1}L) \in \mathcal{L}(D(A^V), V)$, as explained in [15]. Of course, the other statements of the theorem can be reformulated in terms of $\tilde{X}$ in the obvious way.

Finally, we note that because of $H \hookrightarrow V$, $V' \hookrightarrow H$, and the fact that $\tilde{X} \in \mathcal{L}(V, V')$, there holds $\tilde{X} \in \mathcal{L}(H)$ and that for all $x_0 \in H$ we have

$$\inf_{u(\cdot) \in U_{\mathrm{adm}}} J(x_0, u(\cdot)) = \langle x_0, Xx_0 \rangle_V = \langle x_0, \tilde{X}x_0 \rangle_{V,V'} = \langle x_0, \tilde{X}x_0 \rangle_H.$$

Remark 3.3. The first results for the infinite-dimensional LQ-problem were obtained for the case that $L = 0$ and $C_2 = 0$ assuming some exponential detectability condition for C_1 (see, e.g., [4], [1], [11] and references therein). It is not difficult to show that if in our setting $L = 0$ and $C_2 = 0$, then the frequency domain inequality is *implied* by the assumption that there exists a $G \in \mathcal{L}(Y_1, V)$ such that $S_{GC_1}(\cdot)$ is exponentially stable on V. In fact, if the system is finite dimensional, then the frequency domain inequality is in this case equivalent to the statement that the poles of A on the imaginary axis are detectable by C_1 (this is, for instance, the case if $\sigma(A) \cap i\mathbb{R} = \emptyset$). Thus, the frequency domain inequality is much weaker than the exponential detectability assumption. It is straightforward to show that if

$$\sup_{\omega \in \mathbb{R}} \|(i\omega I - A^V)^{-1}\|_{\mathcal{L}(V)} < \infty, \tag{3.14}$$

then the frequency domain condition of item 3 is equivalent to the condition that for all $\omega \in \mathbb{R}$ and all $u \in U$

$$\mathcal{F}((i\omega I - A^V)^{-1}Bu, u) \geq \epsilon_1 \|u\|_U^2 \tag{3.15}$$

for some $\epsilon_1 > 0$. It follows that if $L = 0, C_1 = 0, R = I$, and $S(\cdot)$ is exponentially stable on V, then Theorem 3.1 corresponds to the *bounded real lemma* (this was also considered in [17] for the Pritchard–Salamon class) and if $C_1 = C_2 = 0, R = D + D^*$ it corresponds to the *positive real lemma.*

In the case that $C_2 = 0, R = D_{12}^* D_{12}$, and $L = D_{12}^* C_1$, the frequency domain inequality in (3.9) corresponds to the "invariant zeros condition" related to the H_∞-control problem with state-feedback in [8]. Theorem 3.1 can be used to extend the results in [8] to the Pritchard–Salamon class (see [9]).

Remark 3.4. Theorem 3.1 is a considerable extension of existing results about LQ-theory for systems in the Pritchard–Salamon class:

In [15] the authors consider the infinite horizon LQ-problem (without a priori stability requirements on the state $x(\cdot)$) for the special case that $C_2 = 0$ and $L = 0$, and they prove that the cost can be made finite for any initial condition if and only if there exists a nonnegative solution $\tilde{X} \in \mathcal{L}(V, V')$ to the corresponding Riccati equation (see Remark 3.2 and (3.8) with $C_2 = L = 0$). The result in Theorem 3.1 is different because we always assume that the control is such that $x(\cdot) \in L_2(0, \infty; V)$. However, under some additional detectability assumption it is proved in [15] that their optimal control given by $\bar{u}(\cdot) = -R^{-1}B'\tilde{X}x(\cdot)$ is exponentially stabilizing on V. This result corresponds to the equivalence between 1 and 2 in Theorem 3.1 with $C_2 = 0$ and $L = 0$.

In [17], the equivalence between 1, 2, and 3 is proved for the special case that $C_1 = 0, L = 0$, and $S(\cdot)$ is exponentially stable on V and W.

Proof. We shall prove the implications 1 ⇒ 2, 2 ⇒ 3, and 3 ⇒ 1.

Proof of 1 ⇒ 2. We first assume that $L = 0$ (later we remove this extra assumption by applying some preliminary feedback). The idea is to use Lemma 2.7 and in particular, formula (2.33), to transform the criterion (3.5) in such a way that the new criterion can be treated using Lemma 2.12. Since (A^V, B) is exponentially stabilizable on V, there exists some $\bar{F} \in \mathcal{L}(V, U)$ such that $S_{B\bar{F}}(\cdot)$ is exponentially stable on V. Using Lemma 2.6 it follows that there exist nonnegative definite operators $X_1, X_2 \in \mathcal{L}(V)$ such that for all $x, y \in D(A^V) = D(A^V_{B\bar{F}})$

$$\begin{aligned}&\langle A^V x, X_1 y\rangle_V + \langle x, X_1 A^V y\rangle_V \\ &= -\langle C_1 x, C_1 y\rangle_{Y_1} - \langle (X_1 B\bar{F} + \bar{F}^* B^* X_1)x, y\rangle_V \end{aligned} \tag{3.16}$$

and

$$\begin{aligned}&\langle A^V x, X_2 y\rangle_V + \langle x, X_2 A^V y\rangle_V \\ &= -\langle C_2 x, C_2 y\rangle_{Y_2} - \langle (X_2 B\bar{F} + \bar{F}^* B^* X_2)x, y\rangle_V. \end{aligned} \tag{3.17}$$

Since we assume that $L = 0$, we have

$$J(x_0, u(\cdot)) = \int_0^\infty \left(\|C_1 x(t)\|^2_{Y_1} - \|C_2 x(t)\|^2_{Y_2} + \left\| R^{\frac{1}{2}} u(t) \right\|^2_U \right) dt.$$

Hence, we can use formula (2.33) in the proof of Lemma 2.7 to infer that

$$J(x_0, u(\cdot)) = \langle x_0, (X_1 - X_2)x_0\rangle_V$$

$$+ \int_0^\infty \Big(\langle ((X_2 - X_1)B\bar{F} + \bar{F}^* B^* (X_2 - X_1))x(t), x(t)\rangle_V$$

$$+ 2\langle B^*(X_1 - X_2)x(t), u(t)\rangle_U + \left\|R^{\frac{1}{2}}u(t)\right\|_U^2 \Bigg) dt. \tag{3.18}$$

Note that the integral term is as in Lemma 2.12 (formulas (2.51) and (2.54)), with Q replaced by $(X_2 - X_1)B\bar{F} + \bar{F}^*B^*(X_2 - X_1)$ and L replaced by $B^*(X_1 - X_2)$. Hence, because of our assumption (1), we can apply this lemma on V and infer the existence of a selfadjoint $X_3 \in \mathcal{L}(V)$ such that for all $x, y \in D(A^V)$

$$\langle A^V x, X_3 y\rangle_V + \langle X_3 x, A^V y\rangle_V$$

$$-\langle (X_3 + X_1 - X_2)BR^{-1}B^*(X_3 + X_1 - X_2)x, y\rangle_V$$

$$+ \langle ((X_2 - X_1)B\bar{F} + \bar{F}^*B^*(X_2 - X_1))x, y\rangle_V = 0 \tag{3.19}$$

and $A^V - BR^{-1}B^*(X_3 + X_1 - X_2)$ is the infinitesimal generator of an exponentially stable semigroup on V. Combining (3.16), (3.17), and (3.19), with $X := X_3 + X_1 - X_2$, we conclude that $X \in \mathcal{L}(V)$ is selfadjoint and satisfies

$$\langle A^V x, Xy\rangle_V + \langle Xx, A^V y\rangle_V - \langle XBR^{-1}B^*Xx, y\rangle_V$$

$$+ \langle C_1 x, C_1 y\rangle_{Y_1} - \langle C_2 x, C_2 y\rangle_{Y_2} = 0 \tag{3.20}$$

for all $x, y \in D(A^V)$ and $A^V - BR^{-1}B^*X$ is the generator of an exponentially stable semigroup on V. We note that (3.20) is just (3.8) with $L = 0$.

Next we show how to reduce the general case to the case where $L = 0$. We remove the "cross term" of the cost criterion by applying the preliminary feedback

$$u(\cdot) = Fx(\cdot) + v(\cdot), \tag{3.21}$$

where $F \in \mathcal{L}(W, U)$ is given by

$$F = -R^{-1}L. \tag{3.22}$$

To make this more precise we note that F given by (3.22) is an admissible output operator because L is, and we define the transformed system

$$x_F(t) = S_{BF}(t)x_0 + \int_0^t S_{BF}(t-s)Bv(s)\, ds \tag{3.23}$$

just as in Lemma 2.5. It follows from Lemma 2.3 that C_1, C_2, and L are admissible output operators for this transformed system as well. Furthermore, it follows from Remark 2.4 that (A^V_{BF}, B) is exponentially stabilizable on V and Lemma 2.3 shows that $D(A^V_{BF}) \hookrightarrow V$. Hence, the transformed system also satisfies the a priori assumptions of Theorem 3.1. We define the class of admissible inputs for this system as

$$\bar{U}_{\mathrm{adm}} := \{v(\cdot) \in L_2(0, \infty; U) \text{ such that } x_F(\cdot) \text{ given by (3.23)}$$

$$\text{satisfies } x_F(\cdot) \in L_2(0, \infty; V)\}. \tag{3.24}$$

Using the perturbation results in Lemma 2.5, we see that if $u(\cdot)$ and $v(\cdot)$ are related by (3.21) (or, more specifically, by (2.12) and (2.15)), we have $x(\cdot) = x_F(\cdot)$, $u(\cdot) = Fx_F(\cdot) + v(\cdot)$, and $v(\cdot) = -Fx(\cdot) + u(\cdot)$. Hence, we conclude that $u(\cdot) \in U_{\text{adm}}$ if and only if $v(\cdot) \in \bar{U}_{\text{adm}}$. Next, for all $v(\cdot) \in \bar{U}_{\text{adm}}$, we define the transformed cost function

$$J_F(x_0, v(\cdot)) := \int_0^\infty \mathcal{F}_F(x_F(t), v(t))\, dt,$$

where for $x \in W$ and $v \in U$ we have

$$\mathcal{F}_F(x, v) = \mathcal{F}(x, Fx + v) = \langle C_1 x, C_1 x\rangle_{Y_1} - \langle C_2 x, C_2 x\rangle_{Y_2}$$

$$+2\langle Lx, Fx + v\rangle_U + \langle R(Fx + v), (Fx + v)\rangle_U,$$

so because of $F = -R^{-1}L$, we have

$$\mathcal{F}_F(x, v) =$$

$$\langle C_1 x, C_1 x\rangle_{Y_1} - \langle C_2 x, C_2 x\rangle_{Y_2} - \langle R^{-\frac{1}{2}} Lx, R^{-\frac{1}{2}} Lx\rangle_U + \langle Rv, v\rangle_U.$$

Note that the transformed cost function has the same form as in (3.5), with $L = 0$ and $\langle C_2 x, C_2 x\rangle_{Y_2}$ replaced by $\langle C_2 x, C_2 x\rangle_{Y_2} + \langle R^{-\frac{1}{2}} Lx\ , R^{-\frac{1}{2}} Lx\rangle_U$. Using Lemma 2.5 and the above, we have

$$J(x_0, u(\cdot)) = J_F(x_0, v(\cdot)).$$

Now 1 implies that there exists a unique $\bar{v}(\cdot) \in L_2(0, \infty; U)$ such that

$$\inf_{v(\cdot)\in\bar{U}_{\text{adm}}} J_F(x_0, v(\cdot)) = \min_{v(\cdot)\in\bar{U}_{\text{adm}}} J_F(x_0, v(\cdot)) = J_F(x_0, \bar{v}(\cdot)). \tag{3.25}$$

We have proved the implication 1 ⇒ 2 under the assumption that there is no "cross term" (i.e., $L = 0$), so we conclude from (3.25) that there exists a selfadjoint $X \in \mathcal{L}(V)$ such that for all $x, y \in D(A^V_{BF}) = D(A^V)$, we have

$$\langle A^V_{BF} x, Xy\rangle_V + \langle Xx, A^V_{BF} y\rangle_V - \langle XBR^{-1}B^* Xx, y\rangle_V$$

$$+ \langle C_1 x, C_1 y\rangle_{Y_1} - \langle C_2 x, C_2 y\rangle_{Y_2} - \langle R^{-\frac{1}{2}} Lx, R^{-\frac{1}{2}} Lx\rangle_U = 0 \tag{3.26}$$

and $A^V_{BF} - BR^{-1}B^* X$ is the generator of an exponentially stable semigroup on V.

Since $A^V_{BF} = A^V - BR^{-1}L$, we have proved 2.

Proof of 2 ⇒ 3. Using 2, it is straightforward to show that for all $(\omega, x, u) \in \mathbb{R} \times D(A^V) \times U$ that satisfy $i\omega x = A^V x + Bu$ there holds

$$\mathcal{F}(x, u) = \left\| R^{\frac{1}{2}}(u + R^{-1}(B^* X + L)x)\right\|_U^2. \tag{3.27}$$

Now since $A^V - BR^{-1}(B^* X + L)$ is the generator of an exponentially stable semigroup on V, we see that for all $(\omega, x, u) \in \mathbb{R} \times D(A^V) \times U$ that satisfy $i\omega x = A^V x + Bu$ there holds

$$x = (i\omega I - A^V + BR^{-1}(B^* X + L))^{-1} B(u + R^{-1}(B^* X + L)x) \tag{3.28}$$

and so

$$\|x\|_V \leq c_1 \left\|u + R^{-1}(B^*X + L)x\right\|_U \tag{3.29}$$

for some $c_1 > 0$. Combination of (3.27) and (3.29) (using the fact that R is coercive) implies the existence of some $\epsilon_1 > 0$ such that for all $(\omega, x, u) \in \mathbb{R} \times D(A^V) \times U$ that satisfy $i\omega x = A^V x + Bu$,

$$\mathcal{F}(x,u) \geq \epsilon_1 \|x\|_V^2 . \tag{3.30}$$

For $u \in U$ and $x \in W$ we have

$$u = u + R^{-1}(B^*X + L)x - R^{-1}(B^*X + L)x$$

so that

$$\|u\|_U \leq \left\|R^{-1}(B^*X + L)x\right\|_U + \left\|u + R^{-1}(B^*X + L)x\right\|_U . \tag{3.31}$$

Since $A^V - BR^{-1}(B^*X + L)$ is the generator of an exponentially stable semigroup on V and L is an admissible output operator, we conclude from (3.28) and Lemma 2.9, formula (2.35), that for all $(\omega, x, u) \in \mathbb{R} \times D(A^V) \times U$ that satisfy $i\omega x = A^V x + Bu$ there holds

$$\|Lx\|_U \leq c_2 \left\|u + R^{-1}(B^*X + L)x\right\|_U \tag{3.32}$$

for some $c_2 > 0$. Since $R^{-1}B^*X \in \mathcal{L}(V, U)$, the combination of (3.31) and (3.32) gives

$$\|u\|_U \leq c_3(\|x\|_V + \left\|u + R^{-1}(B^*X + L)x\right\|_U)$$

for some $c_3 > 0$. We can combine this with (3.29) to obtain

$$\|u\|_U \leq c_4 \left\|u + R^{-1}(B^*X + L)x\right\|_U \tag{3.33}$$

for some $c_4 > 0$.

Finally, (3.33), (3.27), and the fact that R is coercive imply the existence of some $\epsilon_2 > 0$ such that

$$\mathcal{F}(x,u) \geq \epsilon_2 \|u\|_U^2 . \tag{3.34}$$

Now 3 follows from (3.30) and (3.34).

Proof of 3 ⇒ 1. The idea is to apply Lemma 2.11 (a similar procedure is used in [12] and [21] for the bounded case). Let $x_0 \in V$ be given. We define

$$H := \Bigg\{ (x(\cdot), u(\cdot)) \in L_2(0,\infty; V) \times L_2(0,\infty; U) \text{ such that}$$

$$u(\cdot) \in U_{\text{adm}} \text{ and } x(t) = S(t)x_0 + \int_0^t S(t-s)Bu(s)\,ds \Bigg\} \tag{3.35}$$

and

$$H_0 := \Bigg\{ (x(\cdot), u(\cdot)) \in L_2(0,\infty; V) \times L_2(0,\infty; U) \text{ such that}$$

$$u(\cdot) \in U_{\text{adm}} \text{ and } x(t) = \int_0^t S(t-s)Bu(s)\,ds \Big\}. \tag{3.36}$$

It is straightforward to show that H_0 is a closed subspace of $L_2(0,\infty;V) \times L_2(0,\infty;U)$ and so H_0 is a Hilbert space (with the obvious inner product determined by the inner products of $L_2(0,\infty;V)$ and $L_2(0,\infty;U)$). Since (A^V, B) is exponentially stabilizable on V, there exists an $F \in \mathcal{L}(V,U)$ such that $S_{BF}(\cdot)$ is exponentially stable on V. We define

$$(x_0(\cdot), u_0(\cdot)) := (S_{BF}(\cdot)x_0, FS_{BF}(\cdot)x_0).$$

Using the perturbation results of Lemma 2.3, it is easy to see that $(x_0(\cdot), u_0(\cdot)) \in H$ and that $H = H_0 + (x_0(\cdot), u_0(\cdot))$ and we have

$$\inf_{u(\cdot)\in U_{\text{adm}}} J(x_0, u(\cdot)) = \inf_{(x(\cdot),u(\cdot))\in H} \int_0^\infty \mathcal{F}(x(t), u(t))\,dt$$

$$= \inf_{(x(\cdot),u(\cdot))\in H_0} \int_0^\infty \mathcal{F}(x(t)+x_0(t), u(t)+u_0(t))\,dt. \tag{3.37}$$

Now for $(x(\cdot), u(\cdot)) \in H_0$ we have

$$\begin{aligned}
&\int_0^\infty \mathcal{F}(x(t)+x_0(t), u(t)+u_0(t))\,dt \\
&\quad= \int_0^\infty (\|C_1x(t) + C_1x_0(t)\|_{Y_1}^2 - \|C_2x(t) + C_2x_0(t)\|_{Y_2}^2 \\
&\qquad + 2\langle L(x(t)+x_0(t)), u(t)+u_0(t)\rangle_U + \langle R(u(t)+u_0(t)), u(t)+u_0(t)\rangle_U)dt \\
&\quad= \int_0^\infty \mathcal{F}(x(t), u(t)) + 2\langle C_1x(\cdot), C_1x_0(\cdot)\rangle_{L_2(0,\infty;Y_1)} \\
&\qquad -2\langle C_2x(\cdot), C_2x_0(\cdot)\rangle_{L_2(0,\infty;Y_2)} + 2\langle Lx(\cdot), u_0(\cdot)\rangle_{L_2(0,\infty;U)} \\
&\qquad +2\langle Lx_0(\cdot), u(\cdot)\rangle_{L_2(0,\infty;U)} + 2\langle Ru(\cdot), u_0(\cdot)\rangle_{L_2(0,\infty;U)} + c
\end{aligned} \tag{3.38}$$

for some $c \in \mathbb{R}$. Next we infer that there exists a selfadjoint operator $T \in \mathcal{L}(H_0)$ and some element $y \in H_0$ such that (3.38) can be reformulated as

$$\int_0^\infty \mathcal{F}(x(t)+x_0(t), u(t)+u_0(t))\,dt = \langle Tx, x\rangle_{H_0} + \langle x, y\rangle_{H_0} + c, \tag{3.39}$$

where $x := (x(\cdot), u(\cdot)) \in H_0$. Indeed, it follows from Lemma 2.7 that for any admissible $C \in \mathcal{L}(W,Y)$ the map from H_0 to $L_2(0,\infty;Y)$ determined by $(x(\cdot), u(\cdot)) \mapsto Cx(\cdot)$ is linear and bounded. Since C_1, C_2, and L are admissible output operators and R is bounded, it is straightforward to conclude from (3.6) that we have a selfadjoint operator $T \in \mathcal{L}(H_0)$, such that for all $x := (x(\cdot), u(\cdot)) \in H_0$ there holds

$$\int_0^\infty \mathcal{F}(x(t), u(t)) = \langle Tx, x\rangle_{H_0}. \tag{3.40}$$

Similarly, one can show that in (3.38),

$$2\langle C_1x(\cdot), C_1x_0(\cdot)\rangle_{L_2(0,\infty;Y_1)} - 2\langle C_2x(\cdot), C_2x_0(\cdot)\rangle_{L_2(0,\infty;Y_2)}$$

$$+ 2\langle Lx(\cdot), u_0(\cdot)\rangle_{L_2(0,\infty;U)} + 2\langle Lx_0(\cdot), u(\cdot)\rangle_{L_2(0,\infty;U)}$$

$$+ 2\langle Ru(\cdot), u_0(\cdot)\rangle_{L_2(0,\infty;U)}$$

is equal to $\langle x, y\rangle_{H_0}$ for some $y \in H_0$. Hence we have (3.39).

It follows from (3.37) and the above that

$$\inf_{u(\cdot)\in U_{\text{adm}}} J(x_0, u(\cdot)) = \inf_{x\in H_0} \langle Tx, x\rangle_{H_0} + \langle x, y\rangle_{H_0} + c. \tag{3.41}$$

To apply Lemma 2.11, we still have to show that T in (3.40) is coercive (of course using the frequency domain inequality of 3). Both components of $(x(\cdot), u(\cdot)) \in H_0$ are Fourier transformable, and it follows from Lemma 2.10 that $\hat{x}(i\omega) \in D(A^V)$ for a.e. $\omega \in \mathbb{R}$ and

$$\begin{aligned} i\omega\hat{x}(i\omega) &= A^V\hat{x}(i\omega) + B\hat{u}(i\omega), \\ \widehat{Cx}(i\omega) &= C\hat{x}(i\omega) \end{aligned}$$

for a.e. $\omega \in \mathbb{R}$, for any admissible $C \in \mathcal{L}(W, Y)$. Hence, for all $x = (x(\cdot), u(\cdot))\in H_0$ we have

$$\langle Tx, x\rangle_{H_0} = \int_0^\infty \mathcal{F}(x(t), u(t))\, dt$$

$$= \frac{1}{2\pi}\int_{-\infty}^{\infty} \mathcal{F}(\hat{x}(i\omega), \hat{u}(i\omega))\, d\omega \qquad \text{(Plancherel's Theorem, see [18, §4.8])}$$

$$\geq \epsilon\frac{1}{2\pi}\int_{-\infty}^{\infty} (\|\hat{x}(i\omega)\|_V^2 + \|\hat{u}(i\omega)\|_U^2)\, d\omega \qquad \text{(using 3)}$$

$$= \epsilon\int_0^\infty (\|x(t)\|_V^2 + \|u(t)\|_U^2)\, dt \qquad \text{(again using Plancherel's Theorem)}$$

$$= \epsilon\, \|(x(\cdot), u(\cdot))\|_{H_0}^2 = \epsilon\, \|x\|_{H_0}^2 .$$

This proves that T is coercive and so Lemma 2.11 implies 1.

Finally, it follows from Lemma 2.12 and the above that the optimal control $\bar{u}(\cdot)$ is given by the state-feedback $\bar{u}(\cdot) = -R^{-1}(B^*X + L)x(\cdot)$ (cf. Lemma 2.3) and that (3.11) is satisfied. The uniqueness of X also follows from that above and Lemma 2.12. □

4. Examples and conclusions. We have solved the nonstandard LQ-problem with stability for the Pritchard–Salamon class. This represents a considerable extension of existing LQ-results for this class in [15] and [17]. As explained in the introduction, many interesting system theoretical problems can be formulated in the nonstandard LQ-framework, for instance, questions arising in optimal control, identification theory, and robust stability theory (see Willems [22]) and so Theorem 3.1 is an important result. Also the solution to the H_∞-control problem for Pritchard–Salamon systems can be solved using Theorem 3.1. In particular, if $C_2 = 0, R = D_{12}^*D_{12}$, and

$L = D_{12}^* C_1$, the frequency domain inequality corresponds to the "invariant zeros condition" in [9, Theorems 4.4, 5.4].

Upon the request of one reviewer we shall now discuss some examples of the theory presented in this paper, in the sense that we shall consider some nonstandard LQ-problems for systems in the Pritchard–Salamon class that are not covered by the bounded theory in [12], [13], [21]. As pointed out in [15], [16] and the introduction, the Pritchard–Salamon class contains many delay systems and several PDE systems. Since the strength of the Pritchard–Salamon class lies in the fact that so many delay systems fit in, we shall extensively treat a delay example, briefly discuss a parabolic PDE example, and make some comments about a flexible beam example.

4.1. Delay systems. The following neutral system with output delays is treated in [15] (for more general delay systems in the Pritchard–Salamon class, like retarded and neutral systems with delays in both the input and the output, we refer to [16], [19]). Consider the system given by

$$(4.1) \qquad \begin{cases} \dfrac{d}{dt}(z(t) - Mz_t) & = & \bar{L}z_t + \bar{B}u(t), \\ y(t) & = & \bar{C}z_t, \end{cases}$$

where $z(t) \in \mathbb{R}^n$; $u(t) \in \mathbb{R}^m$; $y(t) \in \mathbb{R}^p$; z_t is the solution segment defined by

$$z_t(\tau) = z(t+\tau)\ , \quad -h \leq \tau \leq 0\ , \quad h > 0;$$

and $\bar{B} \in \mathbb{R}^{n \times m}$ and $\bar{L}, M, \bar{C}$ are bounded linear functionals from $C(-h, 0\ ;\mathbb{R}^n)$ into $\mathbb{R}^n$ and $\mathbb{R}^p$, respectively. Under some conditions (see [15, Exe. 4.1]), the system given by (4.1) has a unique solution $z(t)$; $t \geq -h$, for every input $u(\cdot) \in L_2^{\mathrm{loc}}(0, \infty; \mathbb{R}^m)$ and every initial condition satisfying

$$\lim_{t \downarrow 0}(z(t) - Mz_t) = \eta_0, \qquad z(\tau) = \phi_0(\tau), \quad -h < \tau < 0,$$

where $x_0 = (\eta_0, \phi_0) \in M_2 = \mathbb{R}^n \times L_2(-h, 0\ ;\mathbb{R}^n)$. Moreover, the evolution of the state

$$x(t) = (z(t) - Mz_t, z_t) \in M_2$$

of the system can be described by

$$(4.2) \qquad x(t) = S(t)x_0 + \int_0^t S(t-s)Bu(s)\,ds,$$

where $B \in \mathcal{L}(\mathbb{R}^m, M_2)$ maps $u \in \mathbb{R}^m$ into the pair $Bu = (\bar{B}u, 0)$ and $S(t) \in \mathcal{L}(M_2)$ is the C_0-semigroup generated by the operator A given by

$$(4.3) \qquad \begin{array}{rcl} D(A) & = & \{x = (\eta, \phi) \in M_2 \mid \phi \in W^{1,2}, \eta = \phi(0) - M\phi\}, \\ Ax & = & (\bar{L}\phi, \dot{\phi}). \end{array}$$

Here $W^{1,2}$ denotes the Sobolev space $W^{1,2}(-h, 0\ ;\mathbb{R}^n)$ of absolute continuous functions in $L_2(-h, 0\ ;\mathbb{R}^n)$ whose derivative is in $L_2(-h, 0\ ;\mathbb{R}^n)$.

Now $D(A)$ can be considered as a Hilbert space by choosing the inner product

$$\langle(\eta, \phi), (\bar{\eta}, \bar{\phi})\rangle_{D(A)} = \langle\phi, \bar{\phi}\rangle_{W^{1,2}},$$

and it follows that $S(\cdot)$ restricts to a C_0-semigroup on $D(A)$.

The output $y(t) = \bar{C}z_t$ of the system may formally be described by

$$y(t) = Cx(t) = C\left(z(t) - Mz_t, z_t\right),$$

where the output operator C is given by

$$C : D(A) \to \mathbb{R}^p, \quad Cx = C(\eta, \phi) = \bar{C}\phi. \tag{4.4}$$

We recall that by assumption $\bar{C}$ is a bounded linear map from $C(-h, 0\,; \mathbb{R}^n)$ to $\mathbb{R}^p$ so that $C \in \mathcal{L}(D(A), \mathbb{R}^p)$ but in general $C \notin \mathcal{L}(M_2, \mathbb{R}^p)$. Using the fact that $S(\cdot)$ restricts to a C_0-semigroup on $D(A)$, a natural choice for W and V is $W = D(A)$ and $V = M_2$, because then $C \in \mathcal{L}(W, \mathbb{R}^p)$ and $B \in \mathcal{L}(\mathbb{R}^m, V)$ and we can choose $U = \mathbb{R}^m$ and $Y = \mathbb{R}^p$. We note that the operator A from $D(A) \subset V$ to V should in fact be denoted by A^V and that the generator of the restriction of $S(\cdot)$ to $W = D(A^V)$ is denoted by A^W (cf. the beginning of §2). In [15] it is explained that now B and C are both admissible in the sense of Definition 2.1, so that the neutral functional differential equation described above can indeed be modeled as a Pritchard–Salamon system $\Sigma(S(\cdot), B, C, 0)$.

Below we shall consider a certain nonstandard LQ-problem for delay systems of the above type, but first we shall discuss the two a priori assumptions of Theorem 3.1, namely (2.6) and (3.3). First of all, we note that condition (2.6) is trivially satisfied because of $W = D(A^V)$. In order to deal with assumption (3.3), we assume in the sequel that $M : C(-h, 0; \mathbb{R}^n) \to \mathbb{R}^n$ has the particular form

$$M\phi = \sum_{j=1}^{\infty} A_{-j}\phi(-h_j) + \int_{-h}^{0} A_{-\infty}(\tau)\phi(\tau)\, d\tau \quad \text{for} \quad \phi \in C(-h, 0; \mathbb{R}^n), \tag{4.5}$$

where $0 < h_j \leq h$, $A_{-j} \in \mathbb{R}^{n \times n}$ for $j \in \mathbb{N}$, $A_{-\infty}(\cdot) \in L_1(-h, 0; \mathbb{R}^{n\times n})$, and $\sum_{j=1}^{\infty} \|A_{-j}\| < \infty$. Furthermore, we assume that

$$\sup\left\{ \mathrm{Re}\lambda : \det\left[I - \sum_{j=1}^{\infty} A_{-j}\exp(-\lambda h_j) \right] = 0 \right\} < 0. \tag{4.6}$$

It is explained in [15] that under the above two assumptions the pair (A^V, B) is exponentially stabilizable on V if and only if

$$\mathrm{rank}\, [\Delta(\lambda), \bar{B}] = n \text{ for all } \lambda \in \mathbb{C} \text{ with } \mathrm{Re}(\lambda) \geq 0, \tag{4.7}$$

where $\Delta(\lambda) = \lambda[I - M(\exp(\lambda \cdot))] - \bar{L}(\exp(\lambda \cdot))$. Now suppose that we have two admissible output operators that have the same form as C:

$$C_1 : D(A^V) \to \mathbb{R}^{p_1}, \quad C_1 x = C_1(\eta, \phi) = \bar{C}_1\phi \tag{4.8}$$

and

$$C_2 : D(A^V) \to \mathbb{R}^{p_2}, \quad C_2 x = C_2(\eta, \phi) = \bar{C}_2\phi, \tag{4.9}$$

where $\bar{C}_1$ and $\bar{C}_1$ are bounded linear functionals from $C(-h, 0; \mathbb{R}^n)$ into $\mathbb{R}^{p_1}$ and $\mathbb{R}^{p_2}$, respectively. We introduce the nonstandard cost function

$$J(x_0, u(\cdot)) := \int_0^{\infty} \mathcal{F}(x(t), u(t))\, dt, \tag{4.10}$$

where

$$\mathcal{F}(x,u) := \langle C_1x, C_1x\rangle_{\mathbb{R}^{p_1}} - \langle C_2x, C_2x\rangle_{\mathbb{R}^{p_2}} + \langle u,u\rangle_{\mathbb{R}^m}. \tag{4.11}$$

The state of the system is given by (4.2), and the cost function can also be written as

$$J(x_0,u(\cdot)) = \int_0^\infty \left(\|y_1(t)\|^2_{\mathbb{R}^{p_1}} - \|y_2(t)\|^2_{\mathbb{R}^{p_2}} + \|u(t)\|^2_{\mathbb{R}^m}\right) dt,$$

where

$$y_1(t) = C_1x(t) \text{ and } y_2(t) = C_2x(t).$$

The following result is an immediate consequence of Theorem 3.1.

THEOREM 4.1. *Let $S(\cdot), B, C_1$, and C_2 be given as above (with M satisfying* (4.5) *and* (4.6)); *and let $J(x_0,u(\cdot))$ be given by* (4.10). *Furthermore, suppose that the stabilizability condition* (4.7) *is satisfied and define U_{adm} according to* (3.4). *Then the following are equivalent.*

1. *For all $x_0 \in V$ there exists a unique $\bar{u}(\cdot) \in L_2(0,\infty;U)$ such that*

$$\inf_{u(\cdot)\in U_{\text{adm}}} J(x_0,u(\cdot)) = \min_{u(\cdot)\in U_{\text{adm}}} J(x_0,u(\cdot)) = J(x_0,\bar{u}(\cdot)). \tag{4.12}$$

2. *There exists a selfadjoint $X \in \mathcal{L}(V)$ such that for all $x,y \in D(A^V)$*

$$\langle A^Vx, Xy\rangle_V + \langle Xx, A^Vy\rangle_V - \langle XBB^*Xx,y\rangle_V$$

$$+ \langle C_1x, C_1y\rangle_{\mathbb{R}^{p_1}} - \langle C_2x, C_2y\rangle_{\mathbb{R}^{p_2}} = 0 \tag{4.13}$$

and $A^V - BB^*X$ is the generator of a C_0-semigroup, exponentially stable on V.

3. *There exists an $\epsilon > 0$ such that for all $(\omega,x,u) \in \mathbb{R}\times D(A^V)\times U$ that satisfy $i\omega x = A^Vx + Bu$ there holds*

$$\|C_1x\|^2_{\mathbb{C}^{p_1}} - \|C_2x\|^2_{\mathbb{C}^{p_2}} + \|u\|^2_{\mathbb{C}^m} \geq \epsilon(\|x\|^2_V + \|u\|^2_{\mathbb{C}^m}). \tag{4.14}$$

Furthermore, if one of these conditions holds, the minimizing $\bar{u}(\cdot)$ in (3.7) *can be given in feedback form*

$$\bar{u}(\cdot) = -B^*Xx(\cdot) \tag{4.15}$$

and

$$\inf_{u(\cdot)\in U_{\text{adm}}} J(x_0,u(\cdot)) = J(x_0,\bar{u}(\cdot)) = \langle x_0, Xx_0\rangle_V. \tag{4.16}$$

Finally, if the X in (3.8) *exists, it is unique.*

Remark 4.2. As explained in Remark 3.3, if

$$\sup_{\omega\in\mathbb{R}} \|(i\omega I - A^V)^{-1}\|_{\mathcal{L}(V)} < \infty, \tag{4.17}$$

then the frequency domain condition of item 3 is equivalent to the condition that for all $\omega \in \mathbb{R}$ and all $u \in \mathbb{C}^m$

$$\mathcal{F}((i\omega I - A^V)^{-1}Bu, u)$$

$$= \|C_1(i\omega I - A^V)^{-1}Bu\|^2_{\mathbb{C}^{p_1}} - \|C_2(i\omega I - A^V)^{-1}Bu\|^2_{\mathbb{C}^{p_2}} + \|u\|^2_{\mathbb{C}^m} \geq \epsilon_1 \|u\|^2_{\mathbb{C}^m}$$

(we note that for our delay system (4.17) will generically hold; in fact, stabilizability of the delay system implies that there are finitely many unstable poles; see [2]). Hence, defining the transfer functions

$$G_1(s) := C_1(sI - A^V)^{-1}B \text{ and } G_2(s) := C_2(sI - A^V)^{-1}B,$$

it follows that the frequency domain condition of item 3 is in this case equivalent to the condition that for all $\omega \in \mathbb{R}$

$$(G_1(i\omega))^*G_1(i\omega) - (G_2(i\omega))^*G_2(i\omega) \geq (\epsilon_1 - 1)I$$

for some $\epsilon_1 > 0$. Hence, the solvability of the LQ-problem depends on a trade-off between the "sizes" of G_1 and G_2 on the imaginary axis.

Finally, we note that if (4.17) holds and $C_2 = 0$, then the frequency domain condition is automatically satisfied (this corresponds to the *standard* LQ-problem, except that here there is no exponential detectability assumption on C_1).

4.2. Parabolic PDEs. Next, we discuss a nonstandard LQ-problem for the class of parabolic PDE systems given in [15].

Let A be a selfadjoint operator on a separable Hilbert space H, and suppose that it has compact resolvent and that its spectrum consists of strictly decreasing real eigenvalues λ_n, $n \in \mathbb{N}$, with eigenvectors $\phi_n \in H$, $\|\phi_n\|_H = 1$. In this case $\{\phi_n, n \in \mathbb{N}\}$ forms an orthonormal basis of H so that for all $x \in H$,

$$\sum_{n=0}^{\infty} \langle x, \phi_n \rangle_H^2 < \infty \quad \text{and} \quad x = \sum_{n=0}^{\infty} \langle x, \phi_n \rangle_H \phi_n.$$

A can be represented as

$$\begin{aligned} D(A) &= \left\{ x \in H \mid \sum_{n=0}^{\infty} \lambda_n^2 \langle x, \phi_n \rangle_H^2 < \infty \right\}, \\ Ax &= \sum_{n=0}^{\infty} \lambda_n \langle x, \phi_n \rangle_H \phi_n, \end{aligned} \tag{4.18}$$

and the C_0-semigroup $S(\cdot)$ generated by A is given by

$$S(t)x = \sum_{n=0}^{\infty} \exp(\lambda_n t) \langle x, \phi_n \rangle_H \phi_n. \tag{4.19}$$

Now let β_n and γ_n be positive sequences satisfying $0 < \beta_n \leq 1 \leq \gamma_n < \infty$ and suppose that W and V are determined by

$$W = \left\{ x \in H \mid \sum_{n=0}^{\infty} \gamma_n \langle x, \phi_n \rangle_H^2 < \infty \right\},$$

$$V' = \left\{ x \in H \mid \sum_{n=0}^{\infty} \beta_n^{-1} \langle x, \phi_n \rangle_H^2 < \infty \right\},$$

with the obvious inner products. Here we assume that H is identified with its dual so that $V' \subset H = H' \subset V$. This means that V can be represented as a space of sequences

$$V = \left\{ x \in \mathbb{R}^{\mathbb{N}} \mid \sum_{n=0}^{\infty} \beta_n x_n^2 < \infty \right\},$$

and the injection $H \subset V$ is given by identifying $x \in H$ with the sequence $\{\langle x, \phi_n \rangle_H \ , n \in \mathbb{N}\}$. Finally, let $B \in \mathcal{L}(\mathbb{R}, V)$ and $C \in \mathcal{L}(W, \mathbb{R})$ be given by

$$Bu = \{b_n u, n \in \mathbb{N}\} \quad \text{and} \quad Cx = \sum_{n=0}^{\infty} c_n \langle x, \phi_n \rangle_H,$$

where the sequences $\{b_n, n \in \mathbb{N}\}$ and $\{c_n, n \in \mathbb{N}\}$ are such that

$$\sum_{n=0}^{\infty} \beta_n b_n^2 < \infty \quad \text{and} \quad \sum_{n=0}^{\infty} \gamma_n^{-1} c_n^2 < \infty. \tag{4.20}$$

It is not difficult to show that B and C are admissible with respect to (W, V) in the sense of Definition 2.1 if

$$\sum_{n=n_0}^{\infty} \frac{\gamma_n b_n^2}{|\lambda_n|} < \infty \quad \text{and} \quad \sum_{n=n_0}^{\infty} \frac{c_n^2}{\beta_n |\lambda_n|} < \infty, \tag{4.21}$$

where $n_0 = 1+\max\{n \in \mathbb{N} \mid \lambda_n \geq 0\}$. Furthermore, given sequences $b_n, c_n, \lambda_n \in \mathbb{R}$ such that λ_n is strictly decreasing and $\lambda_n \downarrow -\infty$, there exist sequences β_n, γ_n such that the inequalities (4.20)–(4.21) are satisfied if and only if

$$\sum_{n=n_0}^{\infty} \frac{b_n c_n}{|\lambda_n|^{1/2}} < \infty \tag{4.22}$$

(see [15, Lemma 4.4]). This last result particularly shows that B and C cannot be "too unbounded" with respect to H. Furthermore, it shows that B can be "more unbounded" as long as C is "less unbounded" and vice versa. We mention that in addition to (2.3)–(2.4), the condition $D(A^V) \hookrightarrow W$ is also satisfied provided that γ_n, β_n are chosen appropriately. Hence, this class of parabolic systems fits into the Pritchard–Salamon framework. A particular example of this class of parabolic systems is given by the PDE model for the temperature distribution of a heated rod with boundary heat control:

$$\begin{cases} \dfrac{\partial z}{\partial t}(\xi, t) = \dfrac{\partial^2 z}{\partial \xi^2}(\xi, t), & t > 0, \quad 0 < \xi < 1; \\[2ex] \dfrac{\partial z}{\partial \xi}(0, t) = u(t), \quad \dfrac{\partial z}{\partial \xi}(1, t) = 0, & t > 0; \\[2ex] y(t) = \int_0^t c(\xi) z(\xi, t)\, d\xi, & t > 0; \\[2ex] z(\xi, 0) = z_0(\xi), & 0 < \xi < 1 \text{ (intitial condition)}, \end{cases} \tag{4.23}$$

where $c(\cdot) \in L_2(0,1)$ (the output $y(t)$ is some averaged measurement of the temperature). Formally, the PDE may be expressed as

$$\frac{\partial z}{\partial t}(\xi, t) = \frac{\partial^2 z}{\partial \xi^2}(\xi, t) - \delta_0 u(t),$$

where δ_0 denotes the Dirac delta impulse at $\xi = 0$ (actually, this can be done in a rigorous way, using distribution theory; see Salamon [20]). Choosing the state-space $H = L_2(0,1)$ with the state function

$$x(t) = z(t,\xi), \qquad t > 0, \quad 0 < \xi < 1,$$

the PDE can be modeled as a system of the form (2.5) with $D = 0$, i.e.,

$$\begin{cases} x(t) &=& S(t)x_0 + \int_0^t S(t-s)Bu(s)\,ds, \\ y(t) &=& Cx(t). \end{cases}$$

Here $S(\cdot)$ is a C_0-semigroup on H and its infinitesimal generator A, the input operator B, and the output operator C are defined by

$$D(A) = \left\{ x \in H \mid x, \frac{dx}{d\xi} \text{ are absolutely continuous }, \right.$$
$$\left. \frac{d^2x}{d\xi^2} \in H, \text{ and } \frac{dx}{d\xi}(0) = \frac{dx}{d\xi}(1) = 0 \right\},$$

$$Ax = \frac{d^2x}{d\xi^2},$$

$$Bu = -\delta_0 u,$$

$$Cx = \langle c(\cdot), x(\cdot)\rangle_H.$$

It follows that A is indeed selfadjoint, that it has compact resolvent, and that it is of the form (4.18), where $\lambda_0 = 0$, $\phi_0 = 1$ and $\lambda_n = -n^2\pi^2$, $\phi_n(\xi) = \sqrt{2}\cos(n\pi\xi)$ for $n \geq 1$. Furthermore, we get $c_n = \langle c(\cdot), \phi_n\rangle_H, n \in \mathbb{N}$, and $b_0 = -1$, $b_n = -\sqrt{2}$ for $n \geq 1$ and condition (4.22) is satisfied if and only if

$$\sum_{n=1}^{\infty} \frac{|c_n|}{n} < \infty. \tag{4.24}$$

Since $c(\cdot) \in H$, we have $\sum_{n=1}^{\infty} c_n^2 < \infty$ and so condition (4.24) is satisfied. In fact, we can choose the sequences γ_n and β_n by $\gamma_n = 1$ and $\beta_n = n^{-2}$, $n \geq 1$. This corresponds to the choice $W = H$ and $V' = W^{1,2}(0,1)$ (the Sobolev space of absolute continuous functions in $L_2(0,1)$ whose derivative is in $L_2(0,1)$).

Now it is of course possible to set up a nonstandard LQ-problem for the above described abstract parabolic systems. The system is given by

$$x(t) = S(t)x_0 + \int_0^t S(t-s)Bu(s)\,ds,$$

and one can choose output operators C_1, C_2, and L with the same admissiblity properties as the output operator C above and construct a cost function of the form

$$J(x_0, u(\cdot)) = \int_0^\infty ((C_1x(t))^2 - (C_2x(t))^2 + Lx(t)u(t) + r(u(t))^2)\,dt.$$

If $b_n \neq 0$ for $n = 1, \dots, n_0 - 1$, the pair (A^V, B) is exponentially stabilizable (see [15]) and so in this case Theorem 3.1 is applicable. In order to make this very general result a bit more explicit we consider the special case of the PDE in (4.23) and we choose the output operators $C_1, C_2 \in \mathcal{L}(W, \mathbb{R})$ as follows:

$$\begin{aligned} C_1x &= K\langle\phi_0(\cdot), x(\cdot)\rangle_H \text{ for some } K \in \mathbb{R}, \\ C_2x &= \langle c(\cdot), x(\cdot)\rangle_H \text{ for some } c(\cdot) \in H \end{aligned}$$

(we take $L = 0$). It follows from the above that with $\gamma_n = 1$ for $n \geq 0$, $\beta_0 = 1$, and $\beta_n = 1/n^2$ for $n \geq 1$ we have a stabilizable Pritchard–Salamon system that satisfies assumption (2.6) and C_1 and C_2 are admissible output operators (recall that $\lambda_0 = 0$, $\phi_0 = 1$; $\lambda_n = -n^2\pi^2$, $\phi_n(\xi) = \sqrt{2}\cos(n\pi\xi)$ for $n \geq 1$; $b_0 = -1$, $b_n = -\sqrt{2}$ for $n \geq 1$, and $c_n = \langle c(\cdot), \phi_n\rangle_H$ for $n \geq 0$). The nonstandard LQ-cost function is given by

$$J(x_0, u(\cdot)) = \int_0^\infty ((C_1x(t))^2 - (C_2x(t))^2 + r(u(t))^2)\,dt.$$

Now let us consider the frequency domain condition of Theorem 3.1 for the problem:

- There exists an $\epsilon > 0$ such that for all $(\omega, x, u) \in \mathbb{R} \times D(A^V) \times \mathbb{C}$ that satisfy $i\omega x = A^V x + Bu$ there holds

$$|C_1x|^2 - |C_2x|^2 + r|u|^2 \geq \epsilon(\|x\|_V^2 + |u|^2). \tag{4.25}$$

The expression $i\omega x = A^V x + Bu$ can be reformulated componentwise as

$$i\omega x_n = (A^V x)_n + (Bu)_n = \lambda_n x_n + b_n u \text{ for all } n \geq 0,$$

and due to the fact that $\lambda_0 = 0$ we need to consider two cases:

$$\omega = 0 \Rightarrow \begin{cases} x_0 \text{ is arbitrary, } x_n = 0 \text{ for } n \geq 1,\ u = 0; \\ \|x\|_V^2 = |x_0|^2; \\ C_1x = Kx_0; \\ C_2x = c_0x_0. \end{cases}$$

and

$$\omega \neq 0 \Rightarrow \begin{cases} x_n = \dfrac{b_n u}{i\omega - \lambda_n} \text{ for } n \geq 0, u \text{ arbitrary}, \\ \|x\|_V^2 = \dfrac{|u|^2}{\omega^2} + \displaystyle\sum_{n=1}^\infty \dfrac{2|u|^2}{n^2(\omega^2 + \lambda_n^2)}, \\ C_1x = Kx_0 = \dfrac{Ku}{i\omega}, \\ C_2x = \langle c(\cdot), x(\cdot)\rangle = \displaystyle\sum_{n=0}^\infty \dfrac{c_n b_n u}{i\omega - \lambda_n} \text{ for } n \geq 0. \end{cases}$$

It follows that a necessary condition for the frequency domain inequality to hold is that $|K| > |c_0|$ (due to the case $\omega = 0$). Sufficient conditions on c_n for the frequency domain inequality to hold may now be obtained by estimating $|C_2 x|$.

Hence, just as in the delay example, Theorem 3.1 is applicable and the frequency domain inequality is the condition that is most easy to verify.

4.3. A beam example. Finally, we discuss the applicability of Theorem 3.1 to a flexible beam example. In [2] an example is given of an Euler–Bernoulli beam with Kelvin–Voigt damping that can be modeled as a Pritchard–Salamon system (it is in fact a simple model for a satellite). The corresponding PDE is given by

$$\begin{cases} \dfrac{\partial^2 w}{\partial t^2}(\xi,t) + \alpha_1 A \dfrac{\partial w}{\partial t}(\xi,t) + \alpha_2 A w(\xi,t) = \dfrac{1}{\rho a}(\delta_0 u_1(t) - \delta_0' u_2(t)),\ t \geq 0, -1 < \xi < 1; \\ \dfrac{\partial^2 w}{\partial \xi^2}(-1,t) = 0 = \dfrac{\partial^2 w}{\partial \xi^2}(1,t) \text{ for } t \geq 0; \\ \dfrac{\partial^3 w}{\partial \xi^3}(-1,t) = 0 = \dfrac{\partial^3 w}{\partial \xi^3}(1,t) \text{ for } t \geq 0; \\ y_1(t) = w(0,t),\ y_2(t) = \dfrac{\partial w}{\partial \xi}(0,t) \text{ for } t \geq 0, \end{cases}$$

where α_1 and α_2 are the damping and stiffness coefficients, ρa is the mass per unit length, and A is the selfadjoint positive operator from $D(A) \subset L_2(-1,1)$ to $L_2(-1,1)$ given by

$$\begin{aligned} D(A) &= \{h \in L_2(-1,1) \mid h', h'', h''', h'''' \in L_2(-1,1) \\ &\qquad h''(-1) = h''(1) = 0,\ h'''(-1) = h'''(1) = 0\}, \\ Ah &= \frac{d^4 h}{d\xi^4}. \end{aligned}$$

Choosing the state $x(t) = (w(\cdot,t), \frac{\partial w}{\partial t}(\cdot,t)) \in D(A^{1/2}) \times L_2(-1,1)$, this PDE can be reformulated as a Pritchard-Salamon system of the form (2.5). For the details we refer to [2]; important elements are the spectral analysis of the system operator $\left(\begin{smallmatrix} 0 & I \\ -\alpha_2 A & -\alpha_1 A \end{smallmatrix}\right)$ and the use of scaled Hilbert spaces, just as in the parabolic PDE example. Furthermore, it is straightforward to formulate nonstandard LQ-problems for this example by choosing appropriate admissible output operators C_1, C_2, and L just as in the delay example and the parabolic PDE example, so we need not pursue this issue any further.

Acknowledgment. I thank the authors of [3] and Martin Weiss for some interesting and helpful discussions.

REFERENCES

[1] A. V. BALAKRISHNAN, *Applied Functional Analysis*, Springer, New York, 1976.

[2] R. F. CURTAIN, *A synthesis of time and frequency domain methods for the control of infinite-dimensional systems: a system theoretic approach*, Control and estimation in distributed parameter systems, Frontiers in Applied Mathematics, H. T. Banks, ed., Society for Industrial and Applied Mathematics, Philadelphia, 1992, pp. 171–224.

[3] R. F. Curtain, H. Logemann, S. Townley, and H. Zwart, *Well-posedness, stabilizablity and admissibility for Pritchard-Salamon systems*, J. Math. Systems Estim. Control, 4 (1994), pp. 493–496.

[4] R. F. Curtain and A. J. Pritchard, *Infinite-Dimensional Linear Systems Theory*, Lecture Notes in Control and Inform. Sci. 8, Springer-Verlag, Berlin, 1978.

[5] R. F. Curtain and H. Zwart, *An introduction to infinite-dimensional linear systems theory*, manuscript.

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

[7] Y.Fourés and I. E. Segal, *Causality and analyticity*, Trans. Amer. Math. Soc., 78 (1955) pp. 385–405.

[8] B. A. M. van Keulen, M. Peters, and R. F. Curtain, *H_∞-control with state-feedback: the infinite-dimensional case*, J. Math. Systems Estimation Control, 3 (1993), pp. 1–39.

[9] B. A. M. van Keulen, *H_∞-Control for Distributed Parameter Systems: A State-Space Approach*, Systems Control Found. Appl., Birkhäuser, Boston, 1993.

[10] I. Lasiecka and R. Triggiani, *Differential and algebraic Riccati equations with applications to boundary/point control problems: Continuous theory and approximation theory*, Lecture Notes in Control Inform. Sci. 164, Springer-Verlag, Berlin, 1991.

[11] J. L. Lions, *Optimal control of systems governed by partial differential equations*, Springer, Berlin, 1971.

[12] J. Louis, *The regulator problem in Hilbert spaces and some applications to stability of nonlinear control systems*, Ph.D. thesis, Facultés Universitaires Notre-Dame de la Paix, Namur, Belgium, 1986.

[13] J. Louis and D. Wexler, *The Hilbert space regulator problem and operator Riccati equation under stabilizability*, Ann. Soc. Sci. Bruxelles Sér. I, 105 (1991), pp. 137–165.

[14] A. Pazy, *Semigroups of linear operators and applications to partial differential equations*, Springer, New York, 1983.

[15] A. J. Pritchard and D. Salamon, *The linear quadratic control problem for infinite-dimensional systems with unbounded input and output operators*, SIAM J. Control Optim., 25 (1987), pp. 121–144.

[16] ———, *The linear quadratic control problem for retarded systems with delays in control and observation*, IMA J. Control Inform., 2 (1985), pp. 335–362.

[17] A. J. Pritchard and S. Townley, *Robustness optimization for abstract, uncertain control systems: unbounded inputs and perturbations*, in Proc. of IFAC Symposium on Distributed Parameter Systems, M. Amouroux and A. El Jai, eds., Pergamon Press, New York, 1990, pp. 117–121.

[18] M. Rosenblum and R. Rovnyak, *Hardy classes and operator theory*, Oxford University Press, London, 1985.

[19] D. Salamon, *Control and observation of neutral systems*, Vol. 91, Pitman, London, 1984.

[20] D. Salamon, *Infinite-dimensional systems with unbounded control and observation: a functional analytic approach*, Trans of the Amer. Math. Soc., 300 (1987), pp. 383–431.

[21] D. Wexler, *On frequency domain stability for evolution equations in Hilbert spaces via the algebraic Riccati equation*, SIAM J. Math Anal., 11 (1980), pp. 969–983.

[22] J. C. Willems, *Least squares stationary optimal control and the algebraic Riccati equation*, IEEE Trans. Automat. Control, 16 (1971), pp. 621–634.

[23] G. Weiss, *Admissible observation operators for linear semigroups*, Israel J. Math., 65 (1989), pp. 17–43.

[24] ———, *Transfer functions of regular systems, Part* 1: *Characterizations of regularity*, submitted for publication.

[25] V. A. Yakubovich, *A frequency theorem for the case in which the state and control spaces are Hilbert spaces with an application to some problems in the synthesis of optimal controls* I, Siberian Math. J., 15 (1974), pp. 457–476.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 33, No. 5, pp. 1357 1391, September 1995

© 1995 Society for Industrial and Applied Mathematics

004

EXACT CONTROLLABILITY AND STABILIZATION OF A VIBRATING STRING WITH AN INTERIOR POINT MASS*

SCOTT HANSEN[†] AND ENRIQUE ZUAZUA[‡]

Abstract. In this article we examine the problems of boundary control and stabilization for a one-dimensional wave equation with interior point masses. We show that singularities in waves are "smoothed one order" as they cross a point mass. Thus in the case of one interior point mass, with, e.g., L^2-Dirichlet control at the left end, the most general reachable space (from 0) that one can expect is $L^2 \times H^{-1}$ to the left of the mass and $H^1 \times L^2$ to the right of the mass. We show that this is in fact the optimal result (modulo certain compatibility conditions). Several related results for both control and stabilization of such systems are also given.

Key words. boundary control, hyperbolic system, hybrid system, vibrating network

AMS subject classifications. 35P10, 35P20, 35L20, 73K03, 93C20

1. Introduction and main results.

In recent years there has been much interest in the topic of control and stabilization of so called "hybrid systems" in which the dynamics of elastic systems and possibly rigid structures are related through some form of coupling. For example, see [1] for serially connected beams, [8]–[10] for beams with end masses, and [4], [12], [14] for networks of strings and beams.

In this article we examine a simple model for an elastic string involving an interior point mass. We obtain a precise description of the space of exact controllability when control is active at one or both ends of the string and also describe the best possible stabilization results via velocity feedback at one or both ends. We refer to [14] for a discussion on the modeling and well-posedness of networks of strings containing, in particular, point masses. Approximate controllability results for these networks were announced in [13].

It will be convenient to regard the string-mass system as two separate strings in which one end of each string is attached to a common point mass. Thus assume the first string occupies $\Omega_1 = (-\ell_1, 0) \subset \mathbb{R}$ and the second one $\Omega_2 = (0, \ell_2) \subset \mathbb{R}$, where ℓ_1 and ℓ_2 are positive.

For simplicity of the exposition we suppose both strings to be homogeneous. The deformations of the first and second string will be described respectively by the functions

$$\begin{aligned} u &= u(x,t), \quad x \in \Omega_1, \quad t > 0, \\ v &= v(x,t), \quad x \in \Omega_2, \quad t > 0. \end{aligned}$$

The position of the mass (which is attached to the strings at the point $x = 0$) is described by the function $z = z(t)$ for $t > 0$.

[*] Received by the editors May 3, 1993; accepted for publication (in revised form) February 22, 1994. This research was supported in part by the Institute for Mathematics and its Applications with funds provided by the National Science Foundation. The authors are grateful to the IMA for the hospitality and support.

[†] Department of Mathematics, Iowa State University, Ames, Iowa 50011.

[‡] Departamento de Matemática Aplicada, Universidad Complutense de Madrid, 28040 Madrid, Spain. The research of this author was partially supported by Dirección General de Investigación Cientifica y Technológica Project PB90-0245 and EEC grant SC1-CT91-0732.

To fix ideas we suppose that the strings satisfy Dirichlet boundary conditions at the end points ($x = -\ell_1, \ell_2$). Then the equations modeling the dynamics of this system in the absence of controls are as follows:

$$\left\{\begin{array}{ll} \rho_1 u_{tt} = \sigma_1 u_{xx}, & x \in \Omega_1,\ t > 0, \\ \rho_2 v_{tt} = \sigma_2 v_{xx}, & x \in \Omega_2,\ t > 0, \\ M z_{tt}(t) + \sigma_1 u_x(0,t) - \sigma_2 v_x(0,t) = 0, & t > 0, \\ u(-\ell_1, t) = v(\ell_2, t) = 0, & t > 0, \\ u(0,t) = v(0,t) = z(t), & t > 0. \end{array}\right. \tag{1.1}$$

The constants $\rho_1 > 0$ and $\rho_2 > 0$ represent the density of each string and $M > 0$ represents the mass of the point mass. The tensions in each string are assumed positive and denoted by σ_1 and σ_2. If the only forces acting on the point mass are those of the strings then $\sigma_1 = \sigma_2$; however, if an external force is present (for example, gravity acting along the x-axis) then the two tensions will be different.

Note that when $M = 0$, system (1.1) describes the motion of a string with a piecewise constant wave speed.

Of course, in order to determine the solution of (1.1) in a unique way we have to add some initial conditions at time $t = 0$ that will be represented by

$$\left\{\begin{array}{ll} u(x,0) = u^0(x), \quad u_t(x,0) = u^1(x), & x \in \Omega_1, \\ v(x,0) = v^0(x), \quad v_t(x,0) = v^1(x), & x \in \Omega_2, \\ z(0) = z^0, \quad z_t(0) = z^1. & \end{array}\right. \tag{1.2}$$

As usual, depending on the regularity properties and the compatibility conditions these initial data satisfy, we may expect a different degree of regularity of solutions.

Let us introduce the energy

$$\begin{aligned} E_M(t) = \frac{1}{2} \int_{-\ell_1}^{0} [\rho_1 |u_t(x,t)|^2 + \sigma_1 |u_x(x,t)|^2]\, dx + \frac{M}{2} |z_t(t)|^2 \\ + \frac{1}{2} \int_0^{\ell_2} [\rho_2 |v_t(x,t)|^2 + \sigma_2 |v_x(x,t)|^2] dx. \end{aligned} \tag{1.3}$$

In the absence of controls, i.e., for solutions of (1.1), this energy is constant in time.

We are interested in the controllability properties of this system when control is active at one or both of the end points of the string-mass system. We will discuss these control problems from two different point of views. The first consists of finding suitable observability estimates and then applying Hilbert's uniqueness method (HUM) (cf. J. L. Lions [6], [7]) and the other is based on the use of nonharmonic Fourier series and moment problems (cf. D. L. Russell [11]).

We will see that the presence of the point mass introduces some important changes in the behavior of the system with respect to the observability properties.

Let us recall what one has concerning the observability of (1.1) in the absence of the mass (i.e., when $M = 0$):

(a) If $T > \ell_1 \sqrt{\rho_1/\sigma_1} + \ell_2 \sqrt{\rho_2/\sigma_2}$ there exists $C(T) > 0$ such that

$$E_0(0) \le C \int_0^T [\,|u_x(-\ell_1, t)|^2 + |v_x(\ell_2, t)|^2] dt. \tag{1.4}$$

(b) If $T > 2(\ell_1\sqrt{\rho_1/\sigma_1} + \ell_2\sqrt{\rho_2/\sigma_2})$ there exists $C(T) > 0$ such that

$$E_0(0) \leq C \int_0^T |v_x(\ell_2, t)|^2 dt. \tag{1.5}$$

We will see that an estimate of the form (1.4) holds[1] for any $M > 0$ but only if $T > 2\max(\ell_1\sqrt{\rho_1/\sigma_1}, \ell_2\sqrt{\rho_2/\sigma_2})$. Thus when controlling at both extremes $x = -\ell_1, \ell_2$ we will get the analogue of the results one can prove (see Theorem 3.1) for two serially connected strings without the point mass but only for $T > 2\max(\ell_1\sqrt{\rho_1/\sigma_1}, \ell_2\sqrt{\rho_2/\sigma_2})$.

However, we will see that the observability inequality (1.5) does not hold when $M > 0$, no matter how large T is. By an explicit computation (see Proposition 2.5) one sees that when a wave starting from initial data

$$\begin{cases} u^0 = \varphi^0 \in H_0^1(\Omega_1), & v^0 = 0, \\ u^1 = \varphi^1 \in L^2(\Omega_1), & v^1 = 0, \\ z^0 = z^1 = 0 & \end{cases}$$

crosses the point mass, part of the wave is reflected off the point mass and part is transmitted. The part which is reflected keeps the same regularity as the initial data, but the part that crosses the mass is regularized by one degree (i.e., $v(\cdot, t) \in H^2(\Omega_2)$ for all $t > 0$). Of course this is due to the presence of the mass and does not occur when $M = 0$.

This phenomena explains why, if we want to observe the initial energy of the solution (in this case the H^1-norm of φ), we need an estimate on $v_{xt}(\ell_2, t)$ in $L^2(0, T)$ and not only on $v_x(\ell_2, t)$.

When $M > 0$ and $T > 2(\ell_1\sqrt{\rho_1/\sigma_1} + \ell_2\sqrt{\rho_2/\sigma_2})$ we are able to prove that

$$E_M(0) + ||v_{xx}(\cdot, 0)||^2_{L^2(\Omega_1)} + ||v_{xt}(\cdot, 0)||^2_{L^2(\Omega_2)} \leq C \int_0^T |v_{xt}(\ell_2, t)|^2 dt \tag{1.6}$$

and this inequality is sharp in the sense that the reverse one holds for all $T > 0$.

As a consequence of (1.6) we deduce that when controlling from only the end $x = \ell_2$ with $L^2(0, T)$ controls, the controllability is achieved in a space smaller than the usual one (the one we have for $M = 0$) since the components corresponding to the first string are of one more degree of regularity. The well-posedness of the system (1.1) in this asymmetric space is due to the presence of the point mass and does not hold when $M = 0$.

It is also interesting to understand this phenomena from the point of view of nonharmonic Fourier series. Let us consider, for simplicity, the case $\rho_1 = \rho_2 = \ell_1 = \ell_2 = \sigma_1 = \sigma_2 = 1$. By a careful study of the spectrum of the elliptic operator involved in system (1.1) we can see that $v_x(\ell_2, t)$ is given by the nonharmonic Fourier series

$$v_x(1, t) = a_0 e^{-i\omega_0 t} + b_0 e^{i\omega_0 t} + \sum_{k \in \mathbb{Z}\setminus\{0\}} (a_k e^{ik\pi t} + b_k e^{i\omega_k t}), \tag{1.7}$$

[1] That an estimate of this type can be obtained was mentioned by Schmidt [13]; however, the optimality of T is much more subtle and was unknown at that time.

where a_k and b_k are complex numbers related to the Fourier coefficients of the initial data and $(i\omega_k)$ is a sequence of eigenvalues such that

$$\left\{\begin{array}{l} |\omega_k - k\pi| \to 0 \text{ as } |k| \to \infty, \\ \omega_k \neq k\pi, \quad \forall k \in \mathbb{Z}. \end{array}\right. \tag{1.8}$$

When $M = 0$ this second sequence $(\omega_k) \cup -\omega_0$ of eigenvalues becomes $(k\pi + \frac{\pi}{2})_{k\in\mathbb{Z}}$ and $v_x(1,t)$ is given by a (harmonic) Fourier series involving the exponentials $(e^{ik\pi t/2})_{k\in\mathbb{Z}}$, which are orthogonal on $L^2(0,4)$.

In terms of (1.7) the inequality (1.5) (which we already mentioned does not hold when $M > 0$) would be equivalent to

$$\sum_{k=-\infty}^{\infty} [\,|a_k|^2 + |b_k|^2] \le C \int_0^T |v_x(1,t)|^2 dt, \tag{1.9}$$

for $T > 4$. However, all results concerning inequality (1.9) for nonharmonic Fourier series existing in the literature require an asymptotic spectral gap (cf. [2], [3], and [16]) that, in view of (1.8), does not hold in our case. Instead of (1.9), employing a result of D. Ullrich [15], we get the following weaker version of (1.9):

$$\sum_{k=-\infty}^{\infty} [\,|a_k + b_k|^2 + (w_k - k\pi)^2 |a_k - b_k|^2] \le C \int_0^T |v_x(1,t)|^2 dt, \tag{1.10}$$

which holds for $T \ge 4$, and is just the Fourier version of the observability inequality (1.6).

When controlling at both extremes $x = -\ell_1, \ell_2$ but with

$$\ell_1 \sqrt{\frac{\rho_1}{\sigma_1}} + \ell_2 \sqrt{\frac{\rho_2}{\sigma_2}} < T < 2\max\left(\ell_1 \sqrt{\frac{\rho_1}{\sigma_1}}, \ell_2 \sqrt{\frac{\rho_2}{\sigma_2}}\right) \tag{1.11}$$

(which is only possible if $\ell_1\sqrt{\rho_1/\sigma_1} \neq \ell_2\sqrt{\rho_2/\sigma_2}$), we again obtain, as above, an asymmetric controllability space since the solution components corresponding the string with the longest propagation time has one more degree of regularity over a portion of that string. Thus we also obtain controls with different regularities, namely, the control on the side of the mass with the smoother solution belongs to $H_0^1(0,T)$ and the control at the other end belongs to $L^2(0,T)$. Therefore, even if we control at both ends when T satisfies (1.11), this phenomena in which the controllability space is asymmetric appears.

We will also briefly discuss the stabilization problem concernings two different situations. First we prove that by introducing boundary damping at both extremes $x = -\ell_1, \ell_2$ the energy of solutions decays exponentially uniformly. Moreover trajectories converge exponentially to a constant equilibrium that can be determined in terms of the initial data. We then consider the case where the boundary damping acts only at one extreme point. In this case the energy of every solution converges to zero but there is not a uniform exponential decay since a sequence of eigenvalues of the system approaches the imaginary axis. This phenomena is similar to that founded by E. B. Lee and Y. C. You [5] and W. Littman and L. Markus [10] when studying the stability properties of strings and beams damped at one extreme through a point mass.

This paper will be devoted to the particular case of two homogeneous strings with a point mass, but the techniques and ideas involved are rather general and may be used to discuss nonhomogeneous strings, other boundary conditions at the extremes, other dynamics at the point mass, and other situations where more than two strings and one mass are present. A more detailed discussion of some of these extensions will be given in the last section.

The rest of the paper is organized as follows. In §2 we give some preliminary results concerning the existence, regularity, and uniqueness of solutions of both the uncontrolled and the controlled system. In §3 we examine the controllability problem for the case where the control acts at both extremes. In §4, using energy methods we prove estimate (1.7) and its consequence concerning controllability. In §5 we carefully analyze the problem of controlling from only one extreme by means of moment problems and nonharmonic Fourier series. In particular we discuss and prove inequality (1.10). In §6 we discuss the boundary-stabilization problem. Finally, in §7, we give some extensions of our main results.

Throughout this paper C will denote a positive constant that may vary from line to line. We will make explicit the dependence of these constants with respect to the various parameters of the problem only when this becomes necessary.

2. Existence, uniqueness and regularity of solutions. In this section we give some preliminary results concerning the existence, uniqueness, and regularity of solutions. First we consider the system (1.1) without controls and then the case of nonhomogeneous boundary conditions.

2.1. Homogeneous boundary conditions. Let us introduce the vector spaces

$$\begin{aligned}\vartheta_1 &= \{\varphi \in H^1(\Omega_1) : \varphi(-\ell_1) = 0\},\\ \vartheta_2 &= \{\psi \in H^1(\Omega_2) : \psi(\ell_2) = 0\},\\ \vartheta &= \{(\varphi, \psi) \in \vartheta_1 \times \vartheta_2 : \varphi(0) = \psi(0)\}\end{aligned}$$

endowed with the norms

$$\begin{aligned}||\varphi||^2_{\vartheta_i} &= \int\limits_{\Omega_i} |\varphi_x(x)|^2\,dx, \qquad i = 1, 2,\\ ||(\varphi, \psi)||^2_{\vartheta} &= ||\varphi||^2_{\vartheta_1} + ||\psi||^2_{\vartheta_2}.\end{aligned}$$

Note that the space ϑ is algebraically and topologically equivalent to $H^1_0(-\ell_1, \ell_2)$. However, since we are considering a system made of two different strings it is convenient to think of ϑ as a subspace of $\vartheta_1 \times \vartheta_2$.

Let us also consider the following closed subspace of $\vartheta \times \mathbb{R}$:

$$W_1 = \{(\varphi, \psi, z) \in \vartheta \times \mathbb{R} : \varphi(0) = \psi(0) = z\},$$

which is densely and continuously embedded in the space

$$W_0 = L^2(\Omega_1) \times L^2(\Omega_2) \times \mathbb{R}.$$

Define the Hilbert space $\mathcal{H}$ by

$$\mathcal{H} = W_1 \times W_0$$

with the product topology. In terms of the vector-valued function

$$y = (u, v, z, \dot{u}, \dot{v}, \dot{z})^t$$

(the superscript ^{t} denotes transposition), we may define an unbounded operator $\mathcal{A}$ on $\mathcal{H}$ by

$$\mathcal{A}y = \begin{pmatrix} 0 & I \\ A & 0 \end{pmatrix} y; \quad A = \begin{pmatrix} \frac{\sigma_1}{\rho_1} d^2 & 0 & 0 \\ 0 & \frac{\sigma_2}{\rho_2} d^2 & 0 \\ \frac{-\sigma_1}{M} d\delta_0 & \frac{\sigma_2}{M} d\delta_0 & 0 \end{pmatrix},$$

where d denotes the (distributional) derivative operator and δ_0 denotes the Dirac delta function with mass at $x = 0$. The domain of $\mathcal{A}$ is given by

$$D(\mathcal{A}) = \{ y \in \mathcal{H} : u \in H^2(\Omega_1), v \in H^2(\Omega_2), (\dot{u}, \dot{v}, \dot{z}) \in W_1 \}.$$

When $\mathcal{D}(\mathcal{A})$ is endowed with the graph-norm topology

$$||y||_{\mathcal{D}(\mathcal{A})} = \{||y||^2_{\mathcal{H}} + ||\mathcal{A}y||^2_{\mathcal{H}}\}^{1/2}$$

it becomes a Hilbert space with dense and continuous embedding in $\mathcal{H}$.

System (1.1)–(1.2) can be written as

$$\text{(2.1)} \qquad \frac{dy}{dt} = \mathcal{A}y, \qquad y(0) = y^0 = (u^0, v^0, z^0, u^1, v^1, z^1)^t.$$

It is easy to see that $\mathcal{A}$ is skew adjoint and m-dissipative on $\mathcal{H}$ and therefore generates a strongly continuous group of isometries on $\mathcal{H}$. Therefore we have the following existence and uniqueness result for (1.1)–(1.2).

PROPOSITION 2.1. (i) *For every* $y^0 = (u^0, v^0, z^0, u^1, v^1, z^1)^t \in \mathcal{H}$ *there exists a unique solution of* (1.1)–(1.2) *in the class*

$$\text{(2.2)} \qquad (u, v, z) \in C([0,T]; W_1) \cap C^1([0,T]; W_0).$$

Furthermore, the energy E_M *remains constant along this solution trajectory.*

(ii) *If* $y^0 \in \mathcal{D}(\mathcal{A})$ *then the corresponding solution has the following additional regularity:*

$$\text{(2.3)} \qquad \begin{cases} u \in C([0,T]; H^2(\Omega_1) \cap C^1([0,T]; H^1(\Omega_1)), \\ v \in C([0,T]; H^2(\Omega_2) \cap C^1([0,T]; H^1(\Omega_2)). \end{cases}$$

Let us denote those solutions satisfying (2.2) by finite-energy solutions.

We can also prove the following regularity result for finite-energy solutions.

PROPOSITION 2.2. *For every* $T > 0$ *there exists some constant* $C(T) > 0$ *such that the following inequality holds for every finite-energy solution:*

$$\text{(2.4)} \qquad \int_0^T [\,|u_x(-\ell_1, t)|^2 + |v_x(\ell_2, t)|^2]dt \leq CE_M(0).$$

Proof. It is well known by now (cf. [7]) that this estimate is of local nature. Therefore it does not depend on whether there is a point mass on the string. However, we have to use the conservation of the energy $E_M(t)$ in (1.3) to obtain the upper bound in terms of the initial energy. $\square$

It is also convenient to consider system (1.1) in the presence of some external distributed force:

$$
\begin{cases}
\rho_1 u_{tt} = \sigma_1 u_{xx} + f(x,t), & x \in \Omega_1,\ 0 < t < T, \\
\rho_2 v_{tt} = \sigma_2 v_{xx} + g(x,t), & x \in \Omega_2,\ 0 < t < T, \\
M z_{tt}(t) + \sigma_1 u_x(0,t) - \sigma_2 v_x(0,t) = h(t), & 0 < t < T, \\
u(0,t) = v(0,t) = z(t), & 0 < t < T, \\
u(-\ell_1, t) = v(\ell_2, t) = 0, & 0 < t < T.
\end{cases}
\tag{2.5}
$$

By standard semigroup methods we have the following proposition.

PROPOSITION 2.3. *For every $y^0 \in \mathcal{H}$ and $(f,g,h) \in L^1(0,T;W_0)$ there exists a unique finite-energy solution of (2.5), (1.2) in the class (2.2). Moreover, there exists a constant $C > 0$ such that*

$$
\int_0^T \left[|u_x(-\ell_1,t)|^2 + |v_x(\ell_2,t)|^2 \right] dt \le C \Big[||y^0||^2_{\mathcal{H}} + ||f||^2_{L^1(0,T;L^2(\Omega_1))} + ||g||^2_{L^1(0,T;L^2(\Omega_2))} + ||h||^2_{L^1(0,T)} \Big].
\tag{2.6}
$$

We will also need the following result.

PROPOSITION 2.4. *Suppose that $y^0 = 0$ and*

$$
f = \frac{\partial F}{\partial t}, \quad g = \frac{\partial G}{\partial t}, \quad h = \frac{dH}{dt},
$$

where $(F,G,H) \in L^1(0,T;W_1)$. Then the solution (u,v,z) of (2.5), (1.2) is such that

$$
(u,v,z) = (U_t, V_t, Z_t)
$$

with

$$
(U,V,Z,U_t,V_t,Z_t)^t \in C([0,T];D(\mathcal{A}))
$$

and therefore, in particular,

$$
\begin{cases}
(u,v,z) \in C([0,T];W_1), \\
(u_t,v_t,z_t) \in C([0,T];W_0) + L^1(0,T;W_1), \\
(U,V) \in C([0,T];H^2(\Omega_1) \times H^2(\Omega_2)).
\end{cases}
\tag{2.7}
$$

Moreover, there exists a constant $C > 0$ such that

$$
\int_0^T [\,|u_x(-\ell_1,t)|^2 + |v_x(\ell_2,t)|^2] dt \le C||(F,G,H)||^2_{L^1(0,T;W_1)}.
\tag{2.8}
$$

Proof. We have $(u,v,z) = (U_t,V_t,Z_t)$ where (U,V,Z) is the solution of (2.5), (1.2) with zero initial data and $(f,g,h) = (F,G,H)$. We have

$$
(U,V,Z,U_t,V_t,Z_t)^t \in C([0,T];D(\mathcal{A})),
$$

and therefore (2.7) holds.

The regularity property (2.8) is more subtle and can be proved proceeding as in [7, Chap. I, Thm. 4.2, p. 46]. □

Let us now study the regularity of solutions where the initial data belong to a space where the regularity is not the same in each of the strings. More precisely, let us consider initial data

$$
y^0 \in \mathcal{H}
\tag{2.9}
$$

such that

$$u^0 \in H^2(\Omega_1), \quad u^1 \in \vartheta_1, \quad u^1(0) = z^1. \tag{2.10}$$

Of course, (2.9)–(2.10) do not imply that $y^0 \in D(\mathcal{A})$ and therefore we cannot apply the regularity Proposition 2.1(ii) provides. However, we can prove the following result.

PROPOSITION 2.5. *Suppose that the initial data y^0 satisfies* (2.9)–(2.10). *Then the solution of* (1.1)–(1.2) *is such that, in addition to* (2.2), *we have*

$$u \in C([0,T]; H^2(\Omega_1)) \cap C^1([0,T]; \vartheta_1). \tag{2.11}$$

Moreover, there exists $C > 0$ such that

$$||u||^2_{L^\infty(0,T;H^2(\Omega_1))} + ||u_t||^2_{L^\infty(0,T;\vartheta_1)} \leq C[E_M(0) + ||u^0||^2_{H^2(\Omega_1)} + ||u^1||^2_{\vartheta_1}] \tag{2.12}$$

for every solution with initial data satisfying (2.9)–(2.10).

Proof. It is sufficient to prove the existence of some $\tau > 0$ and $C > 0$ such that (2.11) and (2.12) hold in the time interval $[0, \tau]$. By scaling the spatial variable in Ω_1 and changing the time scale we may assume $\ell_1 = 1$ and in the wave equation u satisfies, $\rho_1 = \sigma_1 = 1$. Likewise, by changing the length of the second string we may assume $\rho_2 = \sigma_2 = 1$ in the wave equation v satisfies. However, the conditions at the point mass change and we are led to consider the following system:

$$\begin{cases} u_{tt} = u_{xx}, & -1 < x < 0, \quad 0 < t < T, \\ v_{tt} = v_{xx}, & 0 < x < \ell, \quad 0 < t < T, \\ mz_{tt}(t) + u_x(0,t) - \gamma v_x(0,t) = 0, & 0 < t < T, \\ u(-1,t) = v(\ell,t) = 0, & 0 < t < T, \\ u(0,t) = v(0,t) = z(t), & 0 < t < T, \end{cases}$$

with $\gamma > 0$ and $m > 0$, ℓ being the length of the second string.

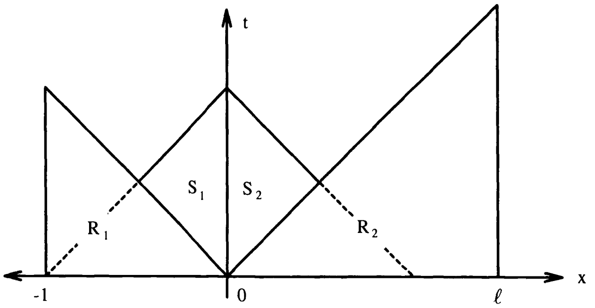


FIG. 1.

The value of u and v in the regions

$$R_1 = \{(x,t) \in (-1,0) \times (0,1) : t < -x\},$$
$$R_2 = \{(x,t) \in (0,\ell) \times (0,\ell) : t < x\},$$

respectively, does not depend on the point mass because of the finite speed of propagation (see Fig. 1).

Therefore, in R_1 and R_2, u and v remain as smooth as in the absence of mass. In particular $u \in C([0,1]; H^2(R_1(t)) \cap C^1([0,1]; H^1(R_1(t)))$ where $R_1(t) = (-1, -t)$. Moreover

$$||u||_{L^\infty(0,1;H^2(R_1(t)))} + ||u_t||_{L^\infty(0,1;H^1(R_1(t)))} \leq C\left[||u^0||_{H^2(-1,0)} + ||u^1||_{H^1(-1,0)}\right].$$

Let us compute u and v in S_1 and S_2, respectively, where

$$S_1 = \left\{(x,t) \in \left(-\frac{\mu}{2},0\right) \times (0,\mu) : |2t-\mu| < \mu + 2x\right\},$$
$$S_2 = \left\{(x,t) \in \left(0,\frac{\mu}{2}\right) \times (0,\mu) : |2t-\mu| < \mu - 2x\right\},$$

with $\mu = \min(\ell, 1)$ ($\mu = 1$ in Fig. 1).

We have from D'Alembert's formula

$$u(x,t) = \frac{1}{2}[z(t-x) + z(t+x)] + \frac{1}{2}\int_{t-x}^{t+x} u_x(0,s)ds \quad \text{in } S_1, \tag{2.13}$$

$$v(x,t) = \frac{1}{2}[z(t-x) + z(t+x)] + \frac{1}{2\gamma}\int_{t-x}^{t+x} [u_x(0,s) + mz''(s)]ds \quad \text{in } S_2. \tag{2.14}$$

Likewise for $(x.t)$ in R_1 and R_2, u and v, respectively, are given in terms of the initial data by D'Alembert's formula. By Proposition 2.1 we know that the solution is continuous, in particular, along the rays $t = |x|$, $(-\mu/2 < x < \mu/2)$. Imposing continuity in the expressions for u and v along these rays leads to

$$z_t(t) + \frac{(1+\gamma)}{m} z(t) = z^1 + \frac{1}{m}[L(t) + \gamma R(t)], \quad z(0) = z^0 \tag{2.15}$$

and

$$u_x(0,t) = z_t(t) + u_x^0(-t) - u^1(-t), \tag{2.16}$$

where

$$L(t) = u^0(t) - \int_0^{-t} u^1(s)ds; \quad R(t) = v^0(t) + \int_0^t v^1(s)ds. \tag{2.17}$$

From the conditions the initial data satisfy we deduce from (2.17) that L and R belong to $H^1(0,\mu)$. It thus follows from (2.15) that $z = z(t)$ belongs to $H^2(0,\mu)$. Then, from (2.10) and (2.16) we deduce that $u_x(0,t) \in H^1(0,\mu)$.

Now, from (2.13) we easily deduce that

$$u \in C\left(\left[0,\frac{\mu}{2}\right]; H^2(-t,0)\right) \cap C^1\left(\left[0,\frac{\mu}{2}\right]; H^1(-t,0)\right). \tag{2.18}$$

Finally, it is easy to check that the expressions we have for u in R_1 (as a solution of the wave equation) and in S_1 (by (2.13)) are such that u_x and u_t are continuous across $x = -t$. This concludes the proof of the proposition. □

Remark 2.1. It is obvious from the proof of this proposition that we cannot replace in the hypothesis (2.10) and in the conclusions (2.11)–(2.12) the space $H^2(-1,0) \times H^1(-1,0)$ by any $H^s(-1,0) \times H^{s-1}(-1,0)$ with $s > 2$. In other words, the most extra regularity degree we may keep in one of the strings is one.

2.2. Nonhomogeneous boundary data. In this section we prove the existence and uniqueness of weak solutions when we introduce $L^2(0,T)$-Dirichlet controls at the extreme points $x = -\ell_1, \ell_2$.

Let us consider the system

$$
(2.19)\qquad \begin{cases}
\rho_1 u_{tt} = \sigma_1 u_{xx}, & x \in \Omega_1,\ 0 < t < T,\\
\rho_2 v_{tt} = \sigma_2 v_{xx}, & x \in \Omega_2,\ 0 < t < T,\\
M z_{tt}(t) + \sigma_1 u_x(0,t) - \sigma_2 v_x(0,t) = 0, & 0 < t < T,\\
u(0,t) = v(0,t) = z(t), & 0 < t < T,\\
u(-\ell_1, t) = p(t), & 0 < t < T,\\
v(\ell_2, t) = q(t), & 0 < t < T,\\
u(x,0) = u^0(x), \quad u_t(x,0) = u^1(x), & x \in \Omega_1,\\
v(x,0) = v^0(x), \quad v_t(x,0) = v^1(x), & x \in \Omega_2,\\
z(0) = z^0, \quad z_t(0) = z^1,
\end{cases}
$$

with $p, q \in L^2(0,T)$, $u^0 \in L^2(\Omega_1)$, $v^0 \in L^2(\Omega_2)$, the initial velocity u^1 and v^1 belonging, respectively, to the dual spaces ϑ_1' and ϑ_2' and $z^0, z^1 \in \mathbb{R}$.

The solution (u, v, z) of (2.19) has to be understood in the sense of transposition. Let us give its precise definition. For that, consider the following system with homogeneous boundary conditions:

$$
(2.20)\qquad \begin{cases}
\rho_1 \varphi_{tt} = \sigma_1 \varphi_{xx} + f, & x \in \Omega_1, \quad 0 < t < T,\\
\rho_2 \psi_{tt} = \sigma_2 \psi_{xx} + g, & x \in \Omega_2, \quad 0 < t < T,\\
M \zeta_{tt}(t) + \sigma_1 \varphi_x(0,t) - \sigma_2 \psi_x(0,t) = h(t), & 0 < t < T,\\
\varphi(-\ell_1, t) = \psi(\ell_2, t) = 0, & 0 < t < T,\\
\varphi(0,t) = \psi(0,t) = \zeta(t), & 0 < t < T,\\
\varphi(x,T) = \varphi_t(x,T) = 0, & x \in \Omega_1,\\
\psi(x,T) = \psi_t(x,T) = 0, & x \in \Omega_2,\\
\zeta(T) = \zeta_t(T) = 0.
\end{cases}
$$

For every $(f, g, h) \in L^1(0,T; W_0)$, in view of the time-reversibility of the system and as a consequence of Proposition 2.3, system (2.20) has a unique solution

$$
(2.21)\qquad \begin{cases}
(\varphi, \psi, \zeta) \in C([0,T]; W_1),\\
(\varphi_t, \psi_t, \zeta_t) \in C([0,T]; W_0)
\end{cases}
$$

satisfying

$$
(2.22)\qquad \int_0^T [\,|\varphi_x(-\ell_1, t)|^2 + |\psi_x(\ell_2, t)|^2] dt \le C ||(f, g, h)||^2_{L^1(0,T;W_0)}.
$$

Multiplying by φ and ψ in the equations satisfied by u and v in (2.19) and integrating formally by parts with respect to x and t, we obtain the following identity:

$$\begin{aligned}\int_0^T\int_{-\ell_1}^0 ufdxdt+\int_0^T\int_0^{\ell_2} vgdxdt+\int_0^T zhdt &= -\rho_1\int_{-\ell_1}^0 u^0\varphi_t(x,0)dx\\ &-\rho_2\int_0^{\ell_2} v^0\psi_t(x,0)dx+\rho_1\langle u^1,\varphi(\cdot,0)\rangle_{\Omega_1}+\rho_2\langle v^1,\psi(\cdot,0)\rangle_{\Omega_2}\\ &+\sigma_1\int_0^T p(t)\varphi_x(-\ell_1,t)dt-\sigma_2\int_0^T q(t)\psi_x(\ell_2,t)dt+Mz^1\zeta(0)-Mz^0\zeta_t(0).\end{aligned}\tag{2.23}$$

We adopt this identity as the definition for weak solutions of (2.19) in the sense of transposition, i.e., (u,v,z) is said to be weak solution of (2.19) (in the sense of transposition) if (2.23) holds for every $(f,g,h)\in L^1(0,T;W_0)$.

In (2.23) observe that the initial velocities (u^1,v^1,z^1) are applied (in the sense of the duality in W_1) to the elements $(\varphi(\cdot,0),\psi(\cdot,0),\zeta(0))$ of W_1. Therefore two initial data that coincide in $W_0\times W_1'$ (note that W_1' is a quotient space of $(\vartheta_1\times\vartheta_2\times\mathbb{R})'$) give rise to the same solution.

We have the following result.

PROPOSITION 2.6. *For every $p,q\in L^2(0,T)$, $(u^0,v^0)\in L^2(\Omega_1)\times L^2(\Omega_2)$, $(u^1,v^1)\in(\vartheta_1'\times\vartheta_2')$, and $z^0,z^1\in\mathbb{R}$ there exists a solution (in the sense of transposition) of* (2.19) *in the class*

$$(u,v,z)\in C([0,T];W_0),\tag{2.24}$$

$$(u_t,v_t,z_t)\in C([0,T];\vartheta_1'\times\vartheta_2'\times\mathbb{R}).\tag{2.25}$$

Moreover, there is a one-to-one correspondence between the initial data as elements of the quotient space $W_0\times W_1'$ and the solutions of (2.19) *in the class* (2.24)–(2.25).

Proof. In view of Proposition 2.3, the right-hand side of (2.23) defines a linear and continuous form on $(f,g,h)\in L^1(0,T;W_0)$. Therefore, there exists a unique

$$(u,v,z)\in L^\infty(0,T;W_0)\tag{2.26}$$

satisfying (2.23). Furthermore, there exists $C>0$ such that

$$\begin{aligned}||(u,v,z)||_{L^\infty(0,T;W_0)}\le C\{&||p||_{L^2(0,T)}+||q||_{L^2(0,T)}+||u^0||_{L^2(\Omega_1)}\\ &+||v^0||_{L^2(\Omega_2)}+||u^1||_{\vartheta_1'}+||v^1||_{\vartheta_2'}+|z^0|+|z^1|\}.\end{aligned}\tag{2.27}$$

When the data are smooth, the solution of (2.19) satisfies (2.24). By a density argument using (2.27), we deduce that (2.24) holds for our weak solution.

Suppose now that

$$(f,g,h)=\left(\frac{\partial F}{\partial t},\frac{\partial G}{\partial t},\frac{dH}{dt}\right)$$

with $(F,G,H)\in\mathcal{D}((0,T);W_1)$ (C^∞ and compactly supported with respect to time). In this case the solution of (2.20) can be written as

$$(\varphi,\psi,\zeta)=\left(\frac{\partial\Phi}{\partial t},\frac{\partial\Psi}{\partial t},\frac{d\Sigma}{dt}\right),$$

where (Φ,Ψ,Σ) is the solution of (2.20) with data (F,G,H) instead of (f,g,h).

In view of Proposition 2.4 we have

$$\begin{aligned}(2.28)\quad &||\varphi_x(-\ell_1,t)||_{L^2(0,T)} + ||\psi_x(\ell_2,t)||_{L^2(0,T)} + ||\varphi(\cdot,0)||_{\vartheta_1} + ||\psi(\cdot,0)||_{\vartheta_2} + |\zeta(0)|\\ &\leq C||(F,G,H)||_{L^1(0,T,W_1)}.\end{aligned}$$

On the other hand,

$$\begin{aligned}\rho_1\varphi_t(x,0) &= \rho_1\Phi_{tt}(x,0) = \sigma_1\Phi_{xx}(x,0) + F(x,0) = \sigma_1\Phi_{xx}(x,0) \in L^2(\Omega_1),\\ \rho_2\psi_t(x,0) &= \rho_2\Psi_{tt}(x,0) = \sigma_2\Psi_{xx}(x,0) + G(x,0) = \sigma_2\Psi_{xx}(x,0) \in L^2(\Omega_2),\\ M\zeta_t(0) &= M\Sigma_{tt}(0) = -\sigma_2\Phi_x(0,0) + \sigma_2\Psi_x(0,0) \in \mathbb{R}\end{aligned}$$

with the bound

$$(2.29)\qquad ||(\varphi_t(\cdot,0),\psi_t(\cdot,0),\zeta_t(0))||_{W_0} \leq C||(F,G,H)||_{L^1(0,T;W_1)}.$$

As a consequence of (2.28)–(2.29) we deduce that

$$(u,v,z) \in W^{1,\infty}(0,T;(\vartheta_1)' \times (\vartheta_2)' \times \mathbb{R}).$$

The continuity in time of (u_t,v_t,z_t) with values in $(\vartheta_1)' \times (\vartheta_2)' \times \mathbb{R}$ can be proved again by density. □

Let us finally consider these weak solutions when the initial and boundary data corresponding to the first string have one more degree of regularity, i.e.,

$$(2.30)\quad p \in H^1(0,T),\ u^0 \in H^1(-\ell_1,0),\ u^1 \in L^2(\Omega_1),\ u^0(0) = z^0,\ p(0) = u^0(-\ell_1).$$

We have the following result.

PROPOSITION 2.7. *Suppose that the initial and boundary data in Proposition* 2.6 *satisfy the further regularity and compatibility conditions* (2.30). *Then, in addition to* (2.24)–(2.25) *we have*

$$(2.31)\qquad u \in C([0,T];H^1(\Omega_1)) \cap C^1([0,T];L^2(\Omega_1)).$$

Furthermore, $u_x(-\ell_1,t) \in L^2(0,T)$ *and there exists some* $C>0$ *such that*

$$\begin{aligned}(2.32)\quad \int_0^T |u_x(-\ell_1,t)|^2 dt \leq C\Big[&||q||^2_{L^2(0,T)} + ||p||^2_{H^1(0,T)} + ||u^0||^2_{H^1(\Omega_1)} + ||u^1||^2_{L^2(\Omega_1)}\\ &+||v^0||^2_{L^2(\Omega_2)} + ||v^1||^2_{(\vartheta_2)'} + |z^0|^2 + |z^1|^2\Big]\end{aligned}$$

Proof. The proof of (2.31) can be carried out in the same way as in Proposition 2.5. Indeed, in R_1 (see Fig. 1), the presence the nonhomogeneous boundary condition does not change the regularity of the solution, which matches that of the initial data. Thus (2.31) holds when restricted to R_1. When studying the regularity of the solution in S_1, the argument is the same as in Proposition 2.5, although since the regularities of the initial data are one degree less, the formulas (2.13)–(2.17) (which no longer have a pointwise interpretation) need to be justified by a density argument.

The regularity property (2.32) is a direct consequence of (2.31) and the local nature of the wave equation. More precisely, the trace regularity

$$(2.33)\qquad \int_0^T |u_x(-\ell_1,t)|^2\,dt \leq C\int_0^T ||u(\cdot,t)||^2_{H^1(\Omega_1)} + ||u_t(\cdot,t)||^2_{L^2(\Omega_1)}\,dt$$

is local and does not depend at all upon whether there is a point mass. Since u has the regularity in (2.31) it follows that the solution map $(p, q, u^0, v^0, z^0, u^1, v^1, z^1) \mapsto u$ is continuous from the space it is defined (with norm defined by the right hand side of (2.32)) into the space in (2.31). Thus

$$\int_0^T \|u(\cdot,t)\|^2_{H^1(\Omega_1)} + \|u_t(\cdot,t)\|^2_{L^2(\Omega_1)}\,dt \le C\Big[\|q\|^2_{L^2(0,T)} + \|p\|^2_{H^1(0,T)} + \|u^0\|^2_{H^1(\Omega_1)}$$
$$+\|u^1\|^2_{L^2(\Omega_1)} + \|v^0\|^2_{L^2(\Omega_2)} + \|v^1\|^2_{(\vartheta_2)'} + |z^0|^2 + |z^1|^2\Big]$$

holds and (2.32) follows. □

3. Control at both extremes. In this section we consider the problem of controlling our system from both ends $x = -\ell_1, \ell_2$. The system now reads

$$(3.1)\qquad \begin{cases} \rho_1 u_{tt} = \sigma_1 u_{xx}, & x \in \Omega_1,\ 0 < t < T,\\ \rho_2 v_{tt} = \sigma_2 v_{xx}, & x \in \Omega_2,\ 0 < t < T,\\ Mz_{tt}(t) + \sigma_1 u_x(0,t) - \sigma_2 v_x(0,t) = 0, & 0 < t < T,\\ u(-\ell_1, t) = p(t), & 0 < t < T,\\ v(\ell_2, t) = q(t), & 0 < t < T,\\ u(0,t) = v(0,t) = z(t), & 0 < t < T,\\ u(x,0) = u^0(x), \quad u_t(x,0) = u^1(x), & x \in \Omega_1,\\ v(x,0) = v^0(x), \quad v_t(x,0) = v^1(x), & x \in \Omega_2,\\ z(0) = z^0, \quad z_t(0) = z^1. \end{cases}$$

We have the following results.

THEOREM 3.1. *Suppose that $T > 2\max(\ell_1\sqrt{\rho_1/\sigma_1}, \ell_2\sqrt{\rho_2/\sigma_2})$. Then, for every*

$$(3.2)\qquad \begin{cases} (u^0, v^0, z^0) \in W_0,\\ (u^1, v^1, z^1) \in (\vartheta_1)' \times (\vartheta_2)' \times \mathbb{R}, \end{cases}$$

there exist controls $p, q \in L^2(0,T)$ such that the solution of (3.1) *satisfies*

$$(3.3)\qquad \begin{cases} u(x,T) = u_t(x,T) = 0, & x \in \Omega_1,\\ v(x,T) = v_t(x,T) = 0, & x \in \Omega_2,\\ z(T) = z_t(T) = 0. \end{cases}$$

Remark 3.1. Concerning Theorem 3.1 we have the following:

(1) As a consequence of Proposition 2.6 the solution of (3.1) (which is defined by transposition) satisfies (2.24)–(2.25).

(2) Due to the linearity and time-reversibility of system (3.1), we deduce that for any initial data as in (3.2) and final data $(\tilde{u}^0, \tilde{v}^0, \tilde{z}^0, \tilde{u}^1, \tilde{v}^1, \tilde{z}^1)$ satisfying the same properties, there exist controls $p, q \in L^2(0,T)$ such that the solution of (3.1) satisfies

$$(3.4)\qquad \begin{cases} u(x,T) = \tilde{u}^0(x), \quad u_t(x,T) = \tilde{u}^1(x), & x \in \Omega_1,\\ v(x,T) = \tilde{v}^0(x), \quad v_t(x,T) = \tilde{v}^1(x), & x \in \Omega_2,\\ z(T) = \tilde{z}^0, \quad z_t(T) = \tilde{z}^1. \end{cases}$$

(3) The controllability space (3.2) is that same as what one obtains when $M = 0$; however, the control time is strictly larger when $\ell_1\sqrt{\rho_1/\sigma_1} \neq \ell_2\sqrt{\rho_2/\sigma_2}$.

THEOREM 3.2. *Suppose that* $T > \ell_1\sqrt{\rho_1/\sigma_1}+\ell_2\sqrt{\rho_2/\sigma_2}$ *and for instance,* $\ell_1\sqrt{\rho_1/\sigma_1} > \ell_2\sqrt{\rho_2/\sigma_2}$. *Then, for every initial data as in* (3.2) *satisfying the additional regularity and compatibility properties*

$$u^0 \in \vartheta_1,\ u^1 \in L^2(\Omega_1),\ u^0(0) = z^0 \tag{3.5}$$

there exist controls $p \in H_0^1(0,T)$ *and* $q \in L^2(0,T)$ *such that the solution of* (3.1) *satisfies* (3.3).

Remark 3.2. Concerning Theorem 3.2 we have the following:

(1) In addition to satisfying (2.24)–(2.25), by Proposition 2.7 the solution of (3.1) also satisfies (2.31).

(2) Due to the linearity, time-reversibility, and well-posedness of the system in the (asymmetric) space (3.2), (3.5), as a consequence of Theorem 3.2, we deduce that we can drive system (3.1) from any initial state in the class (3.2), (3.5) to any terminal state in the same class.

(3) The control time we obtain is the same as in the absence of the mass ($M = 0$). However we only get controllability in the space (3.2) when $T > 2\max\left(\ell_1\sqrt{\rho_1/\sigma_1},\ \ell_2\sqrt{\rho_2/\sigma_2}\right)$.

The proof of Theorem 3.2 will be given at the end of §4.

Applying Lions' HUM (see below), Theorem 3.1 is a direct consequence of the following observability result for solution of the uncontrolled problem:

$$\begin{cases} \rho_1\varphi_{tt} = \sigma_1\varphi_{xx}, & x \in \Omega_1, \quad 0 < t < T, \\ \rho_2\psi_{tt} = \sigma_2\psi_{xx}, & x \in \Omega_2, \quad 0 < t < T, \\ M\zeta_{tt}(t) + \sigma_1\varphi_x(0,t) - \sigma_2\psi_x(0,t) = 0, & 0 < t < T, \\ \varphi(-\ell_1,t) = \psi(\ell_2,t) = 0, & 0 < t < T, \\ \varphi(0,t) = \psi(0,t) = \zeta(t), & 0 < t < T. \end{cases} \tag{3.6}$$

PROPOSITION 3.3. *Let* $T_0 = 2\max(\ell_1\sqrt{\rho_1/\sigma_1}, \ell_2\sqrt{\rho_2/\sigma_2})$ *and suppose that* $T > T_0$. *Then*

(3.7)
$$(T - T_0)\,E_M(0) \le \left(\frac{\max(\ell_1,\ell_2)}{2} + \frac{M}{2(\rho_1+\rho_2)}\right)\int_0^T [\sigma_1|\varphi_x(-1,t)|^2 + \sigma_2|\psi_x(1,t)|^2]dt$$

for every finite-energy solution of (3.6).

Proof of Proposition 3.3. We proceed in several steps.

Step 1. Consider the x-dependent energy:

$$e_1(x) = \frac{1}{2}\int_{(x+\ell_1)\tau_1}^{T-(x+\ell_1)\tau_1} \left[\rho_1|\varphi_t(x,t)|^2 + \sigma_1|\varphi_x(x,t)|^2\right] dt, \quad -\ell_1 \le x \le 0, \tag{3.8}$$

where $\tau_1 = \sqrt{\rho_1/\sigma_1}$. It is easy to check that $e_1(\cdot)$ is nonincreasing. Thus

$$e_1(x) \le e_1(-\ell_1) = \frac{\sigma_1}{2}\int_0^T |\varphi_x(-\ell_1,t)|^2 dt, \quad -\ell_1 \le x \le 0. \tag{3.9}$$

Step 2. Consider

$$e_2(x) = \frac{1}{2} \int_{(\ell_2 - x)\tau_2}^{T-(\ell_2-x)\tau_2} \left[\rho_2 |\psi_t(x,t)|^2 + \sigma_2 |\psi_x(x,t)|^2\right] dt, \quad 0 \le x \le \ell_2, \tag{3.10}$$

where $\tau_2 = \sqrt{(\rho_2/\sigma_2)}$. This energy is nondecreasing and therefore

$$e_2(x) \le e_2(\ell_2) = \frac{\sigma_2}{2} \int_0^T |\psi_x(\ell_2, t)|^2 dt, \quad 0 \le x \le \ell_2. \tag{3.11}$$

Step 3. From (3.9), (3.11) we deduce in particular that

$$\int_\mu^{T-\mu} \int_{-\ell_1}^{0} \left[\rho_1 |\varphi_t(x,t)|^2 + \sigma_1 |\varphi_x(x,t)|^2\right] dxdt + \int_\mu^{T-\mu} \int_0^{\ell_2} \left[\rho_2 |\psi_t(x,t)|^2 + \sigma_2 |\psi_x(x,t)|^2\right] dxdt$$

$$\le \max(\ell_1, \ell_2) \int_0^T \left[\sigma_1 |\varphi_x(-\ell_1, t)|^2 + \sigma_2 |\psi_x(\ell_2, t)|^2\right] dt,$$

where $\mu = \max(\tau_1 \ell_1, \tau_2 \ell_2)$.

Moreover, since $\varphi(0,t) = \psi(0,t) = \zeta(t)$, we have

$$M \int_\mu^{T-\mu} |\zeta_t(t)|^2 dt \le M(r\frac{e_1(0)}{\rho_1} + (2-r)\frac{e_2(0)}{\rho_2}) \,|_{r=\frac{2\rho_1}{\rho_1+\rho_2}}$$

$$\le \frac{M}{\rho_1 + \rho_2} \int_0^T [\sigma_1 |\varphi_x(-1,t)|^2 + \sigma_2 |\psi_x(1,t)|^2] dt.$$

Thus

$$(T - 2\mu) E_M(0) = \int_\mu^{T-\mu} E_M(t) dt$$

$$\le \left(\frac{\max(\ell_1, \ell_2)}{2} + \frac{M}{2(\rho_1 + \rho_2)}\right) \int_0^T [\sigma_1 |\varphi_x(-\ell_1, t)|^2 + \sigma_2 |\psi_x(\ell_2, t)|^2] dt,$$

which is precisely (3.7). □

Remark 3.3. (1) Inequality (3.7) provides explicit constants. In particular, we have explicitly the dependence of the observability constant with respect to the mass. As $M \to 0$ we obtain the usual constant one has for a wave equation with piecewise constant coefficients.

(2) The reverse inequality of (3.7) has been proved in Proposition 2.2.

End of Proof of Theorem 3.1: *Application of HUM.* For any $(\varphi^0, \psi^0, \zeta^0, \varphi^1, \psi^1, \zeta^1)^t \in \mathcal{H}$ we define

$$\Lambda(\varphi^0, \psi^0, \zeta^0, \varphi^1, \psi^1, \zeta^1) = (u_t(0), v_t(0), z_t(0), -u(0), -v(0), -z(0)) \in \mathcal{H}' \text{ (dual of } \mathcal{H}), \tag{3.12}$$

where (u, v, z) is the solution of

$$(3.13) \qquad \begin{cases} \rho_1 u_{tt} = \sigma_1 u_{xx}, & x \in \Omega_1,\ 0 < t < T, \\ \rho_2 v_{tt} = \sigma_2 v_{xx}, & x \in \Omega_2,\ 0 < t < T, \\ M z_{tt}(t) + \sigma_1 u_x(0,t) - \sigma_2 v_x(0,t) = 0, & 0 < t < T, \\ u(-\ell_1, t) = -\varphi_x(-\ell_1, t), & 0 < t < T, \\ v(\ell_2, t) = \psi_x(\ell_2, t), & 0 < t < T, \\ u(0,t) = v(0,t) = z(t), & 0 < t < T, \\ u(x,T) = u_t(x,T) = 0, & x \in \Omega_1, \\ v(x,T) = v_t(x,T) = 0, & x \in \Omega_2, \\ z(T) = z_t(T) = 0, & \end{cases}$$

and $(\varphi, \psi, \zeta, \dot{\varphi}, \dot{\psi}, \dot{\zeta})^t$ is the solution of (2.1) with $y^0 = (\varphi^0, \psi^0, \zeta^0, \varphi^1, \psi^1, \zeta^1)^t$.

In view of Proposition 2.2 we have

$$\varphi_x(-\ell_1, \cdot), \psi_x(\ell_2, \cdot) \in L^2(0,T)$$

and as a consequence of Proposition 2.6, (u, v, z) satisfies (2.24)–(2.25). In particular, $(u_t(0), v_t(0), z_t(0), -u(0), -v(0), -z(0))$ is well defined as an element of $\mathcal{H}'$.

Thus $\Lambda : \mathcal{H} \to \mathcal{H}'$ is continuous and linear.

Using the transposition formula and Proposition 3.3 one obtains

$$\begin{aligned} \langle \Lambda(\varphi^0, \psi^0, \zeta^0, \varphi^1, \psi^1, \zeta^1), (\varphi^0, \psi^0, \zeta^0, \varphi^1, \psi^1, \zeta^1) \rangle &= \int_0^T \left[|\psi_x(\ell_2, t)|^2 + |\varphi_x(-\ell_1, t)^2 \right] dt \\ &\geq C E_M(0). \end{aligned}$$

Taking into account that $(E_M(0))^{1/2}$ is equivalent to the norm of $(\varphi^0, \psi^0, \zeta^0, \varphi^1, \psi^1, \zeta^1)$ in $\mathcal{H}$, we conclude that $\Lambda : \mathcal{H} \to \mathcal{H}'$ is an isomorphism.

Given any $(u^0, v^0, z^0) \in W_0$ and any $(u^1, v^1, z^1) \in \vartheta_1' \times \vartheta_2' \times \mathbb{R}$, we have $(u^1, v^1, z^1) \in W_1'$ and therefore the equation

$$\Lambda(\varphi^0, \psi^0, \zeta^0, \varphi^1, \psi^1, \zeta^1) = (u^1, v^1, z^1, -u^0, -v^0, -z^0)$$

admits a unique solution $(\varphi^0, \psi^0, \zeta^0, \varphi^1, \psi^1, \zeta^1) \in \mathcal{H}$, but this is equivalent to the fact that the solution of (3.13) satisfies

$$\begin{cases} u(x,0) = u^0(x), \quad u_t(x,0) = u^1(x), & x \in \Omega_1, \\ v(x,0) = v^0(x), \quad v_t(x,0) = v^1(x), & x \in \Omega_2, \\ z(0) = z^0, \quad z_t(0) = z^1. & \end{cases}$$

This concludes the proof of Theorem 3.1. □

4. Control at one extreme. In this section we consider the problem of controlling our system from only one extreme-point, for instance, $x = \ell_2$. The system now reads

$$\left\{\begin{array}{ll} \rho_1 u_{tt} = \sigma_1 u_{xx}, & x \in \Omega_1,\ 0<t<T, \\ \rho_2 v_{tt} = \sigma_2 v_{xx}, & x \in \Omega_2,\ 0<t<T, \\ M z_{tt}(t) + \sigma_1 u_x(0,t) - \sigma_2 v_x(0,t) = 0, & 0<t<T, \\ u(0,t) = v(0,t) = z(t), & 0<t<T, \\ u(-\ell_1, t) = 0, & 0<t<T, \\ v(\ell_2, t) = q(t), & 0<t<T, \\ u(x,0) = u^0(x), \quad u_t(x,0) = u^1(x), & x \in \Omega_1, \\ v(x,0) = v^0(x), \quad v_t(x,0) = v^1(x), & x \in \Omega_2, \\ z(0) = z^0, \quad z_t(0) = z^1. & \end{array}\right. \tag{4.1}$$

We have the following result.

THEOREM 4.1. *Suppose that* $T > 2(\ell_1\sqrt{\rho_1/\sigma_1} + \ell_2\sqrt{\rho_2/\sigma_2})$. *Then, for every*

$$\left\{\begin{array}{l} (u^0, u^1) \in \vartheta_1 \times L^2(\Omega_1), \\ (v^0, v^1) \in L^2(\Omega_2) \times (\vartheta_2)', \\ (z^0, z^1) \in \mathbb{R}^2, \end{array}\right. \tag{4.2}$$

such that

$$u^0(0) = z^0 \tag{4.3}$$

there exists a control function $q \in L^2(0,T)$ *such that the solution of* (4.1) *satisfies* (3.3).

Remark 4.1. (1) As a consequence of Proposition 2.6 the solution of (4.1) satisfies (2.24)–(2.25). Furthermore, by Proposition 2.7, it also satisfies (2.31).

(2) Due to the linearity, time-reversibility, and well-posedness of the system in the (asymmetric) space (4.2)–(4.3), as a consequence of Theorem 4.1, we deduce that we can drive system (4.1) from any initial state to any final state in the class (4.2)–(4.3).

(3) The control time we obtain is the same as for the case $M = 0$. However, notice that, when $M = 0$, the controllable space is larger and is given by (3.2). The controllable space we obtain for $M > 0$ is the optimal one.

By HUM (see the proof of Theorem 3.1 for more details), Theorem 4.1 is equivalent to the following observability result for the solutions of the uncontrolled problem (3.6).

PROPOSITION 4.2. *Suppose that* $T > 2\left(\ell_1\sqrt{\rho_1/\sigma_1} + \ell_2\sqrt{\rho_2/\sigma_2}\right)$. *Then there exists* $C = C(T) > 0$ *such that the following holds for every finite-energy solution of* (3.6):

$$\begin{aligned} &||\varphi(\cdot,0)||^2_{L^2(\Omega_1)} + ||\varphi_t(\cdot,0)||^2_{(\vartheta_1)'} + ||\psi(\cdot,0)||^2_{\vartheta_2} + ||\psi_t(\cdot,0)||^2_{L^2(\Omega_2)} \\ &\qquad + |\zeta(0)|^2 + |\zeta_t(0)|^2 \le C\int_0^T |\psi_x(\ell_2,t)|^2 dt. \end{aligned} \tag{4.4}$$

Remark 4.2. This inequality is sharp in the sense that, as pointed out in Proposition 2.7, the reverse holds for all $T > 0$. As a consequence of this double inequality we deduce that the controllable space (4.2)–(4.3) is the optimal one.

Proof. However, first note that by rescaling the spatial variables and renaming the densities ρ_1 and ρ_2 we can obtain the system (4.1) with $\ell_1 = \ell_2 = 1$. Thus it will suffice to prove Proposition 4.2 for the case $\Omega_1 = (-1,0)$ and $\Omega_2 = (0,1)$.

We proceed in several steps.

Step 1. Consider the space-dependent energy $e_2(\cdot)$ defined in (3.10). Since it is nondecreasing we have

$$e_2(x) \le e_2(1) = \frac{\sigma_2}{2}\int_0^T |\psi_x(1,t)|^2 dt, \quad 0 \le x \le 1.$$

This implies, in particular, that

$$\int_{\tau_2}^{T-\tau_2}\int_0^1 \left[\rho_2|\psi_t(x,t)|^2 + \sigma_2|\psi_x(x,t)|^2\right] dxdt \le \sigma_2 \int_0^T |\psi_x(1,t)|^2 dt, \tag{4.5}$$

where $\tau_2 = \sqrt{\rho_2/\sigma_2}$. Taking into account that $\psi(1,t) = 0$ for $\tau_2 \le t \le T - \tau_2$ from Poincaré's inequality and (4.5) we deduce that

$$\int_{\tau_2}^{T-\tau_2} |\psi(0,t)|^2 dt \le C\int_0^T |\psi_x(1,t)|^2 dt$$

and therefore, since $\varphi(0,t) = \psi(0,t) = \zeta(t)$,

$$\int_{\tau_2}^{T-\tau_2} \left[\, |\varphi(0,t)|^2 + |\zeta(t)|^2\right] dt \le C\int_0^T |\psi_x(1,t)|^2 dt. \tag{4.6}$$

On the other hand, since $e_2(0) \le e_2(1)$ we have

$$\int_{\tau_2}^{T-\tau_2} \left[\rho_2|\psi_t(0,t)|^2 + \sigma_2|\psi_x(0,t)|^2\right] dt \le \sigma_2\int_0^T |\psi_x(1,t)|^2 dt \tag{4.7}$$

and therefore, since $\zeta(t) = \psi(0,t)$,

$$\int_{\tau_2}^{T-\tau_2} \rho_2|\zeta_t(t)|^2 dt \le \sigma_2\int_0^T |\psi_x(1,t)|^2 dt. \tag{4.8}$$

Furthermore, since $\sigma_1\varphi_x(0,t) = -M\zeta_{tt}(t) + \sigma_2\psi_x(0,t)$, in view of (4.7)–(4.8) we have

$$\|\varphi_x(0,t)\|^2_{H^{-1}(\tau_2,T-\tau_2)} \le C\int_0^T |\psi_x(1,t)|^2 dt. \tag{4.9}$$

Step 2. As a consequence of (4.6) and (4.9) we have

$$\|\varphi(0,t)\|^2_{L^2(\tau_2,T-\tau_2)} + \|\varphi_x(0,t)\|^2_{H^{-1}(\tau_2,T-\tau_2)} \le C\int_0^T |\psi_x(1,t)|^2 dt. \tag{4.10}$$

The well-posedness of the one-dimensional wave equation as an evolution in x-variable allows us to prove that (see [17], [18] for details)

$$\int_{-1}^{0}\int_{\tau_2-\tau_1 x}^{T-\tau_2+\tau_1 x} |\varphi(x,t)|^2 dtdx \le C\left[\|\varphi(0,t)\|^2_{L^2(\tau_2,T-\tau_2)} + \|\varphi_x(0,t)\|^2_{H^{-1}(\tau_2,T-\tau_2)}\right], \tag{4.11}$$

where $\tau_1 = \sqrt{\rho_1/\sigma_1}$. Combining (4.10) and (4.11), we deduce that

$$\int_{\tau_2+\tau_1}^{T-(\tau_2+\tau_1)}\int_{1}^{0} |\varphi(x,t)|^2 dxdt \le C\int_{0}^{T} |\varphi_x(1,t)|^2 dt. \tag{4.12}$$

Step 3. Set $\mu = \tau_1+\tau_2$ and let $\varepsilon > 0$ be such that

$$T - 2\varepsilon > 2\mu \tag{4.13}$$

and $\eta = \eta(t) \in C^1(\mu, T-\mu)$ such that

$$\begin{cases} 0 \le \eta(t) \le 1, \quad \mu \le t \le T-\mu, \\ \eta(\mu) = \eta(t-\mu) = 0, \\ \eta(t) = 1, \quad \mu+\varepsilon \le t \le T-\mu-\varepsilon, \\ \frac{|\eta_t|^2}{\eta} \in L^\infty(\mu, T-\mu). \end{cases} \tag{4.14}$$

Let us introduce the function $\phi = \phi(x,t)$ such that

$$\begin{cases} -\phi_{xx} = \varphi, \quad -1 < x < 0, \quad \mu \le t \le T-\mu, \\ \phi(-1,t) = \phi_x(0,t) = 0, \quad \mu \le t \le T-\mu. \end{cases} \tag{4.15}$$

We have

$$-\int_{-1}^{0} \varphi_{xx}(x,t)\phi(x,t)dx = \int_{-1}^{0} |\varphi(x,t)|^2 dx - \varphi_x(0,t)\phi(0,t), \quad \mu \le t \le T-\mu \tag{4.16}$$

and

$$\int_{-1}^{0} \varphi_t(x,t)\phi_t(x,t)dx = \int_{-1}^{0} |\phi_{tx}(x,t)|^2 dx = \|\varphi_t(\cdot,t)\|^2_{(\vartheta_1)'}, \quad \mu \le t \le T-\mu. \tag{4.17}$$

We multiply by $\phi\eta$ the equation satisfied by φ and integrate in the region $(x,t) \in \Omega_1 \times (\mu, T-\mu)$ to get, in view of (4.16)–(4.17),

$$\begin{aligned} \rho_1 \int_{\mu}^{T-\mu} \|\varphi_t(\cdot,t)\|^2_{\vartheta_1'}\eta(t)dt = &-\rho_1 \int_{\mu}^{T-\mu}\int_{-1}^{0} \varphi_t\phi\eta_t dxdt + \sigma_1 \int_{\mu}^{T-\mu}\int_{-1}^{0} |\varphi|^2\eta dxdt \\ &- \sigma_1 \int_{\mu}^{T-\mu} \varphi_x(0,t)\phi(0,t)\eta(t)dt. \end{aligned} \tag{4.18}$$

On the other hand,

$$\left|\int_\mu^{T-\mu}\int_{-1}^{0}\varphi_t\phi\eta_t dxdt\right| \le \frac{1}{2}\int_\mu^{T-\mu}\|\varphi_t(\cdot,t)\|^2_{\vartheta_1'}\eta(t)dt + \frac{1}{2}\int_\mu^{T-\mu}\|\phi(\cdot,t)\|^2_{\vartheta_1}\frac{|\eta_t|^2}{\eta}dt. \tag{4.19}$$

Note that

$$\|\phi(\cdot,t)\|_{\vartheta_1} = \|\varphi(\cdot,t)\|_{\vartheta_1'} \le C\|\varphi(\cdot,t)\|_{L^2(\Omega_1)}, \quad \mu \le t \le T-\mu. \tag{4.20}$$

Combining (4.18)–(4.20) we deduce that

$$\int_\mu^{T-\mu}\|\varphi_t(\cdot,t)\|^2_{\vartheta_1'}\eta(t)dt \le C_1\left\{\int_\mu^{T-\mu}\int_{-1}^{0}\varphi^2 dxdt + \sigma_1\left|\int_\mu^{T-\mu}\varphi_x(0,t)\phi(0,t)\eta(t)dt\right|\right\}. \tag{4.21}$$

Since $\sigma_1\varphi_x(0,t) = \sigma_2\psi_x(0,t) - M\zeta_{tt}(t)$, we have

$$\begin{aligned}\sigma_1\int_\mu^{T-\mu}\varphi_x(0,t)\phi(0,t)\eta(t)dt &= \sigma_2\int_\mu^{T-\mu}\psi_x(0,t)\phi(0,t)\eta(t)dt\\ &\quad - M\int_\mu^{T-\mu}\zeta_{tt}(t)\phi(0,t)\eta(t)dt.\end{aligned} \tag{4.22}$$

On the other hand,

$$\int_\mu^{T-\mu}\zeta_{tt}(t)\phi(0,t)\eta(t)dt = -\int_\mu^{T-\mu}\zeta_t(t)\phi_t(0,t)\eta(t)dt - \int_\mu^{T-\mu}\zeta_t(t)\phi(0,t)\eta_t(t)dt. \tag{4.23}$$

Now, observe that

$$|\phi_t(0,t)| \le C\|\phi_t(\cdot,t)\|_{\vartheta_1} \le C\|\varphi_t(\cdot,t)\|_{\vartheta_1'} \quad \mu \le t \le T-\mu.$$

Therefore, there exists $C>0$ such that

$$\left|\int_\mu^{T-\mu}\zeta_t(t)\phi_t(0,t)\eta(t)dt\right| \le \frac{1}{2}\int_\mu^{T-\mu}\|\varphi_t(\cdot,t)\|^2_{\vartheta_1'}\eta(t)dt + C\int_\mu^{T-\mu}|\zeta_t(t)|^2dt. \tag{4.24}$$

Moreover

$$|\phi(0,t)| \le C\|\phi(\cdot,t)\|_{\vartheta_1} \le C\|\varphi(\cdot,t)\|_{L^2(-1,0)}$$

and thus

$$\left|\int_\mu^{T-\mu}\zeta_t(t)\phi(0,t)\eta_t dt\right| \le C\left\{\int_\mu^{T-\mu}\int_{-1}^{0}\varphi^2 dxdt + \int_\mu^{T-\mu}|\zeta_t(t)|^2dt\right\} \tag{4.25}$$

and

$$\left|\int_\mu^{T-\mu}\psi_x(0,t)\phi(0,t)\eta dt\right| \le C\left\{\int_\mu^{T-\mu}\int_{-1}^{0}\varphi^2 dxdt + \int_\mu^{T-\mu}|\psi_x(0,t)|^2dt\right\}. \tag{4.26}$$

Combining (4.21)–(4.26) with $\varepsilon > 0$ small enough, we deduce that

$$\begin{aligned}\int_{\mu+\varepsilon}^{T-\mu-\varepsilon} \|\varphi_t(\cdot,t)\|^2_{\vartheta_1'} dt &\leq \int_{\mu}^{T-\mu} \|\varphi_t(\cdot,t)\|^2_{\vartheta_1'} \eta(t) dt \\ &\leq C\left\{ \int_{\mu}^{T-\mu}\int_{-1}^{0} \varphi^2 dx dt + \int_{\mu}^{T-\mu} [\,|\zeta_t(t)|^2 + |\psi_x(0,t)|^2] dt \right\}.\end{aligned} \tag{4.27}$$

Step 4. As a consequence of (4.5), (4.6), (4.8), (4.12), and (4.27) we have

$$\begin{aligned}\int_{\mu+\varepsilon}^{T-\mu-\varepsilon} \left[\int_{-1}^{0} \varphi^2 dx + \|\varphi_t(\cdot,t)\|^2_{\vartheta_1'} + \int_0^1 \left[\,|\psi_x|^2 + |\psi_t|^2\right] dx + |\zeta(t)|^2 + |\zeta_t(t)|^2 \right] dt \\ \leq C \int_0^T |\psi_x(1,t)|^2 dt.\end{aligned} \tag{4.28}$$

Recall that, in virtue of Proposition 2.7, system (3.6) is well posed in the space

$$\begin{cases} \varphi \in L^2(-1,0), & \varphi_t \in \vartheta_1', \\ \psi \in \vartheta_2, & \psi_t \in L^2(0,1), \\ \zeta, \zeta_t \in \mathbb{R} \end{cases}$$

with compatibility condition

$$\psi(0) = \zeta.$$

Therefore, (4.4) is an immediate consequence of (4.28) and the time-reversibility of system (3.6). □

We are now in a position to prove Theorem 3.2.

Proof of Theorem 3.2. Theorem 3.2 is a direct consequence by HUM (see proof of Theorem 3.1) of the following observability result for the uncontrolled problem (3.6).

PROPOSITION 4.3. *Suppose that* $T > \ell_1\sqrt{\rho_1/\sigma_1} + \ell_2\sqrt{\rho_2/\sigma_2}$ *and* $\ell_1\sqrt{\rho_1/\sigma_1} > \ell_2\sqrt{\rho_2/\sigma_2}$. *Then there exists* $C(T) > 0$ *such that the following holds for every finite energy solution of* (3.6):

$$\begin{aligned}\|\varphi(\cdot,0)\|^2_{L^2(\Omega_1)} + \|\varphi_t(\cdot,0)\|^2_{(\vartheta_1)'} + \|\psi(\cdot,0)\|^2_{\vartheta_2} + \|\psi_t(\cdot,0)\|^2_{L^2(\Omega_2)} \\ + |\xi(0)|^2 + |\xi_t(0)|^2 \leq C\left[\|\varphi_x(-\ell_1,t)\|^2_{H^{-1}(0,T)} + \|\psi_x(\ell_2,t)\|^2_{L^2(0,T)}\right].\end{aligned} \tag{4.29}$$

Proof of Proposition 4.3. As in the proof of Proposition 4.2 we assume that $\ell_1 = \ell_2 = 1$ without loss of generality.

Let τ_1 and τ_2 be the same as in the proof of Proposition 4.2 and let $\mu = \frac{\tau_1+\tau_2}{2}$. For any sufficiently small $\varepsilon > 0$ we have

$$\begin{aligned}\int_{\mu+\varepsilon}^{T-\mu-\varepsilon} \left[\|\varphi(\cdot,t)\|^2_{L^2\left(-1,\frac{\tau_2-\tau_1}{2\tau_1}\right)} + \|\varphi_t(\cdot,t)\|^2_{H^{-1}\left(-1,\frac{\tau_2-\tau_1}{2\tau_1}\right)}\right] dt \\ \leq C\|\varphi_x(-1,t)\|^2_{H^{-1}(0,T)}\end{aligned} \tag{4.30}$$

and

$$\begin{aligned}
&\int_{\mu+\varepsilon}^{T-\mu-\varepsilon} \left[||\varphi(\cdot,t)||^2_{L^2\left(\frac{\tau_2-\tau_1}{2\tau_1},0\right)} + ||\varphi_t(\cdot,t)||^2_{\left(H^1\left(\frac{\tau_2-\tau_1}{2\tau_1},0\right)\right)'} \right] dt \\
&\quad + \int_{\mu+\varepsilon}^{T-\mu-\varepsilon} \int_0^1 [|\psi_t(x,t)|^2 + |\psi_x(x,t)|^2] dx dt + \int_{\mu+\varepsilon}^{T-\mu-\varepsilon} [|\xi(t)|^2 + |\xi_t(t)|^2] dt \\
&\quad \leq C \int_0^T |\psi_x(1,t)|^2 dt.
\end{aligned} \tag{4.31}$$

The inequality (4.30) is a standard estimate that holds for the wave equation (which we may apply here due to the finite speed of propagation) while (4.31) is proved in the same way that (4.28) was.

Since

$$\int_{\mu+\varepsilon}^{T-\mu-\varepsilon} ||\varphi_t(\cdot,t)||^2_{(\vartheta_1)'} \leq C \int_{\mu+\varepsilon}^{T-\mu-\varepsilon} \left[||\varphi_t(\cdot,t)||^2_{H^{-1}\left(-1,\frac{\tau_2-\tau_1}{2\tau_1}\right)} + ||\varphi_t(\cdot,t)||^2_{\left(H^1\left(\frac{\tau_2-\tau_1}{2\tau_1},0\right)\right)'} \right] dt,$$

the inequalities (4.30) and (4.31) can be combined and we deduce easily (4.29) as in Step 4 of the proof of Proposition 4.2. □

5. Representation of controllability spaces by Fourier series.

In this section we give a characterization of the controllable and observable spaces for (4.1) in terms of nonharmonic Fourier series. For simplicity, we limit our analysis to the case where

$$\ell_1 = \ell_2 = \sigma_1 = \sigma_2 = \rho_1 = \rho_2 = 1, \tag{5.1}$$

although a similar analysis is valid, for example, when the above parameters are rational.

5.1. Spectral analysis.

We begin with a spectral analysis of the operator $\mathcal{A}$ in (2.1). Since $\mathcal{A}$ is skew adjoint with compact inverse, $\sigma(\mathcal{A})$ consists of a discrete sequence of imaginary eigenvalues. We seek nontrivial solutions y for

$$\mathcal{A}y = i\omega y, \quad \omega \in \mathbb{R}. \tag{5.2}$$

We obtain nontrivial solutions only when $\omega \in S_1 \cup S_2$, where

$$\begin{cases} S_1 = (k\pi)_{k\in\mathbb{Z}\setminus\{0\}}, \\ S_2 = (\omega_k)_{k=0^-,0^+,\pm1,\pm2,\ldots}, \end{cases} \tag{5.3}$$

where $\omega_{0^+} = -\omega_{0^-}$ and for $k \in \mathbb{N}$, $\omega_k = -\omega_{-k}$, where ω_{k-1} is the kth positive root of

$$M\omega = 2\cot\omega. \tag{5.4}$$

It is easy to see from (5.4) that $|\omega_k - k\pi| \to 0$ as $k \to \infty$. In fact, if we let

$$\delta_k = \omega_k - k\pi, \quad k \in \mathbb{N}, \tag{5.5}$$

a simple calculation using Taylor's formula applied to (5.4) gives

$$\delta_k = \frac{2}{Mk\pi} + \mathcal{O}(k^{-2}) \quad \text{as } k \to \infty. \tag{5.6}$$

The eigenfunctions of $\mathcal{A}$ corresponding to the eigenvalue ω are given by

$$\varphi_\omega(x) = \begin{pmatrix} U_\omega \\ V_\omega \\ Z_\omega \\ \dot{U}_\omega \\ \dot{V}_\omega \\ \dot{Z}_\omega \end{pmatrix} = \begin{pmatrix} \frac{a}{\omega} \sin \omega(x+1) \\ \frac{1}{\omega} \sin \omega(x-1) \\ -\frac{(1-a)}{M\omega^2} \cos \omega \\ ia \sin \omega(x+1) \\ i \sin \omega(x-1) \\ -\frac{i(1-a)}{M\omega} \cos \omega \end{pmatrix}, \tag{5.7}$$

where $a = 1$ when $\omega \in S_1$ and $a = -1$ when $\omega \in S_2$.

Under the above normalization there exist positive constants C_1, C_2 for which

$$C_1 \leq \|\varphi_\omega\|_{\mathcal{H}} \leq C_2 \quad \forall \omega \in \sigma(\mathcal{A}). \tag{5.8}$$

Remark 5.1. The form of the eigenfunctions in (5.7) shows that each spectral class, S_1 and S_2, has its own physical significance; namely, if $\omega \in S_1$ then φ_ω describes a sinusoidal motion of the string which does not move the mass, while if $\omega \in S_2$ then φ_ω describes an even, piecewise sinusoidal motion of the string with a jump in the spatial derivative at the point mass. It is insightful to note that, although $(\varphi_\omega)_{i\omega \in \sigma(\mathcal{A})}$ forms an orthogonal set, $\|V_{k\pi} - V_{\omega_k}\|_{H^1(\Omega_2)} + \|\dot{V}_{k\pi} - \dot{V}_{\omega_k}\|_{L^2(\Omega_2)} \to 0$ as $k \to \infty$. It thus becomes increasingly difficult to distinguish such consecutive modes (as the frequency gets large) by only observing the v (or only the w) portion of the state. This explains why, in terms of the eigenfunctions, we obtain an asymmetric observability (and hence also controllability) space when we only observe (or control) at an endpoint.

5.2. Reduction to moment problem. Let us consider a general class of control problems which will include the problem (4.1).

Let $b \in \mathcal{D}(\mathcal{A})'$ and consider the system

$$\dot{y} = \mathcal{A}y + bg(t), \quad y(0) = 0, \tag{5.9}$$

where $g \in L^2(0,T)$. A unique mild solution $y \in C([0,T]; D(\mathcal{A})')$ is given by

$$y(t) = \Phi_t g, \tag{5.10}$$

where for $0 \leq t \leq T$, $\Phi_t : L^2(0,T) \to \mathcal{D}(\mathcal{A})'$ is given by

$$\Phi_t h = \int_0^t e^{\mathcal{A}(t-s)} bh(s) ds. \tag{5.11}$$

This notion of mild solution coincides with that of the weak solutions in the sense of transposition introduced in §2. (Note that $\mathcal{D}(\mathcal{A})' = W_0 \times W_1'$.) We adopt here this notation and terminology of the theory of semigroups for the sake of clarity and brevity.

The central problem for controllability of (5.9) is to determine the range of Φ_t. By integrating (5.10) against φ_ω we obtain the modal solutions

$$y_\omega(t) = \int_0^t e^{i\omega(t-s)} b_\omega g(s) ds, \quad \omega \in S_1 \cup S_2, \tag{5.12}$$

where

$$y_\omega(t) = (y(t), \varphi_\omega), \quad b_\omega = (b, \varphi_w),$$

and $(\cdot,\cdot)$ denotes the duality pairing of $\mathcal{D}(\mathcal{A})'$ and $\mathcal{D}(\mathcal{A})$ relative to $\langle\cdot,\cdot\rangle_{\mathcal{H}}$. Thus $x = \sum_{\omega\in\sigma(\mathcal{A})} x_\omega \varphi_\omega \in \mathcal{R}(\Phi_T)$ if and only if there exist $g \in L^2(0,T)$ for which

$$x_\omega = \int_0^T e^{i\omega s} b_w g(s) ds \quad \forall \omega \in S_1 \cup S_2. \tag{5.13}$$

Hence the problem of determining $\mathcal{R}(\Phi_T)$ has been replaced by one of determining the moment space (i.e., the sequences (x_ω) associated with some $g \in L^2(0,T)$ in (5.13)) of the moment problem (5.13).

5.3. Some general results on moment problems. Before determining the moment space of (5.13) we first give some background and some results of independent interest.

A Riesz basis for a Hilbert space X is the image of an orthonormal basis through a bounded, invertible operator $B : X \to X$.

The following is due to Ullrich [15].

THEOREM 5.1. *Let $(\sigma_n)_{n\in\mathbb{Z}}$ be a sequence of distinct, noninteger, complex numbers with* $\lim_{|n|\to\infty} |\sigma_n - n| = 0$. *Then*

$$(e^{int})_{n\in\mathbb{Z}} \cup \left(\frac{e^{int} - e^{i\sigma_n t}}{n - \sigma_n}\right)_{n\in\mathbb{Z}} \tag{5.14}$$

forms a Riesz basis for $L^2(0,4\pi)$.

COROLLARY 5.2. *For each $T \geq 4\pi$ the moment problem*

$$\int_0^T e^{int} f(t)\, dt = a_n, \quad \int_0^T e^{i\sigma_n t} f(t)\, dt = b_n; \quad n \in \mathbb{Z}, \tag{5.15}$$

has a solution $f \in L^2(0,T)$ if and only if

$$\left(\frac{a_n - b_n}{n - w_n}\right) \in \ell^2, \quad (a_n) \in \ell^2, \quad (b_n) \in \ell^2. \tag{5.16}$$

Proof. It is obvious that if the result is true when $T = 4\pi$, then it is true for $T > 4\pi$. Hence we may assume $T = 4\pi$. By Theorem 5.1 we know that

$$\int_0^T e^{int} f(t) dt = c_k, \quad \int_0^T \frac{e^{int} - e^{i\sigma_n t}}{n - \sigma_n} f(t)\, dt = d_k \tag{5.17}$$

has a solution if and only if $(c_k) \in \ell^2$ and $(d_k) \in \ell^2$. Putting $a_n = c_n - d_n(n - \sigma_n)$ and $b_n = c_n$, $n \in \mathbb{Z}$ we recover (5.15).

Since

$$(c_n) \in \ell^2 \text{ and } (d_n) \in \ell^2 \Leftrightarrow \text{(5.16) holds}$$

the result holds. $\square$

When applied to the moment problem (5.13), Corollary 5.2 completely describes the moment space. However it will be more convenient to use a dual version of Corollary 5.2, which we give below after we develop some notation.

Let

$$\mathcal{M} = \{(a_n) \cup (b_n) : (a_n), (b_n) \text{ satisfy (5.16)}\}.$$

Then $\mathcal{M}$ becomes a Hilbert space when endowed with the inner product

$$(5.18)\quad \langle (a_n)\cup(b_n),(\hat{a}_n)\cup(\hat{b}_n)\rangle_{\mathcal{M}} = \langle (a_n),(\hat{a}_n)\rangle_{\ell^2} + \left\langle \left(\frac{a_n-b_n}{\sigma_n-n}\right), \left(\frac{\hat{a}_n-\hat{b}_n}{\sigma_n-n}\right)\right\rangle_{\ell^2}.$$

One easily computes $\mathcal{M}'$, the dual space of $\mathcal{M}$ relative to the ℓ^2 inner product, to be the Hilbert space of sequences $(c_k)_{k\in\mathbb{Z}} \cup (d_k)_{k\in\mathbb{Z}}$ with $(c_k+d_k)\in\ell^2$ and $([\sigma_k-k][c_k-d_k])\in\ell^2$ with corresponding scalar product given by

$$(5.19)\quad \begin{aligned}\langle (c_k)\cup(d_k),(\hat{c}_k)\cup(\hat{d}_k)\rangle_{\mathcal{M}'} &= \langle([\sigma_k-k][c_k-d_k]),([\sigma_k-k][\hat{c}_k-\hat{d}_k])\rangle_{\ell^2}\\ &\quad+\langle(c_k+d_k),(\hat{c}_k+\hat{d}_k)\rangle_{\ell^2}.\end{aligned}$$

COROLLARY 5.3. *For any $T>0$ there exists $M>0$ such that for any $N\in\mathbb{N}$*

$$(5.20)\quad \int_0^T\left|\sum_{k=-N}^{N} a_k e^{ikt}+b_k e^{i\sigma_k t}\right|^2 dt \le M\|(a_k)\cup(b_k)\|^2_{\mathcal{M}'}.$$

Furthermore, there exists $m>0$ such that for any $N\in\mathbb{N}$

$$(5.21)\quad \int_0^{4\pi}\left|\sum_{k=-N}^{N} a_k e^{ikt}+b_k e^{i\sigma_k t}\right|^2 dt \ge m\|(a_k)\cup(b_k)\|^2_{\mathcal{M}'}.$$

Proof. Let $\mathcal{C}_T : L^2(0,T)\to\mathcal{M}$ by

$$\mathcal{C}_T f = (a_k)\cup(b_k):\ (5.15)\text{ holds.}$$

By Theorem 5.1 and the proof of Corollary 5.2, $\mathcal{C}_T$ is continuous for all $T>0$ and becomes onto when $T\ge 4\pi$. Equation (5.20) is simply the statement that $\mathcal{C}_T^*$ is continuous for $T>0$ while (5.21) states that $\mathcal{C}_T^*$ is bounded away from zero for $T\ge 4\pi$. □

5.4. Fourier description of observable space. We close this section with a description of the observable space in terms of nonharmonic Fourier series. In particular, we obtain a slight improvement of Theorem 4.1 for the case where (5.1) holds and also examine the dependence of the observability in terms of the frequency of the initial data.

PROPOSITION 5.4. *Assume* (5.1) *and let $y=(u,v,z,\dot{u},\dot{v},\dot{z})^t$ be a finite-energy solution of* (2.1) *with*

$$y^0=\sum_{k\in\mathbb{Z}\setminus\{0\}} a_k\varphi_{k\pi}+\sum_{\omega_k\in S_2} b_k\varphi_{\omega_k}.$$

Then for any $T>0$ there exists $C>0$ for which

$$(5.22)\quad \int_0^T |v_x(1,t)|^2dt \le C\left(b_{0^-}^2+b_{0^+}^2+\sum_{k\in\mathbb{Z}\setminus\{0\}}\delta_k^2(b_k-a_k)^2+(b_k+a_k)^2\right).$$

Furthermore, there exists $c>0$ for which

$$(5.23)\quad \int_0^4 |v_x(1,t)|^2dt \ge c\left(b_{0^-}^2+b_{0^+}^2+\sum_{k\in\mathbb{Z}\setminus\{0\}}\delta_k^2(b_k-a_k)^2+(b_k+a_k)^2\right).$$

In particular, Theorem 4.1 *and Proposition* 4.2 *remain true when* $T = 4$ *and* (5.1) *holds.*

Proof. For $h = (u, v, z, \dot{u}, \dot{v}, \dot{z})^t \in \mathcal{D}(\mathcal{A})$ define $b^* : \mathcal{D}(\mathcal{A}) \to \mathbb{R}$ by

$$b^* h = v_x(1).$$

Then b can be viewed as an element of $\mathcal{D}(\mathcal{A})'$ and we are in the setting of (5.9). Under the normalization of the eigenfunctions taken in (5.7) we have

$$b^* \varphi_w = 1 \quad \forall w \in S_1 \cup S_2.$$

It follows that for $T > 0$

$$\int_0^T |b^* y|^2 dt = \int_0^T \left| \sum_{k \in \mathbb{Z} \setminus \{0\}} a_k e^{ik\pi t} + \sum_{\omega_k \in S_2} b_k e^{i\omega_k t} \right|^2 dt. \tag{5.24}$$

A Riesz basis of exponentials $(e^{ir_k t})_{k \in \mathbb{Z}}$ is stable with respect to the perturbation $r_0 \to \tilde{r}_0$ provided $r_0 \neq r_k, k = \pm 1, \pm 2, \ldots$, (see [16]). Therefore Theorem 5.1 and its corollaries remain true when the exponential 1 is replaced by $\exp t\omega_{0-}$. Thus (5.22) and (5.23) follow from (5.24) and Corollary 5.3.

Define E as the completion of $\mathcal{H}$ with respect to the norm

$$\left\| \sum_{k \in \mathbb{Z} \setminus \{0\}} a_k \varphi_{k\pi} + \sum_{\omega_k \in S_2} b_k \varphi_{\omega_k} \right\|_E = \|(a_k) \cup (b_k)\|_{\mathcal{M}'}. \tag{5.25}$$

Corollary 5.3 and (5.22) show that the map $E \to \mathbb{R}$ defined by

$$y^0 \to \left(\int_0^T |b^* y|^2 dt \right)^{1/2}$$

defines a norm on E for any $T \geq 4$. By Proposition 4.2, an equivalent norm on E is given by the left side of (4.4). It thus follows that the inequality (4.4) remains valid when ((5.1) holds and) $T \geq 4$. Consequently Theorem 4.1 remains true when ((5.1) holds and) $T \geq 4$. □

Let us now give a characterization of the observable space E defined in (5.25) in terms of the eigenfunctions.

Let $p_0 = \phi_{\omega_{0-}}$, $q_0 = \phi_{\omega_{0+}}$ and for $k \in \mathbb{Z} \setminus \{0\}$ define

$$p_k = \frac{\varphi_{k\pi} + \varphi_{\omega_k}}{2}, \quad q_k = \frac{\varphi_{k\pi} - \varphi_{\omega_k}}{2\delta_k}.$$

LEMMA 5.5. $(p_k)_{k \in \mathbb{Z}} \cup (q_k)_{k \in \mathbb{Z}}$ *forms a Riesz basis for the space* E *defined by* (5.25). *In particular,*

$$E = \left\{ \sum_{k \in \mathbb{Z}} c_k p_k + d_k q_k : (c_k) \in \ell^2, (d_k) \in \ell^2 \right\}.$$

Proof. From (5.6) and (5.19), an equivalent representation for $\mathcal{M}'$ as it applies to our problem is

$$\mathcal{M}' = \left\{ (a_k) \cup (b_k) : b_{0+}^2 + b_{0-}^2 + \sum_{k \in \mathbb{Z} \setminus \{0\}} ((a_k - b_k)\delta_k)^2 + (a_k + b_k)^2 < \infty \right\}.$$

Thus the topology obtained by making $(p_k) \cup (q_k)$ form an orthonormal basis must also be equivalent to the natural one given by (5.25). □

We can now easily prove the following result, which describes the dependence of the observability upon the frequency.

PROPOSITION 5.6. *Assume* (5.1) *and let* $y = (u, v, z, \dot{u}, \dot{v}, \dot{z})^t$ *be a solution of* (2.1) *with*

$$y^0 = \sum_{k=-N}^{N} c_k p_k + d_k q_k.$$

Then there exists $C > 0$ *independent of* $N > 0$ *for which*

$$\left(\int_0^4 |v_x(1,t)|^2 dt \right)^{1/2} \geq \frac{C}{N} \|y^0\|_{\mathcal{H}}. \tag{5.26}$$

Furthermore, the bound is sharp in the sense that if $y^0 = q_N$ *then the inequality in* (5.26) *is reversible for some larger value of* C.

Proof. First note that $(p_k) \cup (q_k)$ form an orthogonal system on $\mathcal{H}$. In particular,

$$\|y^0\|^2_{\mathcal{H}} \leq C_1 \sum_{k=-N}^{N} |c_k|^2 + \frac{|d_k|^2}{\delta_k^2}, \tag{5.27}$$

where C_1 is independent of N. By Proposition 5.4 and Lemma 5.5 there exists $C_2 > 0$, independent of N for which

$$\int_0^4 |v_x(1,t)|^2 dt \geq C_2 \sum_{k=-N, k\neq 0}^{N} |c_k|^2 + |d_k|^2. \tag{5.28}$$

Let $\Delta_N = \min\{1, \delta_1^2, \delta_2^2, \ldots \delta_N^2\}$. Since

$$\sum_{k=-N}^{N} |c_k|^2 + |d_k|^2 \geq \Delta_N \sum_{k=-N}^{N} |c_k|^2 + \frac{|d_k|^2}{\delta_k^2}, \tag{5.29}$$

(5.26) follows from (5.6). To show the optimality, it is enough to see that all the inequalities are reversible when $y^0 = q_N$. For (5.27) this is obvious, for (5.28) the reversibility follows from Lemma 5.5, and for (5.29) we have equality for N sufficiently large by (5.6). Thus the reverse inequality of (5.26) holds for all N by making C sufficiently large. □

Remark 5.2. In the previous proof, the constants C_1 and C_2 can be shown to be independent of M for $M \in (1, \infty)$. Thus (5.29) shows that the constant C in the statement of Proposition 5.6 is proportional to $\Delta_N^{1/2}$, i.e., to M^{-1} as $M \to \infty$. Thus for solutions of a fixed energy level, the frequencies one can observe above a certain threshhold level vary inversely with the mass. As $M \to 0$, the numbers Δ_N in (5.29) approach 1 for all N, but not uniformly. Thus as $M \to 0$, for fixed N, we obtain the same observability constant as with the string without a point mass.

6. Some results on boundary stabilization. In this section we examine the problem of stabilization by velocity feedback at one or both ends. Thus we will be interested in the decay properties of solutions of the system

$$\tag{6.1}\begin{cases} \rho_1 u_{tt} = \sigma_1 u_{xx}, & x \in \Omega_1,\ t > 0, \\ \rho_2 v_{tt} = \sigma_2 v_{xx}, & x \in \Omega_2,\ t > 0, \\ M z_{tt}(t) + \sigma_1 u_x(0,t) - \sigma_2 v_x(0,t) = 0, & t > 0, \\ u(0,t) = v(0,t) = z(t), & t > 0, \\ u(x,0) = u^0(x),\ u_t(x,0) = u^1(x), & x \in \Omega_1, \\ v(x,0) = v^0(x),\ v_t(x,0) = v^1(x), & x \in \Omega_2, \\ z(0) = z^0,\ z_t(0) = z^1 \end{cases}$$

under the following two types of boundary conditions:

$$\tag{6.2}\begin{cases} u(-\ell_1, t) = 0, & t > 0, \\ \sigma_2 v_x(\ell_2, t) + \gamma v_t(\ell_2, t) = 0, & t > 0 \end{cases}$$

and

$$\tag{6.3}\begin{cases} \sigma_1 u_x(-\ell_1, t) - \gamma u_t(-\ell_1, t) = 0, & t > 0, \\ \sigma_2 v_x(\ell_2, t) + \gamma v_t(\ell_2, t) = 0, & t > 0, \end{cases}$$

where γ is positive.

In (6.2) we are introducing some damping on the system at the extreme-point $x = \ell_2$ and in (6.3) at both extremes $x = -\ell_1, \ell_2$.

The effect of the damping can be seen by noting that (formally) the energy of solutions of (6.1)–(6.2) and (6.1), (6.3) satisfies

$$\frac{dE_M}{dt}(t) = -\gamma |v_t(\ell_2, t)|^2$$

and

$$\frac{dE_M}{dt}(t) = -\gamma [\, |u_t(-\ell_1, t)|^2 + |v_t(\ell_2, t)|^2],$$

respectively.

Standard semigroup theory allows us to prove the following two facts:

(i) If $y^0 = (u^0, v^0, z^0, u^1, v^1, z^1)^t$ and

$$\tag{6.4}\begin{cases} y^0 \in H^1(\Omega_1) \times H^1(\Omega_2) \times \mathbb{R} \times L^2(\Omega_1) \times L^2(\Omega_2) \times \mathbb{R}, \\ u^0(0) = v^0(0) = z^0 \quad (u^0(-\ell_1) = 0 \text{ in case of (6.2) }), \end{cases}$$

then (6.1)–(6.2) and (6.1), (6.3) have a unique solution in the class

$$\tag{6.5}\begin{cases} (u, v, z) \in C([0, \infty); H^1(\Omega_1) \times H^1(\Omega_2) \times \mathbb{R}), \\ (u_t, v_t, z_t) \in C([0, \infty); L^2(\Omega_1) \times L^2(\Omega_2) \times \mathbb{R}) \end{cases}$$

and the following energy identities hold:

$$\tag{6.6}\begin{cases} \text{(i) } E_M(t_2) - E_M(t_1) = -\gamma \int_{t_1}^{t_2} |v_t(\ell_2, t)|^2 dt \quad \text{(for (6.1)–(6.2))}, \\ \text{(ii) } E_M(t_2) - E_M(t_1) = -\gamma \int_{t_1}^{t_2} [\, |u_t(-\ell_1, t)|^2 + |v_t(\ell_2, t)|^2] dt \quad \text{(for (6.1), (6.3))}. \end{cases}$$

(ii) If the initial data satisfy the following additional regularity and compatibility conditions:

$$\tag{6.7}\begin{cases} (u^0, v^0) \in H^2(\Omega_1) \times H^2(\Omega_2); \quad (u^1, v^1) \in H^1(\Omega_1) \times H^1(\Omega_2), \\ u^1(0) = v^1(0) = z^1, \\ \sigma_2 v_x^0(\ell_2) + \gamma v^1(\ell_2) = 0, \\ u^1(-\ell_1) = 0 \quad (\text{for (6.1)–(6.2)}), \\ \sigma_1 u_x^0(-\ell_1) - \gamma u^1(-\ell_1) = 0 \text{ (for (6.1), (6.3))}, \end{cases}$$

then the solutions have the following added regularity:

$$\tag{6.8}\begin{cases} (u, v) \in C([0, \infty); H^2(\Omega_1) \times H^2(\Omega_2)), \\ (u_t, v_t) \in C([0, \infty); H^1(\Omega_1) \times H^1(\Omega_2)), \\ z \in C^2([0, \infty); \mathbb{R}). \end{cases}$$

In what follows, solutions in the class (6.5) will be referred to as finite-energy solutions and those in (6.8) as smooth solutions.

There is an important difference between boundary conditions (6.2) and (6.3). For solutions of (6.1)–(6.2) the energy E_M is coercive and thus the only equilibrium configuration is the zero one. However in the system (6.1), (6.3) the energy is not coercive and for every real constant k, $(u, v, z) = (k, k, k)$ defines a solution.

In both systems we may expect the energy to decay to zero and this is the case. However, in system (6.1)–(6.3) additional work is required to show that every solution converges to an equilibrium.

There is another important difference between system (6.1)–(6.2) and (6.1), (6.3). In (6.1), (6.3) we will prove a uniform exponential decay of the energy; however, we will see that this does not hold for the system (6.1)–(6.2) since the system possesses a sequence of eigenfrequencies that approach the imaginary axis.

We have divided this section in three parts. In the first one we show that the energy of solutions of (6.1)–(6.2) converges to zero. In the second one we prove the uniform exponential decay of energy of solutions of (6.1), (6.3) and the fact that every trajectory converges toward an equilibrium. In the last part we study the spectral properties of system (6.1)–(6.2) and prove the nonuniform decay of energy.

6.1 Strong convergence to the equilibrium for (6.1)–(6.2). We have the following result.

THEOREM 6.1. *Every trajectory associated with a finite-energy solution of* (6.1)–(6.2) *converges strongly to zero in the finite-energy space* (6.4).

Proof. It is sufficient to prove that

$$\tag{6.9} E_M(t) \to 0 \text{ as } t \to \infty.$$

First observe that due to the density of the data satisfying (6.7) in the finite-energy space (6.4), the decreasing character of the energy (as in (6.6)) and the linearity of the system, if (6.9) holds for all smooth solutions then it also holds for all finite-energy solutions. Thus we will need to consider only smooth solutions.

Next we observe that $(u_t, v_t, z_t, u_{tt}, v_{tt}, z_{tt})$ is a solution (6.1)–(6.2) with the initial data $(u^1, v^1, z^1, \rho_1^{-1}\sigma_1 u_{xx}^0, \rho_2^{-1}\sigma_2 v_{xx}^0, M^{-1}(\sigma_2 v_x^0 - \sigma_1 u_x^0))$ which satisfies the compatibility condition $u^1(0) = v^1(0) = z^1$. Thus the trajectory $(u_t, v_t, z_t, u_{tt}, v_{tt}, z_{tt})$ is a finite energy solution and hence due to the fact that this energy is nonincreasing we deduce that

$$(u_t, v_t, z_t) \in L^\infty(0,\infty; H^1(\Omega_1) \times H^1(\Omega_2) \times \mathbb{R}),$$
$$(u_{tt}, v_{tt}, z_{tt}) \in L^\infty(0,\infty; L^2(\Omega_1) \times L^2(\Omega_2), \times\mathbb{R}),$$

and then by elliptic regularity that

$$(u,v) \in L^\infty(0,\infty; H^2(\Omega_1) \times H^2(\Omega_2)).$$

This shows that trajectories

$$\{(u(t), v(t), z(t), u_t(t), v_t(t), z_t(t))\}_{t\geq 0}$$

are relatively compact in the finite-energy space. Let ω be its ω-limit set with respect to the strong topology of the finite-energy space.

Since the energy E_M is a Lyapunov function, by LaSalle's invariance principle we deduce that ω is reduced to states for which the corresponding solution has constant energy for all $t > 0$. It is easy to see that the only solution satisfying this property is the zero solution. This shows that $\omega = \{0\}$ and concludes the proof. □

6.2. Uniform exponential decay for (6.1), (6.3). We have the following result.

THEOREM 6.2. *There exist $C > 0$ and $\omega_0 > 0$ such that*

$$E_M(t) \leq C E_M(0) e^{-\omega_0 t} \quad \forall t > 0 \tag{6.10}$$

holds for every finite-energy solution of (6.1), (6.3).

Moreover,

(6.11)
$$\begin{cases} \|(u(\cdot,t) - k)\|^2_{H^1(\Omega_1)} + \|(v(\cdot,t) - k)\|^2_{H^1(\Omega_2)} + |z(t) - k|^2 \leq C E_M(0) e^{-\omega_0 t} \\ \text{with } k \in \mathbb{R} \text{ such that } 2\gamma k = \int\limits_{-\ell_1}^{0} u^1 dx + \int\limits_{0}^{\ell_2} v^1 dx + M z^1 + \gamma u^0(-\ell_1) + \gamma v^0(\ell_2). \end{cases}$$

Proof. It is sufficient to prove the result for smooth solutions. Let us first prove (6.10).

We have

$$E_M(T) - E_M(0) = -\gamma \int\limits_0^T [\,|u_t(-\ell_1,t)|^2 + |v_t(\ell_2,t)|^2]dt \quad \forall T > 0. \tag{6.12}$$

Therefore it is sufficient to prove the existence of $C > 0$ and $T > 0$ such that

$$E_M(T) \leq C \int\limits_0^T [\,|u_t(-\ell_1,t)|^2 + |v_t(\ell_2,t)|^2]dt \tag{6.13}$$

for every smooth solution. Indeed from (6.12) and (6.13), (6.10) holds easily by using the semigroup property.

On the other hand, (6.13) can be proved the same way as we proved inequality (3.7) in Proposition 3.3.

Let us now prove (6.11). By differentiating the quantity

$$I(t) = \int\limits_{-\ell_1}^{0} \rho_1 u_t(x,t)dx + \int\limits_{0}^{\ell_2} \rho_2 v_t(x,t)dx + M z_t(t) + \gamma u(-\ell_1,t) + \gamma v(\ell_2,t) \tag{6.14}$$

and using (6.1), (6.3), one sees that $I(t)$ remains constant in time.

Given a solution of (6.1), (6.3) we decompose it as follows:

$$(u, v, z, u_t, v_t, z_t) = (\tilde{u}, \tilde{v}, \tilde{z}, u_t, v_t, z_t) + (\bar{u}, \bar{v}, \bar{z}, 0, 0, 0), \tag{6.15}$$

where

$$\bar{u} = \bar{v} = \bar{z} = k \quad \text{and} \quad \tilde{u} = u - k, \quad \tilde{v} = v - k, \quad \tilde{z} = z - k \tag{6.16}$$

with k a constant such that

$$2\gamma k = I(0). \tag{6.17}$$

Clearly $(\bar{u}, \bar{v}, \bar{z})$ is a stationary solution of (6.1), (6.3). On the other hand, the quantity $I(t)$ associated with $(\tilde{u}, \tilde{v}, \tilde{z}, u_t, v_t, z_t)$ is identically zero. It is easy to check that $E_M(\cdot)$ is coercive over the subspace of the finite-energy space in which $I = 0$. From (6.10) we then deduce that the component $(\tilde{u}, \tilde{v}, \tilde{z}, u_t, v_t, z_t)$ of our solutions decays exponentially to zero in the energy space, i.e., (6.11) holds. □

6.3. Nonuniform energy decay. Let $\mathcal{A}$ be the differential operator in (2.1), however, on the space

$$\widetilde{\mathcal{H}} = \widetilde{W_1} \times W_0, \tag{6.18}$$

where

$$\widetilde{W_1} = \{(u, v, z) \in H^1(\Omega_1) \times H^1(\Omega_2) \times \mathbb{R} : u(-\ell_1) = 0, \ u(0) = v(0) = z\} \tag{6.19}$$

with domain

(6.20)

$$\widetilde{\mathcal{D}}(\mathcal{A}) = \{y \in \widetilde{\mathcal{H}} : u \in H^2(\Omega_1), v \in H^2(\Omega_2), \ (\dot{u}, \dot{v}, \dot{z}) \in \widetilde{W_1} : \sigma_2 v_x(\ell_2) + \gamma \dot{v}(\ell_2) = 0\}.$$

We have the following result, which describes the amount of damping in (6.1)–(6.2), in terms of $\sigma(\mathcal{A})$.

THEOREM 6.3. *Let $\mathcal{A}$ be defined by (6.18)–(6.20), with $\gamma > 0$. Then there exists a sequence $(s_k)_{k\in\mathbb{Z}} \subset \sigma(\mathcal{A})$ for which*

(i) $\left| \sqrt{\dfrac{\sigma_1}{\rho_1}} \dfrac{k\pi}{\ell_1} - \operatorname{Im} s_k \right| \to 0$ *as* $|k| \to \infty$

(ii) *there exists c_1, $c_2 > 0$ for which* $-\dfrac{c_1}{k^2} > \operatorname{Re} s_k > -\dfrac{c_2}{|k|} \quad \forall k \in \mathbb{Z}.$

If $\lambda \in \sigma(\mathcal{A}) \setminus \{(s_k)\}$ then there exists $c_3 > 0$ for which $\operatorname{Re} \lambda < -c_3$.

Remark 6.1. The second inequality of (ii) implies that the energy of solutions does not decay uniformly to zero in bounded sets of $\mathcal{H}$. But as a consequence of a general result due to W. Littman and L. Markus [8], in view of the first inequality of (ii), one can obtain uniform polynomial decay rates of energy for solutions with initial data belonging to a bounded set of the domain of some power of $\mathcal{A}$.

Proof. We assume without loss of generality (by rescaling) that $\ell_1 = \ell_2 = 1$.

We seek nontrivial solutions of

$$\mathcal{A}y = i\omega y. \tag{6.21}$$

Let

$$\alpha = \sqrt{\frac{\sigma_1}{\rho_1}}, \quad \beta = \sqrt{\frac{\sigma_2}{\rho_2}}.$$

Using (6.21) and the boundary conditions, one can easily calculate

$$\text{(6.22)} \qquad \begin{cases} u = a\sin\dfrac{\omega}{\alpha}(x+1), \\ v = b\psi_\omega(x); \quad \psi_\omega(x) = \dfrac{\gamma}{\rho_2}\sin\dfrac{\omega}{\beta}(x-1) + i\beta\cos\dfrac{\omega}{\beta}(x-1), \\ z = \dfrac{a\sigma_1}{M\alpha\omega}\cos\dfrac{\omega}{\alpha} - \dfrac{b\sigma_2}{M\omega^2}\dfrac{d\psi_\omega}{dx}(0), \end{cases}$$

where a, b, ω are to be determined. The boundary conditions in (6.19) imply that

$$\text{(6.23)} \qquad a\sin\frac{\omega}{\alpha} = b\psi_\omega(0) = z.$$

The possibility that $z = 0$ leads to only trivial solutions. Thus we may set $a = \psi_\omega(0)$, $b = \sin\omega/\alpha$, and obtain from (6.23)

$$\text{(6.24)} \qquad M\omega = \frac{\sigma_1\cos\omega/\alpha}{\alpha\sin\omega/\alpha} - h(\omega); \quad h(\omega) = \frac{\sigma_2}{\omega\psi_\omega(0)}\frac{d\psi_\omega}{dx}(0).$$

Now there are three separate cases to consider:

$$\gamma < \sqrt{\rho_2\sigma_2}, \quad \gamma = \sqrt{\rho_2\sigma_2}, \quad \gamma > \sqrt{\rho_2\sigma_2}.$$

In each case the result is the same and the idea is the same. Thus let us consider only the first case, $\gamma < \sqrt{\rho_2\sigma_2}$.

Let $\mathbb{K}$ denote the roots of $\psi_\omega(0) = 0$. A simple calculation shows

$$\text{(6.25)} \qquad \mathbb{K} = \left\{\beta\left(k\pi + \frac{\pi}{2} + i\tanh^{-1}\frac{\gamma}{\beta\rho_2}\right)\right\}_{k\in\mathbb{Z}}.$$

For $S \subset \mathbb{C}$, define

$$N_\delta(S) = \{\lambda \in \mathbb{C} : \operatorname{dist}(\lambda, S) < \delta\}.$$

From (6.25) it follows that $N_{\delta_0}(\mathbb{K}) \cap N_{\delta_0}(\alpha\pi\mathbb{Z}) = \emptyset$ for δ_0 small enough. Furthermore, since $h(\omega)$ and $\cot\omega/\alpha$ are periodic and analytic outside $N_{\delta_0}(\mathbb{K})$ and, respectively, $N_{\delta_0}(\alpha\pi\mathbb{Z})$, $h(\omega)$ is uniformly bounded in $N_{\delta_0}(\alpha\pi\mathbb{Z})$ and $\cot\omega/\alpha$ is uniformly bounded in $N_{\delta_0}(\mathbb{K})$. Since the moduli of both sides of (6.24) are the same, it follows that for any $\varepsilon > 0$, only finitely many eigenvalues lie outside $N_\varepsilon(\alpha\pi\mathbb{Z}) \cup N_\varepsilon(\mathbb{K})$. Rouché's theorem can be used to show that for any $\varepsilon > 0$ if $|\omega|$ is sufficiently large, there is a unique root of (6.24) in each component of $N_\varepsilon(\alpha\pi\mathbb{Z})$ and $N_\varepsilon(\mathbb{K})$. (See [8] for an example of this calculation.)

A simple calculation shows that $\operatorname{Im} h(\omega) < -a < 0$ for all $\omega \in \mathbb{R}$. Since $h(\omega)$ is periodic and analytic in a neighborhood of the real axis, we may assume there exist positive numbers a_0 and l for which

$$\text{(6.26)} \qquad \operatorname{Im}\, h(\omega) < -a_0 \quad \forall\omega \in \mathbb{C} : |\operatorname{Im}\,\omega| \le l.$$

Let $\alpha k\pi + s + i\tau$ denote a root of (6.24) in $N_\varepsilon(\alpha\pi\mathbb{Z})$. By setting the imaginary parts of (6.24) equal and then using (6.26), we find that for ε sufficiently small (and hence $|k|$ sufficiently large) we have

$$\text{(6.27)} \qquad \left(\tau - \frac{\sigma_1}{2a_0}\right)^2 + s^2 < \frac{\sigma_1^2}{4a_0^2}.$$

Thus for $|\omega|$ sufficiently large, these roots lie in disks of radii $\sigma_1/2a_0$ tangent to the real axis at the points in $\{\alpha\pi\mathbb{Z}\}$.

Due to the upperbound on $|h(\omega)|$ in $N_{\delta_0}(\alpha\pi\mathbb{Z})$, since the moduli of both sides of (6.24) are the same, for $|\omega|$ sufficiently large we must have

$$\frac{M}{2}|\omega| < \left|\frac{\sigma_1}{\alpha}\cot\frac{\omega}{\alpha}\right| < 2M|\omega|. \tag{6.28}$$

Since $\cot\omega/\alpha$ is periodic and has a first-order pole at 0, (6.28) implies there exists c_1, $c_2 > 0$, and $k \in \mathbb{Z}$ for which

$$\frac{c_1}{|\omega|} < |\omega - \alpha k\pi| < \frac{c_2}{|\omega|}, \quad |\omega| \text{ sufficiently large.} \tag{6.29}$$

Intersecting the sets described by (6.27) and (6.29) and recalling (6.21), we obtain (i) and (ii). For any other roots of (6.24), either $\omega \in N_{\delta_0}(\mathbb{K})$ or ω is one of the finitely many outside of both $N_{\delta_0}(\mathbb{K})$ and $N_\varepsilon(\alpha k\mathbb{Z})$. In either case, by (6.25) and the dissipativity of $\mathcal{A}$ we have Im $\omega >$ const > 0. Thus Re $\lambda =$ Re $i\omega < -c_3 < 0$. ||

7. Extensions. In this final section we discuss two important generalizations of the results in the previous sections. First we consider a string with n point masses and describe the manner in which the previous results extend to this case. Second we analyze again the one-mass problem, but this time with a string having spatially varying coefficients.

7.1. n masses. All the regularity results carry over to strings with n masses due to the local nature of these results. In particular Proposition 2.5 and the remark that follows it imply that the degree of regularity of a travelling wave solution increases exactly one order as it crosses each mass. (Of course, part of the wave is *reflected* at each mass point with no increase in regularity.) Furthermore, the method used to prove controllability (Theorem 4.1) relies only upon the regularity result (Proposition 2.7) and the use of characteristics (as in construction of the energy functions e_1 and e_2 in (3.7) and (3.9)) and hence applies equally well to strings with n masses. As such, analogous controllability and observability results hold for strings with n masses. For example, if we consider the problem of Dirichlet control of the rightmost end of the string, which has $n - 1$ masses (and n intervals), the control space one obtains is the same on the first (farthest right) interval as in Theorem 4.1. For each successive interval the regularity increases one order, i.e., on the kth interval I_k (from the right), the position of the string (respectively, velocity) is in $H^{k-1}(I_k)$ (respectively, $H^{k-2}(I_k)$). In addition, boundary conditions and compatibility conditions up to the proper order for that interval must be imposed to obtain a well-posed system. A detailed examination of this problem is beyond the scope of this paper, however.

7.2. Variable coefficients. All the results of the previous sections remain valid for strings with spatially variable coefficients. To see this, let us consider the following system:

$$\left\{\begin{array}{lll} \rho(x)w_{tt}(x,t) - \frac{\partial}{\partial x}(\sigma(x)w_x(x,t)) = 0, & x \in \Omega, & t > 0, \\ Mz_{tt}(t) = \sigma(0^+)w_x(0^+,t) - \sigma(0^-)w_x(0^-,t), & t > 0, & \\ w(0^-,t) = w(0^+,t) = z(t), & t > 0, & \\ w(-1,t) = 0, \quad w(1,t) = q(t), & t > 0, & \\ w(x,0) = w^0(x), \quad w_t(x,0) = w^1(x), & x \in \Omega, & \\ z(0) = z^0, \quad z_t(0) = z^1, & & \end{array}\right. \tag{7.1}$$

where $\Omega = (-1,0)\cup(0,1)$ and the notation 0^+, 0^- refers to right- and left-hand limits at 0.

We have the following result.

THEOREM 7.1. *Suppose that $\rho, \sigma \in H^2((-1,0)\cup(0,1))$ and that both functions are bounded below by a positive constant. Then, if*

$$T > 2\int_{-1}^{1}\left(\frac{\rho(\tau)}{\sigma(\tau)}\right)^{1/2} d\tau,$$

the conclusion of Theorem 4.1 *holds.*

Sketch of proof. Let us introduce the following change of variables:

$$\begin{aligned}
&s = \int_0^x \rho(\tau)^{1/2}\sigma(\tau)^{-1/2} d\tau,\\
&\tilde{\Omega} = (-\ell_1, 0)\cup(0,\ell_2); \quad \ell_1 = -s(-1), \quad \ell_2 = s(1),\\
&b(s) = \rho(x)^{-1}\frac{d}{dx}\sqrt{\rho(x)\sigma(x)}\ |_{x=x(s)},\\
&\tilde{w}(s,t) = \exp\left(\int_0^s b(r)/2\, dr\right) w(x(s),t),\\
&a = \sqrt{\rho(0^+)\sigma(0^+)}\frac{b(0^+)}{2} - \sqrt{\rho(0^-)\sigma(0^-)}\frac{b(0^-)}{2}.
\end{aligned}$$

Then (7.1), in the absence of control ($q = 0$), becomes

$$(7.2)\quad \begin{cases}
\tilde{w}_{tt} - \left[\tilde{w}_{ss} - \left(\frac{b'(s)}{2} + \frac{b^2(s)}{4}\right)\tilde{w}\right] = 0, & s \in \tilde{\Omega},\ t > 0,\\
Mz_{tt}(t) + az(t) = \sqrt{\rho(0^+)\sigma(0^+)}\tilde{w}_s(0^+,t) & \\
\qquad -\sqrt{\rho(0^-)\sigma(0^-)}\tilde{w}_s(0^-,t), & t > 0,\\
\tilde{w}(0^-,t) = \tilde{w}(0^+,t) = z(t), & t > 0,\\
\tilde{w}(-\ell_1,t) = 0 = \tilde{w}(\ell_2,t), & t > 0,\\
\tilde{w}(s,0) = \exp\int_0^s b(r)/2\, dr w^0(x(s)), & s \in \tilde{\Omega},\\
\tilde{w}_t(s,0) = \exp\int_0^s b(r)/2\, dr w^1(x(s)), & s \in \tilde{\Omega},\\
z(0) = z^0, \quad z_t(0) = z^1. &
\end{cases}$$

Therefore, it is sufficient to prove the controllability of this system at time $T > 2(\ell_1 + \ell_2)$ with a control at the right end $x = \ell_2$.

First let us examine the system (7.2), but without the potential term. This is equivalent to

$$(7.3)\quad \begin{cases}
w_{tt} - w_{xx} = 0, & x \in \tilde{\Omega},\ t > 0,\\
Mz_{tt}(t) + az(t) = \gamma w_x(0^+,t) - w_x(0^-,t), & t > 0,\\
w(0^-,t) = w(0^+,t) = z(t), & t > 0,\\
w(-\ell_1,t) = 0 = w(\ell_2,t), & t > 0,\\
w(x,0) = w^0(x), \quad w_t(x,0) = w^1(x), & x \in \tilde{\Omega},\\
z(0) = z^0, \quad z_t(0) = z^1. &
\end{cases}$$

We assume $M, \gamma, \ell_1, \ell_2$ to be arbitrary positive numbers, and a is assumed only to be real. One can check that the proof of Proposition 2.5 remains valid for the system (7.3); when $Mz_{tt} + az$ replaces Mz_{tt} in (2.14) one still obtains (2.18). It hence follows that all regularity results, in particular Proposition 2.7, remain valid for (7.3).

Finally, the assumptions on the coefficients imply the potential $(\frac{b'(s)}{2} + \frac{b^2(s)}{4})$ remains in $L^2(\widetilde{\Omega})$. We can thus write the solution of (7.2) in terms of an integral equation given by the variation of constants formula for (7.3) and for small enough t determine a fixed point having the same regularity as the solution of (7.3) with initial data as in the hypothesis of Proposition 2.7. The remaining parts of the proof of Theorem 4.1 only rely upon the use of characteristics, and hence can be done directly for the variable coefficient problem (7.1). □

REFERENCES

[1] G. Chen, M. Delfour, A. M. Krall, and G. Payre, *Modeling, stabilization and control of serially connected beams*, SIAM J. Control Optim., 25 (1987), pp. 526–546.

[2] A. Haraux, *Séries lacunaires et contrôle semi-interne des vibrations d'une plaque rectangulaire*, J. Math. Pures Appl., 68 (1989), pp. 457–465.

[3] A. E. Ingham, *Some trigonometrical inequalities with applications to the theory of series*, Math. Z., 41 (1936), pp. 367–379.

[4] J. Lagnese, G. Leugering, and E. J. P. G. Schmidt, *On the analysis and control of hyperbolic systems associated with vibrating networks*, Proc. Roy. Soc. Edinburgh Sect. A, 124 (1994), pp. 77–104.

[5] E.B. Lee and Y.C. You, *Stabilization of a hybrid (string/point mass) system*, in Proc. 5th Int. Conf. on Systems Engineering, Dayton, OH, 1987.

[6] J.L. Lions, *Exact controllability, stabilization and perturbations for distributed systems*, SIAM Rev., 30 (1988), pp. 1–68.

[7] ———, *Contrôlabilité Exacte, Perturbations et Stabilisation de Systèmes Distribués*, Tome I and II, Masson, Paris, 1988.

[8] W. Littman and L. Markus, *Some recent results on control and stabilization of flexible structures*, in Stabilization of Flexible Structures, A.V. Balakrishnan and J.P. Zolesio eds., Optimization Software Inc., Los Angeles, 1988, pp. 151–161.

[9] W. Littman and L. Markus, *Exact boundary controllability of a hybrid system of elasticity*, Arch. Rational Mech. Anal., 103 (1988), pp. 193–236.

[10] ———, *Stabilization of a hybrid system of elasticity*, Ann. Mat. Pura Appl. (IV), CLII (1988), pp. 281–330.

[11] D. L. Russell, *Controllability and stabilizability theory for partial differential equations. Recent progress and open questions*, SIAM Rev., 20 (1978), pp. 639–739.

[12] E. J. P. G. Schmidt, *On the modelling and exact controllability of networks of vibrating strings*, SIAM J. Control Optim., 30 (1992), pp. 229–245.

[13] ———, *Controllability of string networks*, SIAM Conference on Control, Minneapolis, MN, September 1992.

[14] E. J. P. G. Schmidt and Wei Ming, *On the modelling and analysis of networks of vibrating strings and masses*, Report 91-13, Dept. Math. and Stats., McGill Univ., Montreal, 1991.

[15] D. Ullrich, *Divided differences and systems of nonharmonic Fourier series*, Proc. Amer. Math. Soc., 80 (1980), pp. 47–57.

[16] R. M. Young, *An Introduction to Nonharmonic Fourier Series*, Academic Press, New York, 1980.

[17] E. Zuazua, *Contrôlabilité exacte d'une équation des ondes surlinéaire à une dimension d'espace*, C.R. Acad. Sci. Paris, 311 (1990), pp. 285–290.

[18] ———, *Exact controllability for semilinear wave equations in one space dimension*, Ann. Inst. H. Poincaré, Anal. Non Lineáire, 10 (1993), pp. 109–129.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 33, No. 5, pp. 1392 1410, September 1995

© 1995 Society for Industrial and Applied Mathematics
005

OPTIMAL STOPPING OF A DISCRETE MARKOV PROCESS BY TWO DECISION MAKERS*

KRZYSZTOF SZAJOWSKI†

Abstract. In this paper a problem of optimal stopping of the discrete time Markov process by two decision makers in a competitive situation is considered. The zero-sum game approach is adopted. The gain function depends on the states chosen by both decision makers. The random assignment mechanism is used when both want to accept the realization of the Markov process at the same moment. The construction of the value function and the optimal strategies for the players are given. Examples related to the generalization of the best choice problem are solved.

Key words. optimal stopping problem, game variant, Markov process, random priority, secretary problem, zero-sum two-person game

AMS subject classifications. 60G40, 90D15

1. Introduction.

This paper deals with a class of the following two-person decision problems. Let $(X_n, \mathcal{F}_n, \mathbf{P}_x)_{n=0}^N$ be a homogeneous Markov process defined on probability space $(\Omega, \mathcal{F}, \mathbf{P})$ with fixed state space $(\mathbb{E}, \mathcal{B})$. The decision makers, henceforth called Player 1 and Player 2, observe the process sequentially. They want to accept the most profitable state of the process from their point of view.

We adopt the zero-sum game model for the problem. In view of this approach, the preferences of each player are described by gain function $f : \mathbb{E} \times \mathbb{E} \rightarrow \Re$. The function depends on the state chosen by both players. It would be natural to consider the stopping times with respect to $(\mathcal{F}_n)_{n=0}^N$ as the strategies of the player if the players could obtain the state which they want. Since there is only one random sequence $(X_n)_{n=0}^N$ on a trial, at each moment n only one player can obtain realization x_n of X_n. The problem of assigning an object to the players when both want to accept the same one at the same moment is solved by adopting the random mechanism; i.e., a lottery chooses the player who benefits. The player chosen by the lottery obtains realization x_n, and the player thus deprived of the acceptance of x_n at $n < N$ may select any later realization. The realization can be accepted only when it appears. No recall is allowed. We can think about the decision process as an investigation of objects with characteristics described by the Markov process. Both players together can accept at most two objects.

The above-described decision model is a generalization of the problems considered by the author in [20] and by Radzik and the author in [13]. The related questions, when Player 1 has permanent priority, have been considered by many authors in the zero-sum game or the non–zero-sum game setting. One can mention, for example, the recent papers of Ano [1], Enns and Ferenstein [4], Ferenstein [5], and Sakaguchi [16]. Many papers on the subject were inspired by the secretary problem (see T. S. Ferguson [6] for an extensive bibliography on the problem). Sakaguchi [16] considered the non–zero-sum two-person game related to the full-information best-choice problem with random priority. The review of the related models can be found in [20].

In this paper a formal model of the random priority is derived. The lottery is taken into account in the sets of the strategies of the players. The most interesting

* Received by the editors April 2, 1993; accepted for publication (in revised form) February 18, 1994. This research was supported in part by Komitet Badan Naukowych (Polish Research Committee) grant 2 1163 91 01.

† Institute of Mathematics, Technical University of Wrocław, Wybrzeże Wyspiańskiego 27, PL-50-370 Wrocław, Poland (szajow@im.pwr.wroc.pl).

question concerns the influence of the level of priority on the value of the problem or the probability of obtaining the required state of the process (or, in other words, the required object). The tip of the problem is shown by the examples. The random-priority game approach to the generalized secretary problem is considered. At first, random priority is added to the problem of choosing the best or the second best (BOS) but a better one than that of an applicant who is chosen by the opponent (the game considered in [20]). The dependence of the value, the strategies, and the probability of obtaining the required object by Player 1 on the priority is investigated. The second example tries to answer the question of what is the relation between the priority level and the aims of the players. The simplest problem with asymmetric aims of the players is considered. The first player's aim is to choose the best applicant (BA), and the second player wants to accept the BOS but a better one than the opponent. The numerical solution provides that the game is fair when Player 1 has priority $p \cong 0.7579$ (in the limiting case when $N \to \infty$).

The organization of the paper is as follows. In §2, the formal model is formulated in a rigorous way, and the general solution of the problem is given. The ideas used by Yasuda [21] and in [20] are employed. In §§3 and 4, the above-mentioned examples are solved. Section 5 contains some final remarks. Among other things, the probabilities of the success in the considered games are given.

2. Random priority and stopping the Markov process. Let a homogeneous Markov chain $(X_n, \mathcal{F}_n, \mathbf{P}_x)_{n=0}^N$ be defined on a probability space $(\Omega, \mathcal{F}, \mathbf{P})$ with a state space $(I\!\!E, \mathcal{B})$, and let $f : I\!\!E \times I\!\!E \to \Re$ be a $\mathcal{B} \times \mathcal{B}$ real-valued measurable function. Horizon N is finite. The players observe the Markov chain, and they try to accept the "best realization" according to function f and a possible selection of another player. Each realization x_n of X_n can be accepted by only one player, and each player can accept at most one realization. If the players have not accepted previous realizations, they evaluate the state of the Markov chain at instant n and have two options: either to accept the observed state of the process at moment n or to reject it. If both players want to accept the same realization, the following random priority mechanism is applied. Let $\xi_1, \xi_2, \dots, \xi_N$ be a sequence of independently and identically distributed random variables with the uniform distribution on $[0, 1]$, and let $\alpha = (\alpha_1, \alpha_2, \dots, \alpha_N)$ be a given vector of real numbers, $\alpha_i \in [0, 1]$. When both players want to accept realization x_n of X_n, then Player 1 obtains x_n if $\xi_n \leq \alpha_n$; otherwise Player 2 benefits. If Player 1 rejects the applicant, then Player 2 turns to exercise one of his options, which also consists of accepting the observed state of the Markov chain or rejecting it. If one of the players accepts realization x_n of X_n, then the other one is informed about it and he continues to play alone. If, in the above decision process, Player 1 and Player 2 have accepted states x and y, respectively, then Player 2 pays $f(x, y)$ to Player 1. When only Player 1 (Player 2) accepts state x (y), then Player 1 obtains $f_1(x) = \sup_{y \in I\!\!E} f(x, y)$ ($f_2(y) = \inf_{x \in I\!\!E} f(x, y)$) by assumption. If both players finish the decision process without any accepted state, then they gain 0.

Let $\mathcal{S}^N$ be the aggregation of Markov times with respect to $(\mathcal{F}_n)_{n=0}^N$. We admit that $\mathbf{P}_x(\tau \leq N) < 1$ for some $\tau \in \mathcal{S}^N$ (i.e., there is a positive probability that the Markov chain will not be stopped). The elements of $\mathcal{S}^N$ are possible strategies for the players with the restriction that Player 2 cannot stop at the same moment as Player 1. If the players declare willingness to accept the same object, the random device decides which player is endowed. Let us formalize these considerations. Denote $\mathcal{S}_k^N = \{\tau \in \mathcal{S}^N : \tau \geq k\}$. Let Λ_k^N and M_k^N be copies of $\mathcal{S}_k^N$ ($\mathcal{S}^N = \mathcal{S}_0^N$). One can define the set of strategies $\tilde{\Lambda}^N = \{(\lambda, \{\sigma_n^1\}) : \lambda \in \Lambda^N, \sigma_n^1 \in \Lambda_{n+1}^N \text{ for every } n\}$ and $\tilde{M}^N = \{(\mu, \{\sigma_n^2\}) : \mu \in M^N, \sigma_n^2 \in M_{n+1}^N \text{ for every } n\}$ for Players 1 and 2, respectively. Denote $\tilde{\mathcal{F}}_n = \sigma(\mathcal{F}_n, \xi_1, \xi_2, \ldots, \xi_n)$ and let $\tilde{\mathcal{S}}^N$ be the set of stopping times with respect to $(\tilde{\mathcal{F}}_n)_{n=0}^N$. Define $\tau_1 = \lambda \mathbb{I}_{\{\lambda<\mu\}} + (\lambda \mathbb{I}_{\{\xi_\lambda \le \alpha_\lambda\}} + \sigma_\mu^1 \mathbb{I}_{\{\xi_\lambda > \alpha_\lambda\}}) \mathbb{I}_{\{\lambda=\mu\}} + \sigma_\mu^1 \mathbb{I}_{\{\lambda>\mu\}}$ and $\tau_2 = \mu \mathbb{I}_{\{\lambda>\mu\}} + (\mu \mathbb{I}_{\{\xi_\mu > \alpha_\mu\}} + \sigma_\lambda^2 \mathbb{I}_{\{\xi_\mu \le \alpha_\mu\}}) \mathbb{I}_{\{\lambda=\mu\}} + \sigma_\lambda^2 \mathbb{I}_{\{\lambda \le \mu\}}$.

LEMMA 2.1. *Random variables τ_1 and τ_2 are Markov times with respect to $(\tilde{\mathcal{F}}_n)_{n=0}^N$ and $\tau_1 \neq \tau_2$.*

Proof. We have

$$\{\tau_1 = n\} = \{\lambda = n, \mu \ge n\} \cup \bigcup_{i=0}^{n-1} \{\mu = i, \sigma_i^1 = n\}$$
$$\cup \{\mu = n, \lambda = n, \xi_n \le \alpha_n\} \in \tilde{\mathcal{F}}_n \text{ for every } n$$

and

$$\{\tau_2 = n\} = \{\mu = n, \lambda > n\} \cup \bigcup_{i=0}^{n} \{\lambda = i, \sigma_i^2 = n\}$$
$$\cup \{\mu = n, \lambda = n, \xi_n \le \alpha_n\} \in \tilde{\mathcal{F}}_n \text{ for every } n;$$

hence τ_1 and τ_2 are Markov times and $\tau_1 \neq \tau_2$. □

Let $E_x f_1^+(X_n) < \infty$ and $E_x f_2^-(X_m) < \infty$ for $n, m = 0, 1, \ldots, N$ and $x \in \mathbb{E}$. Let $s \in \tilde{\Lambda}^N$ and $t \in \tilde{M}^N$. Define $\tilde{R}(x, s, t) = E_x f(X_{\tau_1}, X_{\tau_2})$ as the expected gain of Player 1. In this way the normal form of the game $(\tilde{\Lambda}^N, \tilde{M}^N, \tilde{R}(x, s, t))$ is defined. This game is denoted by $\mathcal{G}$. The game $\mathcal{G}$ is a model of the considered bilateral stopping problem for the Markov process.

DEFINITION 2.2. *Pair (s^*, t^*), $s^* \in \tilde{\Lambda}^N$, $t^* \in \tilde{M}^N$, is an equilibrium point in the game $\mathcal{G}$ if for every $x \in \mathbb{E}$, $s \in \tilde{\Lambda}^N$, and $t \in \tilde{M}^N$ we have*

$$\tilde{R}(x, s, t^*) \le \tilde{R}(x, s^*, t^*) \le \tilde{R}(x, s^*, t).$$

The aim is to construct the equilibrium pair (s^*, t^*). The following approach is proposed. When one of the players accepts realization x_n at moment n, the second player will try to maximize his gain without any disturbance from another player. It means that if there is no acceptance of states until moment n, the players must take into account the potential danger from a future decision of the opponent before accepting or rejecting realization x_n of X_n. To this end, they consider the following auxiliary game $\mathcal{G}_a$.

Define $s_0(x, y) = S_0(x, y) = f(x, y)$ and

$$s_n(x, y) = \inf_{\tau \in \mathcal{S}^n} E_y f(x, X_\tau),$$
$$S_n(x, y) = \sup_{\tau \in \mathcal{S}^n} E_x f(X_\tau, y)$$

for all $x, y \in \mathbb{E}$, $n = 1, 2, \ldots, N$. By the theory of optimal stopping for the Markov processes [17], the function $s_n(x, y)$ $(S_n(x, y))$ can be constructed by the recursive procedure as $s_n(x, y) = Q_{\min}^n f(x, y)$ $(S_n(x, y) = Q_{\max}^n f(x, y))$, where $Q_{\min} f(x, y) = f(x, y) \wedge T_2 f(x, y)$ $(Q_{\max} f(x, y) = f(x, y) \vee T_1 f(x, y))$ and $T_2 f(x, y) = E_y f(x, X_1)$ $(T_1 f(x, y) = E_x f(x, y))$. ($\wedge$ and $\vee$ denote minimum and maximum, respectively.) Operations $\wedge$ and T_2 ($\vee$ and T_1) preserve measurability. This can be proved in a standard way. Hence $s_n(x,y)$ $(S_n(x,y))$ are $\mathcal{B}\otimes\mathcal{B}$ measurable (cf. [3]). If Player 1 is the first to accept x at moment n, then his expected gain is

$$h(n,x) = E_x s_{N-n-1}(x, X_1) \tag{1}$$

for $n = 0, 1, \ldots, N-1$ and $h(N,x) = f_1(x)$. When Player 2 is the first, then the expected gain of Player 1 is

$$H(n,x) = E_x S_{N-n-1}(X_1, x) \tag{2}$$

for $n = 0, 1, \ldots, N-1$ and $H(N,x) = f_2(x)$. Functions $h(n,x)$ and $H(n,x)$ are well defined. They are $\mathcal{B}$-measurable of the second variable; $h(n, X_1)$ and $H(n, X_1)$ are integrable with respect to $\mathbf{P}_x$. Let Λ^N and M^N be sets of strategies in $\mathcal{G}_a$ for Player 1 and Player 2, respectively. For $\lambda \in \Lambda^N$ and $\mu \in M^N$, define the payoff function

$$r(\lambda,\mu) = \begin{cases} h(\lambda, X_\lambda)(\mathbb{I}_{\{\lambda<\mu\}} + \mathbb{I}_{\{\lambda=\mu,\xi_\lambda\le\alpha_\lambda\}}) & \\ \quad +H(\mu, X_\mu)(\mathbb{I}_{\{\lambda>\mu\}} + \mathbb{I}_{\{\lambda=\mu,\xi_\mu>\alpha_\mu\}}) & \text{if } \lambda \le N \text{ or } \mu \le N, \\ 0 & \text{otherwise,} \end{cases} \tag{3}$$

where $\mathbb{I}_A$ is a characteristic function of set A. As a solution of the game we search for equilibrium pair (λ^*, μ^*) such that

$$R(x, \lambda, \mu^*) \le R(x, \lambda^*, \mu^*) \le R(x, \lambda^*, \mu) \qquad \text{for all } x \in \mathbb{E}, \tag{4}$$

where $R(x,\lambda,\mu) = E_x r(\lambda,\mu)$. By (3) we can observe that $\mathcal{G}_a$ with the sets of strategies Λ^N and M^N is equivalent to Neveu's stopping problem [12] considered by Yasuda [21] if the sets of strategies are extended to the set of stopping times not greater than $N+1$ and the payoff function is (3). Because the Markov process is observed here, one can define a sequence $v_n(x)$, $n = 0, 1, \ldots, N+1$, on $\mathbb{E}$ by setting $v_{N+1}(x) = 0$ and

$$v_n(x) = \operatorname{val}\begin{bmatrix} h(n,x)\alpha_n + (1-\alpha_n)H(n,x) & h(n,x) \\ H(n,x) & Tv_{n+1}(x) \end{bmatrix} \tag{5}$$

for $n = 0, 1, \ldots, N$, where $Tv_\cdot(x) = E_x v_\cdot(X_1)$ and val A denotes a value of the two-person zero-sum game with payoff matrix A (see [10], [21]).

To prove the correctness of the construction let us observe that the payoff matrix in (5) has the form

$$A = \begin{array}{cc} & \begin{array}{cc} \text{zall s} & \text{f} \end{array} \\ \begin{array}{c} \text{s} \\ \text{f} \end{array} & \begin{bmatrix} (a-b)\alpha + b & a \\ b & c \end{bmatrix} \end{array}, \tag{6}$$

where a, b, c, and α are real numbers and $\alpha \in [0,1]$. By direct checking we have the following lemma.

LEMMA 2.3. *The two-person zero-sum game with payoff matrix A given by* (6) *has an equilibrium point (ϵ, δ) in pure strategies, where*

$$(\epsilon,\delta) = \begin{cases} (s,s) & \text{if } a \ge b, \\ (s,f) & \text{if } c \le a < b, \\ (f,s) & \text{if } a < b \le c, \\ (f,f) & \text{if } a < c < b. \end{cases}$$

Notice that v_{N+1} is measurable. Let us assume that functions v_i, $i = N, N-1, \dots, n+1$, are measurable. Denote

$$\begin{aligned}
A_n^{\mathrm{ss}} &= \{x \in \mathbb{E} : h(n,x) \geq H(n,x)\},\\
A_n^{\mathrm{sf}} &= \{x \in \mathbb{E} : h(n,x) < H(n,x), h(n,x) \geq Tv_{n+1}(x)\},\\
A_n^{\mathrm{fs}} &= \{x \in \mathbb{E} : h(n,x) < H(n,x), H(n,x) \leq Tv_{n+1}(x)\},
\end{aligned}$$

and

$$A_n^{\mathrm{ff}} = \mathbb{E} \setminus (A_n^{\mathrm{ss}} \cup A_n^{\mathrm{sf}} \cup A_n^{\mathrm{fs}}).$$

By sets $A_n^{\mathrm{ss}}, A_n^{\mathrm{sf}}$, and $A_n^{\mathrm{fs}} \in \mathcal{B}$ and Lemma 2.3 we have

$$\begin{aligned}
v_n(x) = {} & [(h(n,x) - H(n,x))\alpha_n + H(n,x)]\mathbb{I}_{A_n^{\mathrm{ss}}}(x) + h(n,x)\mathbb{I}_{A_n^{\mathrm{sf}}}(x)\\
& + H(n,x)\mathbb{I}_{A_n^{\mathrm{fs}}(x)} + Tv_{n+1}(x)\mathbb{I}_{A_n^{\mathrm{ff}}(x)};
\end{aligned}$$

hence $v_n(x)$ is $\mathcal{B}$-measurable.

Define $\lambda^* = \inf_n\{X_n \in A_n^{\mathrm{ss}} \cup A_n^{\mathrm{sf}}\}$ and $\mu^* = \inf_n\{X_n \in A_n^{\mathrm{ss}} \cup A_n^{\mathrm{fs}}\}$.

THEOREM 2.4. *Game $\mathcal{G}_a$ with payoff function* (3) *and sets of strategies Λ^N and M^N for Players* 1 *and* 2, *respectively, has a solution. Pair (λ^*, μ^*) is the equilibrium point and $v_0(x)$ is the value of the game.*

Proof. The theorem follows from the results in [21]. The form of the equilibrium point is obtained by Lemma 2.3. □

Let us construct an equilibrium pair for game $\mathcal{G}$. Define (see [3])

$$\sigma_n^{1^*} = \inf\{m > n : S_{N-m}(X_m, X_n) = f(X_m, X_n)\}, \tag{7}$$

$$\sigma_n^{2^*} = \inf\{m > n : s_{N-m}(X_n, X_m) = f(X_n, X_m)\}. \tag{8}$$

Let (λ^*, μ^*) be an equilibrium point in $\mathcal{G}_a$.

THEOREM 2.5. *Game $\mathcal{G}$ has a solution. Pair (s^*, t^*) such that $s^* = (\lambda^*, \{\sigma_n^{1^*}\})$ and $t^* = (\mu^*, \{\sigma_n^{2^*}\})$ is the equilibrium point. The value of the game is $v_0(x)$.*

Proof. Let

$$\tau_1^* = \lambda^* \mathbb{I}_{\{\lambda^* < \mu^*\}} + (\lambda^* \mathbb{I}_{\{\xi_{\lambda^*} \leq \alpha_{\mu^*}\}} + \sigma_{\mu^*}^{1^*} \mathbb{I}_{\{\xi_{\lambda^*} > \alpha_{\lambda^*}\}})\mathbb{I}_{\{\lambda^* = \mu^*\}} + \sigma_{\mu^*}^{1^*} \mathbb{I}_{\{\lambda^* > \mu^*\}}$$

and

$$\tau_2^* = \mu^* \mathbb{I}_{\{\lambda^* > \mu^*\}} + (\mu^* \mathbb{I}_{\{\xi_{\lambda^*} > \alpha_{\lambda^*}\}} + \sigma_{\lambda^*}^{2^*} \mathbb{I}_{\{\xi_{\lambda^*} \leq \alpha_{\lambda^*}\}})\mathbb{I}_{\{\lambda^* = \mu^*\}} + \sigma_{\lambda^*}^{2^*} \mathbb{I}_{\{\lambda^* < \mu^*\}}.$$

We obtain by the properties of conditional expectation and by (7) and (8)

$$\begin{aligned}
\tilde{R}(x, s^*, t^*) = E_x f(X_{\tau_1^*}, X_{\tau_2^*}) = {} & E_x[\mathbb{I}_{\{\lambda^* < \mu^*\} \cup \{\lambda^* = \mu^*, \xi_{\lambda^*} \leq \alpha_{\mu^*}\}} f(X_{\lambda^*}, X_{\sigma_{\lambda^*}^{2*}})\\
& + \mathbb{I}_{\{\lambda^* > \mu^*\} \cup \{\lambda^* = \mu^*, \xi_{\lambda^*} > \alpha_{\lambda^*}\}} f(X_{\sigma_{\mu^*}^{1*}}, X_{\mu^*})]\\
= {} & E_x \mathbb{I}_{\{\lambda^* < \mu^*\} \cup \{\lambda^* = \mu^*, \xi_{\lambda^*} \leq \alpha_{\mu^*}\}} E_{X_{\lambda^*}} f(X_{\lambda^*}, X_{\sigma_{\lambda^*}^{2*}})\\
& + E_x \mathbb{I}_{\{\lambda^* > \mu^*\} \cup \{\lambda^* = \mu^*, \xi_{\lambda^*} > \alpha_{\lambda^*}\}} E_{X_{\mu^*}} f(X_{\sigma_{\mu^*}^{1*}}, X_{\mu^*})\\
= {} & R(x, \lambda^*, \mu^*).
\end{aligned}$$

Let $t = (\mu, \{\sigma_n^2\}) \in \tilde{M}^N$. We obtain

$$\tilde{R}(x, s^*, t^*) = R(x, \lambda^*, \mu^*) \leq R(x, \lambda^*, \mu)$$

$$\begin{aligned}
&= E_x[\mathbb{I}_{\{\lambda^*<\mu\}\cup\{\lambda^*=\mu,\xi_{\lambda^*}\le\alpha_\mu\}} h(\lambda^*, X_{\lambda^*}) \\
&\qquad + \mathbb{I}_{\{\lambda^*>\mu\}\cup\{\lambda^*=\mu,\xi_{\lambda^*}>\alpha_{\lambda^*}\}} H(\mu, X_\mu)] \\
&= E_x[\mathbb{I}_{\{\lambda^*<\mu\}\cup\{\lambda^*=\mu,\xi_{\lambda^*}\le\alpha_\mu\}} E_{X_{\lambda^*}} f(X_{\lambda^*}, X_{\sigma^{2*}_{\lambda^*}}) \\
&\qquad + \mathbb{I}_{\{\lambda^*>\mu\}\cup\{\lambda^*=\mu,\xi_{\lambda^*}>\alpha_{\lambda^*}\}} H(\mu, X_\mu)] \\
&\le E_x[\mathbb{I}_{\{\lambda^*<\mu\}\cup\{\lambda^*=\mu,\xi_{\lambda^*}\le\alpha_\mu\}} E_{X_{\lambda^*}} f(X_{\lambda^*}, X_{\sigma^2_{\lambda^*}}) \\
&\qquad + \mathbb{I}_{\{\lambda^*>\mu\}\cup\{\lambda^*=\mu,\xi_{\lambda^*}>\alpha_{\lambda^*}\}} E_{X_\mu} f(X_{\sigma^{1*}_\mu}, X_\mu)] \\
&= E_x[\mathbb{I}_{\{\lambda^*<\mu\}\cup\{\lambda^*=\mu,\xi_{\lambda^*}\le\alpha_\mu\}} f(X_{\lambda^*}, X_{\sigma^2_{\lambda^*}}) \\
&\qquad + \mathbb{I}_{\{\lambda^*>\mu\}\cup\{\lambda^*=\mu,\xi_{\lambda^*}>\alpha_{\lambda^*}\}} f(X_{\sigma^{1*}_\mu}, X_\mu)] \\
&= E_x f(X_{s^*}, X_t) = \tilde{R}(x, s^*, t)
\end{aligned}$$

Similarly one can show that for every $s \in \tilde{\Lambda}^N$ we have $\tilde{R}(x,s,t^*) \le \tilde{R}(x,s^*,t^*)$. Hence (s^*,t^*) is the equilibrium pair for $\mathcal{G}$. □

3. Random priority in choosing BOS but a better one than the opponent. Let us formulate the problem. Two employers, Player 1 and Player 2, are to view a group of N applicants for vacancies in their enterprises sequentially. Each of the applicant has some characteristic unknown to the employer. Let $\mathbb{K} = \{x_1, x_2, \ldots, x_N\}$ be the set of characteristics, assuming that the values are different. The employer observes a permutation $\eta_1, \eta_2, \ldots, \eta_N$ of the elements of $\mathbb{K}$ sequentially. We assume that all permutations are equally likely. Let Z_k denote the absolute rank of the object with the characteristics η_k, i.e.,

$$Z_k = \min\left\{ r : \eta_k = \bigwedge_{1\le i_1<\cdots<i_r\le N} \; \bigvee_{1\le j\le r} \eta_{i_j} \right\},$$

($\bigwedge$ and $\bigvee$ denote minimum and maximum, respectively.) The object with the smallest characteristics has the rank 1. The decisions of the employer at each time n are based on the relative ranks $Y_1, Y_2, \ldots, Y_N$ of the applicants and the previous decisions of the opponent, where

$$Y_k = \min\left\{ r : \eta_k = \bigwedge_{1\le i_1<\cdots<i_r\le k} \; \bigvee_{1\le j\le r} \eta_{i_j} \right\}.$$

The random priority assignment model is applied when both players want to accept the same applicant. We assume that $\alpha_n = p$, $p \in [0,1]$ for every n. If the applicant is viewed, the employer must either accept or reject her. Once accepted, the applicant cannot be rejected; once rejected, the applicant cannot be reconsidered. Each employer can accept at most one applicant. The aim of the players is to accept BOS but a better one than that chosen by the opponent. Both players together can accept at most two objects. It makes the problem resemble the optimal double stop of the Markov process (cf. [9], [18], [3]). It is a generalization of the best-choice problem. Excellent reviews of the extensions and modifications of the best-choice problem or the secretary problem are given by Gilbert and Mosteller [8], Freeman [7], and Rose [15]. The review of the game extensions of the problem are mentioned in [20]. We adopt the following payoff function here. The player obtains $+1$ from the other if he has chosen the required applicant, -1 when the opponent has done so, and 0 otherwise.

Let us describe the mathematical model of the problem. With sequential observation of the applicants we connect the probability space $(\Omega, \mathcal{F}, \mathbf{P})$. The elementary events are a permutation of the elements of $I\!\!K$, and the probability measure $\mathbf{P}$ is the uniform probability on Ω. The observable sequence of relative ranks Y_k, $k = 1, 2, \ldots, N$, defines a sequence of σ-fields $\mathcal{F}_k = \sigma(Y_1, \ldots, Y_k)$, $k = 1, 2, \ldots, N$. The random variables Y_k are independent and $\mathbf{P}(Y_k = i) = 1/k$. Denote by $\mathcal{S}^N$, the set of all Markov times τ with respect to the σ-fields $\{\mathcal{F}_k\}_{k=1}^N$. The problem considered can be formulated as follows. For $s \in \tilde{\Lambda}^N$ and $t \in \tilde{M}^N$ denote $A_i = \{\omega : X_{\tau_i} = 1\} \cup \{\omega : X_{\tau_i} = 2, X_{\tau_j} \neq 1\}$. Define the payoff function $g(s,t) = I\!\!I_{A_1} - I\!\!I_{A_2}$ and the expected payoff $G(s,t) = Eg(s,t)$. We are looking for (s^*, t^*) such that for every $s \in \tilde{\Lambda}^N$ and $t \in \tilde{M}^N$

$$G(s, t^*) \leq G(s^*, t^*) \leq G(s^*, t).$$

It is obvious that the essential decisions of the players can be made when applicants with relative rank 1 or 2 have appeared. We will call them candidates. For further consideration it is convenient to define the following random sequence $(W_k)_{k=1}^N$. Let $W_1 = (1, Y_1) = (1,1)$, $\rho_1 = 1$. Define

$$\rho_t = \inf\{r > \rho_{t-1} : Y_r \in \{1,2\}\}, \qquad t > 1,$$

$(\inf \emptyset = \infty)$ and $W_t = (\rho_t, Y_{\rho_t})$. If $\rho_t = \infty$, then we put $W_t = (\infty, \infty)$. The Markov chain $(W_t, \mathcal{G}_t, \mathbf{P}_{(1,1)})_{t=1}^N$ with state space $I\!\!E = \{(s,l) : l \in \{1,2\}, s = 1, 2, \ldots, N\} \cup \{(\infty, \infty)\}$ and $\mathcal{G}_t = \sigma(W_1, W_2, \ldots, W_t)$ is homogeneous. The one-step transition probabilities are the following:

$$\begin{aligned} p(r,s) &= \mathbf{P}\{W_{t+1} = (s, l_s) \mid W_t = (r, l_r)\} \\ &= \begin{cases} \frac{1}{2} & \text{if } r = 1,\ s = 2, \\ \frac{r(r-1)}{s(s-1)(s-2)} & \text{if } 2 \leq r < s, \\ 0 & \text{if } r \geq s \text{ or } (r = 1,\ s \neq 2), \end{cases} \end{aligned} \tag{9}$$

$p(\infty, \infty) = 1$, $p(r, \infty) = 1 - 2\sum_{s=r+1}^N p(r,s)$ for $l_s, l_r \in \{1,2\}$ and $1 \leq r \leq s \leq N$. We will call this Markov chain the auxiliary Markov chain (AMC).

The solution of the two-decision-makers problem will partially use the solution of the problem of choosing BOS (see [8], [2], [14]). The problem can be treated as an optimal stopping problem for AMC with the following payoff function:

$$f_{\mathrm{BOS}}(r, l_r) = \begin{cases} \frac{r(2N-r-1)}{N(N-1)} & \text{if} \quad l_r = 1, \\ \frac{r(r-1)}{N(N-1)} & \text{if} \quad l_r = 2. \end{cases} \tag{10}$$

Let $\mathcal{T}^N = \{\tau \in \mathcal{S}^N : \tau = r \Rightarrow Y_r \in \{1,2\}\}$. It is a set of stopping times with respect to $\mathcal{G}_t$, $t = 1, 2, \ldots, N$. We search $\tau^* \in \mathcal{S}^N$ such that

$$\mathbf{P}\{Z_{\tau^*} \in \{1,2\}\} = \sup_{\tau \in \mathcal{S}^N} \mathbf{P}\{Z_\tau \in \{1,2\}\} = \sup_{\sigma \in \mathcal{T}^N} E_{(1,1)} f_{\mathrm{BOS}}(W_\sigma).$$

Denote $\Gamma(r,s) = \{(t, l_t) : t > r,\ l_t = 1\} \cup \{(t, l_t) : t > s,\ l_t = 2\}$. Let $r < s$ and $c(r,s) = E_{(r,l_r)} f_{\mathrm{BOS}}(W_\sigma)$, where $\sigma = \inf\{t : W_t \in \Gamma(r,s)\}$. Denote $c(r) = E_{(r,l_r)} f_{\mathrm{BOS}}(W_{\sigma_1}) = 2\frac{r(N-r)}{N(N-1)}$, where $\sigma_1 = \inf\{t : W_t \in \Gamma(r,r)\}$. We have

$$c(r,s) = \frac{r}{N(N-1)} \sum_{i=r+1}^{s-1} \frac{2N-i-1}{i-1} + \frac{r}{s-1} c(s-1) \tag{11}$$

for $r < s$, $r, s = 1, 2, \ldots, N$ ($\sum_r^s = 0$ if $s < r$). Define $r_a = \inf\{1 \leq r \leq N : f_{\mathrm{BOS}}(r,2) \geq c(r,r)\}$ and $r_b = \inf\{1 \leq r \leq r_a : f_{\mathrm{BOS}}(r,1) \geq c(r,r_a)\}$. Denote

$$\tilde{c}_{\mathrm{BOS}}(r,l_r) = \sup_{\tau \in \mathcal{S}_{r+1}^N} \mathbf{P}\{Z_\tau \in \{1,2\} \mid Y_r = l_r\}.$$

We have

$$\tilde{c}_{\mathrm{BOS}}(r,l_r) = \tilde{c}_{\mathrm{BOS}}(r) = \begin{cases} c(r) & \text{if } r_a \leq r \leq N, \\ c(r,r_a) & \text{if } r_b \leq r < r_a, \\ c(r_b-1,r_a) & \text{if } 1 \leq r < r_b. \end{cases} \tag{12}$$

The optimal stopping time for the one-decision-maker problem of choosing BOS is $\sigma^* = \inf\{t : W_t \in \Gamma(r_b, r_a)\} \in \mathcal{T}^N$ or $\tau^* = \inf\{r : (r, Y_r) \in \Gamma(r_b, r_a)\} \in \mathcal{S}^N$. We have $a = \lim_{N\to\infty} \frac{r_a}{N} = \frac{2}{3}$, $b = \lim_{N\to\infty} \frac{r_b}{N} \cong 0.3470$ and $\lim_{N\to\infty} \tilde{c}_{\mathrm{BOS}}(1) \cong 0.5736$ (cf. [11], [2], [19]).

To solve the two-person competitive stopping problem described at the beginning of the section let us perform a strategy of the players when one of them accepts some observation at moment r with relative rank $Y_r = l_r$. It is enough to focus our attention on a situation when Player 1 has accepted and Player 2 is alone in the remaining decision process. Player 2 will use a strategy $\sigma_r^{2^*} = \varsigma^*(r,l_r)$. The strategy $\varsigma^*(r,l_r)$ is such that

$$h(r,l_r) = E_{(r,l_r)} g((r,\sigma_\mu^1),(\mu,\varsigma^*(r,l_r))) = \inf_{\sigma \in \mathcal{S}_{r+1}^N} E_{(r,l_r)} g((r,\sigma_\mu^1),(\mu,\sigma)), \tag{13}$$

where the expectation is taken with respect to $\mathbf{P}_{(r,l_r)}$ of AMC. To perform strategy $\varsigma^*(r,l_r)$ let us consider the possible essential situations. Let $W_t = (r,l_r)$. Since Player 2 minimizes his expected loss (cf. (13)), he can do so by stopping on some object with relative rank 1 or 2. If $l_r = 1$, then he cannot change the payoff by stopping on the objects with relative rank 2 before another one having relative rank 1 has appeared. Let $W_s = (m,1)$ and $W_u = (n,2)$ for $u = t+1, t+2, \ldots, s-1$. Player 1 can be the winner in this case if $W_{s+1} = (\infty,\infty)$ and Player 2 does not accept the mth object. We see that it is the first moment after the acceptance decision of Player 1 when Player 2 can change the gain of Player 1. We want to know if it is optimal for Player 2 to stop at $(m,1)$. If he stops, he has -1 with the probability $f_{\mathrm{BOS}}(m,1)$ (see (10)). When he passes over and will behave optimally in the future, he has -1 with probability $\tilde{c}_{\mathrm{BOS}}(m)$. Since he minimizes his loss, his optimal strategy in $(m,1)$ is the same as in the previously mentioned one-player problem. If it happens that $n < m < r_b$, then according to the optimal strategy in the problem of one player choosing BOS, the mth object will not be accepted and Player 2 will behave according to σ^*. It means Player 1 will have $+1$ if the nth object is the best or it happens that his candidate is the absolute second and the best one will not be chosen by Player 2 (because she has appeared before r_b). If $l_r = 2$, then the optimal behaviour of Player 2 is to use the optimal strategy for the problem of one player choosing BOS. Hence, by (9), (12), and (10) we have

$$\begin{aligned} h_1(r,l_r) &= \mathbf{P}\{A_1 \mid \tau_1 = r, \tau_2 > \tau_1, Y_r = l_r\} \\ &= \begin{cases} \frac{r}{N} + \begin{cases} 0 & \text{if } r \geq r_b \\ \sum_{s=r+1}^{r_b-1} \frac{r}{s(s-1)} \frac{s(s-1)}{N(N-1)} & \text{if } r < r_b \end{cases} & \text{for } l_r = 1, \\ \frac{r(r-1)}{N(N-1)} & \text{for } l_r = 2 \end{cases} \end{aligned}$$

and

$$h_2(r,l_r) = \mathbf{P}\{A_2 \mid \tau_1 = r, \tau_2 > \tau_1, Y_r = l_r\}$$
$$= \begin{cases} \sum_{s=r+1}^{N} \frac{r}{s(s-1)} \max\left\{\frac{s(2N-s-1)}{N(N-1)}, \tilde{c}_{\mathrm{BOS}}(s)\right\} & \text{if } l_r = 1, \\ \tilde{c}_{\mathrm{BOS}}(r) & \text{if } l_r = 2, \end{cases}$$

where $\sigma_n^{i^*} = \varsigma^*(n, Y_n)$, $s = (\lambda, \{\sigma_n^{1^*}\})$, and $t = (\mu, \{\sigma_n^{2^*}\})$. The optimal strategy ς^*, after the first acceptance has been made at moment r on $Y_r = l_r$, has the form

$$\varsigma^*(r,l_r) = \begin{cases} \begin{cases} \vartheta_r & \text{if } \vartheta_r \geq r_b \\ \sigma^*_{\vartheta_r} & \text{if } \vartheta_r < r_b \end{cases} & \text{for } l_r = 1, \\ \sigma_r^* & \text{for } l_r = 2, \end{cases} \tag{14}$$

where $\vartheta_r = \inf\{s > r : Y_s = 1\}$ and $\sigma_r^* = \inf\{s > r : (s, Y_s) \in \Gamma(r_b, r_a)\}$.

Let us consider the auxiliary game described in §2. The above comments and the solution of the problem of one player choosing BOS suggest that it is enough to consider AMC. We will use it for further calculations. For presentation of strategies the sequence of relative ranks $(Y_n)_{n=1}^N$ is more convenient. To avoid some misunderstanding the arguments of the functions h, H, and v are the moment n and the relative rank Y_n as well as the state of process W_t. We have also $h(r,l_r) = -H(r,l_r) = h_1(r,l_r) - h_2(r,l_r)$. By backward induction we construct the strategies and the value of the game. The results of this calculation and consideration can be presented as follows.

Denote for $p \in [.5, 1]$

$$\begin{aligned} w(r,s,t,u;p) = &\sum_{j=r+1}^{s-1} p(r,j)[H(j,1) + w(j,s,t,u;p)] \\ &+ \sum_{j=s}^{t-1} p(r,j)[h(j,1)(2p-1) + w(j,j+1,t,u;p)] \\ &+ \sum_{j=t}^{u-1} p(r,j)[h(j,1)(2p-1) + H(j,2)] \\ &+ (2p-1)\sum_{j=u}^{N} p(r,j)[h(j,1) + h(j,2)] \end{aligned}$$

and $B_{st}(k) = \{(r,l_r) : s \leq r \leq t,\ l_r = k\}$. Solving (5) recursively we have $r_a = \min\{1 \leq r \leq N : h(r,2) \geq 0\}$ and obtain that there exist $r_{\alpha(p)} = \min\{r < r_a : H(r,2) \leq \tilde{v}(r;p)\}$, $r_\gamma = \min\{r < r_{\alpha(p)} : h(r,1) \geq 0\}$, and $r_{\beta(p)} = \min\{r < r_\gamma : H(r,1) \leq \tilde{v}(r;p)\}$ such that

$$v(r,l_r;p) = \begin{cases} h(r,l_r)(2p-1) & \text{if } (r,l_r) \in B_{r_\gamma N}(1) \cup B_{r_a N}(2), \\ H(r,l_r) & \text{if } (r,l_r) \in B_{r_{\beta(p)} r_\gamma - 1}(1) \cup B_{r_{\alpha(p)} r_a - 1}(2), \\ \tilde{v}(r;p) & \text{if } (r,l_r) \in B_{1 r_{\beta(p)} - 1}(1) \cup B_{1 r_{\alpha(p)} - 1}(2), \end{cases} \tag{15}$$

where

$$\tilde{v}(r;p) = \tilde{v}(r,l_r;p) = Tv(r,l_r;p) = \begin{cases} w(r,r+1,r+1,r+1;p) & \textit{if } r_a \leq r \leq N, \\ w(r,r+1,r+1,r_a;p) & \textit{if } r_{\alpha(p)} \leq r < r_a, \\ w(r,r+1,r_{\alpha(p)},r_a;p) & \textit{if } r_\gamma \leq r < r_{\alpha(p)}, \\ w(r,r_\gamma,r_{\alpha(p)},r_a;p) & \textit{if } r_{\beta(p)} \leq r < r_\gamma, \\ w(r_{\beta(p)},r_\gamma,r_{\alpha(p)},r_a;p) & \textit{if } 1 \leq r < r_{\beta(p)}. \end{cases} \tag{16}$$

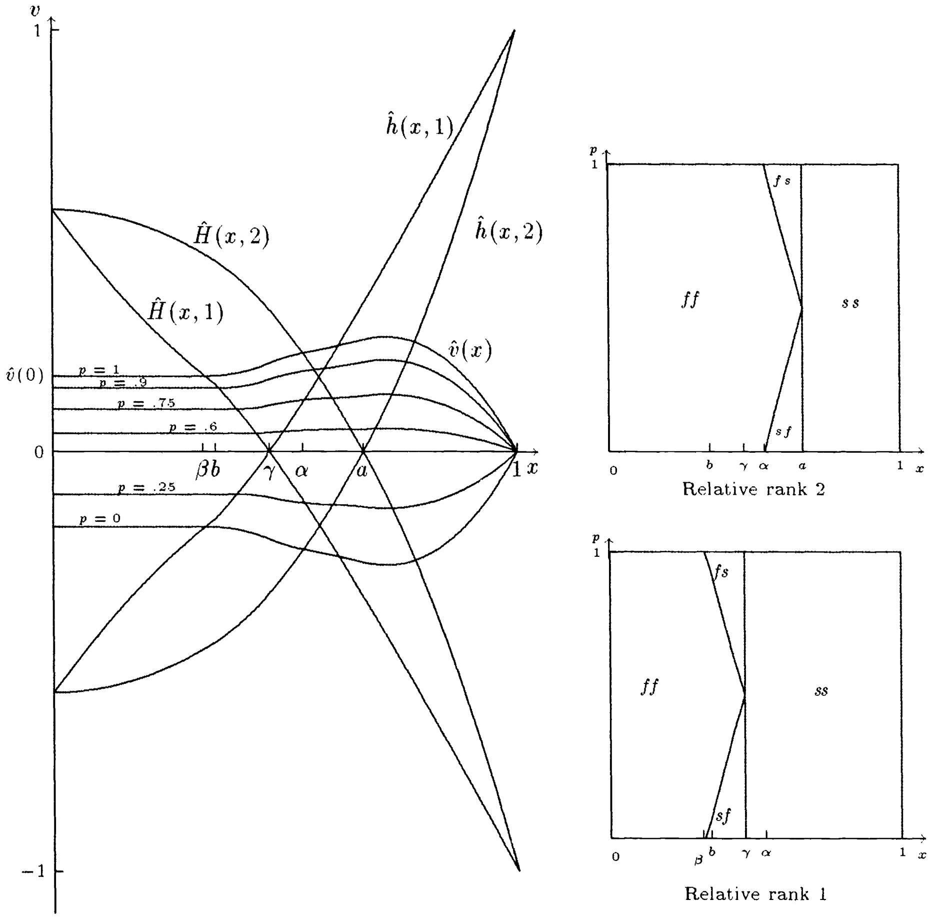


FIG. 1. *The gain functions, the value function, and strategies of the first stop in the* BOS *vs.* BOS *game.*

The optimal first-stop strategy is given by the sets (see §2 and Fig. 1) $A_t^{ss} = B_{r_\gamma N}(1) \cup B_{r_a N}(2)$, $A_t^{fs} = B_{r_{\beta(p)} r_\gamma}(1) \cup B_{r_{\alpha(p)} r_a}(2)$, $A_t^{ff} = I\!\!E \setminus (A_t^{ss} \cup A_t^{fs})$, $t = 1, 2, \ldots, N$.

The above solution was obtained in the following way. Let us assume $p \in [0.5, 1]$. For $n = N$ we have that ss is an equilibrium point. Assume for induction that ss is the equilibrium for $n, n+1, \ldots, N$ and $l_n = 1, 2$. It is not easy to check whether this assumption implies that ss is the equilibrium at $n-1$ and $l_{n-1} = 1, 2$. Instead of looking for a solution for every N, we will be content with the limit of the value of the game and the asymptotic behavior of the equilibrium strategies. By monotonicity of $h(r, l_r)$ and $H(r, l_r)$ on r this approach gives the solution for large N. To this end we find the limit of functions $h(r, l_r)$, $H(r, i_r)$ and under the induction assumption the limit of $\tilde{v}(r, l_r; p)$ when $N \to \infty$ in such a way that $\frac{r}{N} \to x$. Let us denote these

TABLE 1
The value of the game BOS *vs.* BOS *and decision point as a function of priority.*

Priority p	$\alpha(p)$	$\beta(p)$	Value of the game
1.00	0.5365	0.3263	0.1789
0.95	0.5485	0.3421	0.1651
0.90	0.5608	0.3557	0.1505
0.85	0.5736	0.3682	0.1350
0.80	0.5866	0.3809	0.1186
0.75	0.5998	0.3940	0.1013
0.70	0.6131	0.4073	0.0831
0.65	0.6266	0.4209	0.0639
0.60	0.6400	0.4349	0.0436
0.55	0.6534	0.4493	0.0223
0.50	0.6667	0.4639	0

limits by $\hat{h}(x,l)$, $\hat{H}(x,l)$, and $\hat{v}(x,l)$, respectively. Next, we are looking for x_{ss} such that the strategy ss is the equilibrium for the game

$$\begin{array}{c|cc} & \text{s} & \text{f} \\ \hline \text{s} & \hat{h}(x,l)(2p-1) & \hat{h}(x,l) \\ \text{f} & \hat{H}(x,l) & \hat{v}(x,l) \end{array} \tag{17}$$

for every $x \in [x_{ss}, 1]$, $l = 1, 2$. In the presence of the above, the value x_{ss} is the value of x nearest to 1 at which the equilibrium will change to the other one. We have $x_{ss} = a$. One can show that if $\epsilon > 0$ is small enough, then for $x \in (x_{ss} - \epsilon, x_{ss})$ the equilibrium point is ss at $(x,1)$ and fs at $(x,2)$. We can also say, by the above consideration, that for large N there exists r_{ss} such that for every $r \geq r_{ss}$ and $l = 1,2$ the equilibrium is ss and for $r = r_{ss} - 1$ at one of the state $(r_{ss} - 1, 1)$ or $(r_{ss} - 1, 2)$ the strategy ss is not the equilibrium. One can check that for $p \in (0.5, 1]$ we have that at $(r_{ss} - 1, 2)$ the strategy fs is the equilibrium strategy and at the state $(r_{ss} - 1, 1)$ the strategy ss is the equilibrium. Let us assume for backward induction that for $r, r+1, \ldots, r_{ss} - 1$ the equilibrium strategies are the same. By analysis of (17) for $x < x_{ss}$ similarly as for $x \in [x_{ss}, 1]$, one can find $x_{fs} \leq x_{ss}$ such that for $x \in [x_{fs}, x_{ss})$ at the state $(x, 1)$ the equilibrium is ss and at $(x, 2)$ the equilibrium is fs. When we iterate the above consideration up to $r = 1$ (or $x = 0$ in limit case), the construction of the strategy and the value of the game will be finished. The strategy is presented by constants $\beta(p)$, γ, $\alpha(p)$, and a. The numeric value of $\gamma \cong 0.4639$ and $a = \frac{2}{3}$. The parameters α and β depend on the value of the priority parameter p, and some values of p are given in Table 1 (see also Fig. 1). The value of the game also depends on p (see Table 1 and Fig. 1). By asymmetry of the problem with respect to $p = 0.5$ we have that the value of the problem fulfills the relation $\tilde{v}(r;p) = -\tilde{v}(r; 1-p)$. We have also following relations between stopping sets for equilibrium strategies. We have that A^{fs} for p and A^{sf} for $1-p$ are equal.

The analytical description of the value function follows.

Denote

$$\begin{aligned} \hat{w}_1(x_2, y_1, y_2, z; p) &= 2x_2 \ln \frac{x_2}{y_1} + x_2(y_1 - x_2) + x_2[(\ln x_2)^2 - (\ln y_1)^2] \\ &\quad + (2p-1)\left[2x_2 \ln \frac{y_2}{y_1} + x_2(y_1 - y_2) + x_2[(\ln y_2)^2 - (\ln y_1)^2]\right] \\ &\quad + x_2 y_2 \left[(4(2p-1) - 3z + 2\ln z)\left(\frac{1}{y_2} - \frac{1}{z}\right) - (2p-1)\ln \frac{z}{y_2}\right. \end{aligned}$$

$$+ 4(p-1)\left[\frac{\ln y_2}{y_2} - \frac{\ln z}{z}\right] + 2(2p-1)\left(\frac{1}{z} + \frac{\ln z}{z} - \ln ez\right)\Bigg],$$

Let p_0 be the solution of the equation $\hat{h}(b,1) = \hat{w}_1(b,\gamma,\alpha(p),a;p)$ in $[0.5,1]$. Such a solution exists since $\hat{w}_1$ is a nondecreasing function of p and $\hat{h}(b,1) < \hat{w}_1$ for $p = 1$. We have $p_0 \cong 0.9358$, and for $p \in [0.5,1]$ we define

$$\begin{aligned}\hat{w}(x_1,x_2,y_1,y_2,z;p) = & \left[(2 - x_2 + x_1)(x_2 - x_1) + (4 - x_2 + 2\ln x_2)x_1 \ln\frac{x_1}{x_2}\right] I\!I_{\{p>p_0\}} \\ & + 2x_1 \ln\frac{x_2}{y_1} + x_1(y_1 - x_2) + x_1[(\ln x_2)^2 - (\ln y_1)^2] \\ & + (2p-1)\left[2x_1 \ln\frac{y_2}{y_1} + x_1(y_1 - y_2) + x_1[(\ln y_2)^2 - (\ln y_1)^2]\right] \\ & + x_1 y_2 \Bigg[[4(2p-1) - 3z + 2\ln z]\left(\frac{1}{y_2} - \frac{1}{z}\right) - (2p-1)\ln\frac{z}{y_2} \\ & + 4(p-1)\left(\frac{\ln y_2}{y_2} - \frac{\ln z}{z}\right) + 2(2p-1)\left(\frac{1}{z} + \frac{\ln z}{z} - \ln ez\right)\Bigg].\end{aligned}$$

We have

$$(18)\ \hat{v}(x;p) = \lim_{N\to\infty} \tilde{v}(r;p) = \begin{cases} \hat{w}(x,x,x,x,x;p) & \textit{if } a \le x \le 1, \\ \hat{w}(x,x,x,x,a;p) & \textit{if } \alpha(p) \le x < a, \\ \hat{w}(x,x,x,\alpha(p),a;p) & \textit{if } \gamma \le x < \alpha(p), \\ \hat{w}(x,x,\gamma,\alpha(p),a;p) & \textit{if } \max(b,\beta(p) \le x < \gamma, \\ \hat{w}(x,\max(\beta(p),b),\gamma,\alpha,a;p) & \textit{if } \min(\beta(p),b) \le x < b, \\ \hat{w}(\beta,\max(\beta(p),b),\gamma,\alpha,a;p) & \textit{if } 0 \le x < \min(\beta(p),b). \end{cases}$$

So we can formulate the following theorem.

THEOREM 3.1. *In the competitive two-person problem of choosing* BOS *but better than the opponent the asymptotically optimal strategy of the first stop is described by the sets* A_t^{ss}, A_t^{fs}, *and* A_t^{ff}. *The second stop is according to* (14). *The value function of the problem is given by* (15), *the expected value with respect to* $\mathbf{P}_{(r,l_r)}$ *of* AMC *by* (16), *and its limit by* (18).

4. The best vs. the BOS game. Let us consider once more the game investigated in §3, but let us assume that the aim of Player 1 is to accept the BA. The meaning of the most of the denotations is the same as in §3. The changes of notation will be given.

We denote by A_i the random event that the ith player is a winner. For $s \in \tilde{\Lambda}^N$ and $t \in \tilde{M}^N$ we have $A_1 = \{\omega : X_{\tau_1} = 1\}$ and $A_2 = \{\omega : X_{\tau_2} = 1\} \cup \{\omega : X_{\tau_2} = 2,\ X_{\tau_1} \neq 1\}$. The payoff function in this game is $g_1(s,t) = I\!I_{A_1} - I\!I_{A_2}$.

First of all, let us consider the position when one of the players has accepted some object at moment r with the relative rank $Y_r = l_r$. We construct the strategy of the player who has not accepted any state yet. Since the aims of the players are different, we have to consider independently the situations when Player 1 has stopped as the first and when Player 2 has done so. We introduce the useful denotation

$$h_{ik}(r,l_r) = \mathbf{P}(A_k | \tau_i = r, \tau_j > \tau_i, Y_r = l_r)$$

for $k,i,j = 1,2$, $i \neq j$, $r = 1,2,\ldots,N$, $l_r = 1,2$.

Let Player 1 stop the process as the first at the moment r on the object with $Y_r = l_r$. As he wants to accept the object with the absolute rank 1, it is obvious that he will stop on the relatively first object. He will also probably accept, in some circumstances, the relatively second objects to disturb Player 2 in the realization of his aims. We will see that this supposition is true. Player 2 staying alone will use a strategy $\sigma_r^{2^*} = \varsigma^*(r, l_r)$ defined in (14) with $g(s,t) = g_1(s,t)$. Let $W_t = (r, l_r)$. Since Player 2 minimizes his expected loss and he would like to choose the BOS object, he can do so by stopping on some object with relative rank 1 or 2. The optimal strategy ς^* is given by (13). Consequently,

$$h(r, l_r) = h_{11}(r, l_r) - h_{12}(r, l_r),$$

where we have $h_{11}(r, l_r) = \frac{r}{N}$ for $l_r = 1$ and 0; otherwise, $h_{12}(r, l_r) = h_2(r, l_r)$.

Let us assume that Player 2 has stopped the process as the first on some object at moment r with relative rank $Y_r = l_r$. Player 1 will use a strategy $\sigma_r^{1^*} = \delta^*(r, l_r)$. The strategy $\delta^*(r, l_r)$ is such that

$$H(r, l_r) = E_{(r,l_r)} g_1((\lambda, \delta^*(r, l_r)), (r, \sigma_\lambda^2)) = \sup_{\sigma \in \mathcal{S}_{r+1}^N} E_{(r,l_r)} g_1((\lambda, \sigma), (r, \sigma_\lambda^2)).$$

Let $W_t = (r, l_r)$. Since Player 1 maximizes his expected gain and would like to choose the best object, he can do so by stopping on some object with relative rank 1. Denote $\tilde{c}_{\mathrm{BA}}(r) = \sup_{\tau \in \mathcal{S}_{r+1}^N} \mathbf{P}\{Z_\tau = 1 | Y_r = l_r\}$, $r_c = \inf\{1 \le r \le N : \sum_{i=r+1}^N \frac{1}{i-1} \le 1\}$, and $\tau_r^* = \inf\{s > r : Y_s = 1, s \ge r_c\}$. The optimal strategy δ^* of Player 1, after the first acceptance at the moment r on $Y_r = l_r$ by Player 2, has the form

$$\delta^*(r, l_r) = \begin{cases} \begin{cases} \vartheta_r & \text{if } \vartheta_r \ge r_c \\ \tau_{\vartheta_r}^* & \text{if } \vartheta_r < r_c \end{cases} & \text{for } l_r = 1, \\ \tau_r^* & \text{for } l_r = 2, \end{cases} \tag{19}$$

where ϑ_r is the first moment after r when $Y_r = 1$. We have

$$H(r, l_r) = h_{21}(r, l_r) - h_{22}(r, l_r),$$

where

$$h_{21}(r, l_r) = \sum_{s=r+1}^{N} p(r,s) \left[\max\left\{\frac{s}{N}, \tilde{c}_{\mathrm{BA}}(r)\right\} + \tilde{c}_{\mathrm{BA}}(r)\right] = \tilde{c}_{\mathrm{BA}}(r)$$

and

$$\begin{aligned} h_{22}(r, l_r) &= \mathbf{P}\{A_2 \mid \tau_2 = r,\ \tau_1 > \tau_2,\ Y_r = l_r\} \\ &= \begin{cases} \frac{r}{N} + \begin{cases} 0 & \text{if } r \ge r_c \\ \sum_{s=r+1}^{r_c - 1} \frac{r}{s(s-1)} \frac{s(s-1)}{N(N-1)} & \text{if } r < r_c \end{cases} & \text{for } l_r = 1, \\ \frac{r(r-1)}{N(N-1)} & \text{for } l_r = 2 \end{cases} \end{aligned}$$

Denote $h_p(r, l_r) = p h(r, l_r) + (1-p) H(r, l_r)$. Define $r_d = \min\{1 \le r \le N : h(r, 2) \ge H(r, 2)\}$ and $r_\ell = \min\{1 \le r \le r_d : h(r,1) \ge H(r,1)\}$. During the recursive construction of $\tilde{v}(r, l_r; p)$ and the strategy according to Theorems 2.4 and 2.5 (see also (5)) for a large N we get that there exist $r_{\nu(p)} = \min\{r < r_d : H(r,2) \le \tilde{v}(r;p)\}$ and $\tilde{p}_1 = \min\{0 \le p \le 1 : h(\ell, 1) < \tilde{v}(\ell; p)\}$. For $p \ge \tilde{p}_1$ there exists $r_{\kappa(p)} = \min\{r \le r_\ell :$

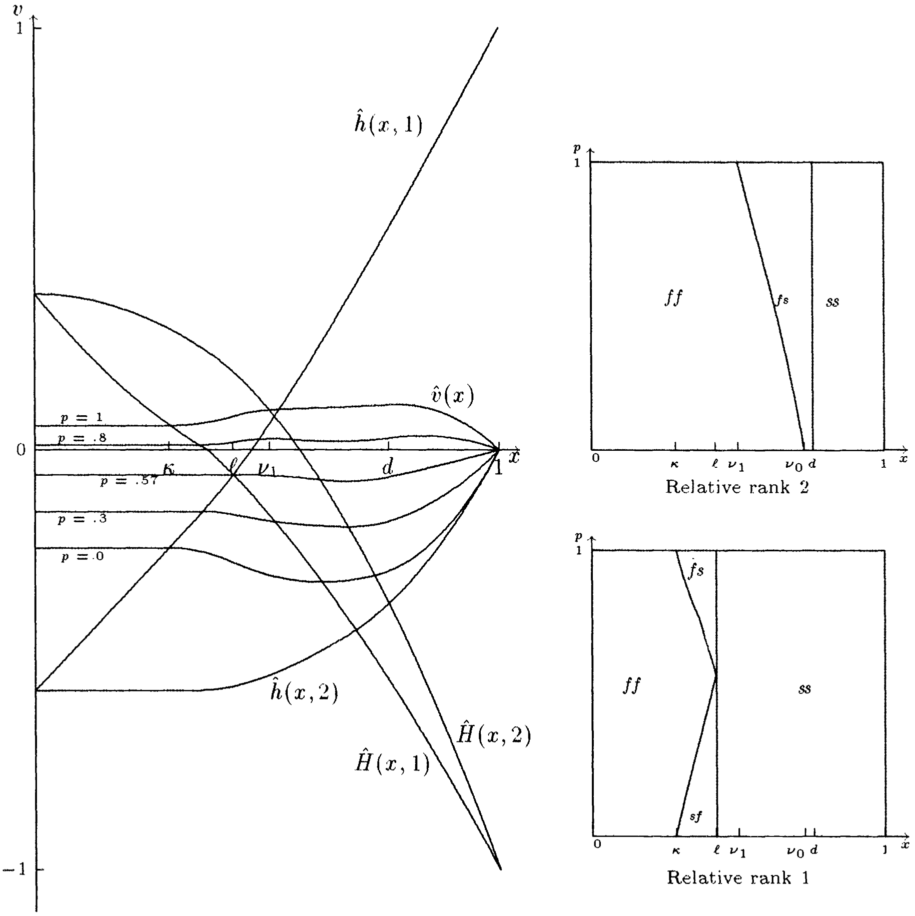


FIG. 2. *The gain function, the value functions, and strategies of the first stop in the random priority game BA vs. BOS but a better one than the opponent.*

$H(r,1) \le \tilde{v}(r;p)\}$, and for $p < \tilde{p}_1$ there exists $r_{\kappa(p)} = \min\{r \le r_\ell : h(r,1) \ge \tilde{v}(r;p)\}$. These points r_d, r_ℓ, $r_{\nu(p)}$, and $r_{\kappa(p)}$ are such that

$$
(20) \qquad v(r,l_r;p) = \begin{cases} h_p(r,l_r) & \textit{if } (r,l_r) \in B_{r_\ell N}(1) \cup B_{r_d N}(2), \\ H(r,l_r) & \textit{if } (r,l_r) \in B_{r_{\nu(p)} r_d - 1}(2), \\ H(r,l_r)\mathbb{I}_{\{p \ge \tilde{p}_1\}} & \\ \quad + h(r,l_r)\mathbb{I}_{\{p < \tilde{p}_1\}} & \textit{if } (r,l_r) \in B_{r_{\kappa(p)} r_\ell - 1}(1), \\ \tilde{v}(r;p) & \textit{if } (r,l_r) \in B_{1 r_{\kappa(p)} - 1}(1) \cup B_{1 r_{\nu(p)} - 1}(2), \end{cases}
$$

where

$$\tilde{v}(r;p) = Tv(r,l_r;p) = \begin{cases} w(r,r+1,r+1,r+1;p) & \text{if } r_d \le r \le N, \\ w(r,r+1,r+1,r_d;p) & \text{if } r_{\nu(p)} \le r < r_d, \\ w(r,r+1,r_{\nu(p)},r_d;p) & \text{if } r_\ell \le r < r_{\nu(p)}, \\ w(r,r_\ell,r_{\nu(p)},r_d;p) & \text{if } r_{\kappa(p)} \le r < r_\ell, \\ w(r_{\kappa(p)},r_\ell,r_{\nu(p)},r_d;p) & \text{if } 1 \le r < r_{\kappa(p)} \end{cases} \tag{21}$$

and

$$\begin{aligned} w(r,s,t,u;p) = {} & \sum_{j=r+1}^{s-1} \frac{r}{j(j-1)}[H(j,1)\mathbb{I}_{\{p\ge\tilde{p}_1\}} + h(j,1)\mathbb{I}_{\{p<\tilde{p}_1\}}] \\ & + \sum_{j=s}^{t-1} \frac{r}{j(j-1)} h_p(j,1) + \sum_{j=t}^{u-1} \frac{r(t-2)}{j(j-1)(j-2)}[h_p(j,1) + H(j,2)] \\ & + \sum_{j=u}^{N} \frac{r(t-2)}{j(j-1)(j-2)}[h_p(j,1) + h_p(j,2)] \end{aligned}$$

for $r \le s \le t \le u$. The optimal first-stop strategy is given by sets $A_t^{ss} = B_{r_\ell N}(1) \cup B_{r_d N}(2)$, $A_t^{fs} = (\mathbb{I}_{\{p\ge\tilde{p}_1\}} B_{r_{\kappa(p)} r_\ell - 1}(1)) \cup B_{r_{\nu(p)} r_d - 1}(2)$, $A_t^{sf} = \mathbb{I}_{\{p<\tilde{p}_1\}} B_{r_{\kappa(p)} r_\ell - 1}(1)$, $A_t^{ff} = \mathbb{E} \setminus (A_t^{ss} \cup A_t^{fs} \cup A_t^{sf})$, $t = 1,2,\dots,N$ (see §§2 and 3 and Fig. 2). Here we adopt the convention that for every set A we have $1 \cdot A = A$ and $0 \cdot A = \emptyset$, where $\emptyset$ is the empty set.

The function $w(r,s,t,u;p)$ also depends on r_b and r_c. Let $r \le s \le r \le t \le u$. When $N \to \infty$ and $\frac{r}{N} \to x_1$, $\frac{s}{N} \to x_2$, $\frac{t}{N} \to y_1$, and $\frac{u}{N} \to y_2$, we get

$$\begin{aligned} \hat{w}(x_1,x_2,y_1,y_2;p) &= \lim_{N\to\infty} w(r,s,t,u;p) \\ &= \hat{w}_{21}(x_1,x_2,y_1,y_2;p)\mathbb{I}_{\{p\ge p_1\}} + \hat{w}_{22}(x_1,x_2,y_1,y_2;p)\mathbb{I}_{\{p<p_1\}}, \end{aligned}$$

where

$$\begin{aligned} \hat{w}_{21}(x_1,x_2,y_1,y_2;p) = x_1 & \left[\left(\ln\frac{x_1}{x_2} - \frac{1}{2}((\ln x_2)^2 - (\ln x_1)^2) \right) \mathbb{I}_{\{x_1>c\}} \right. \\ & \left. - \left(\frac{1}{2} + \ln x_2 + \frac{1}{2}(\ln x_2)^2 \right) \mathbb{I}_{\{x_1\le c\}} \right] \\ & + \frac{x_1 \mathbb{I}_{\{x_1>c\}} + c\mathbb{I}\{x_1 \le c\}}{x_2} \hat{w}_1(x_2,y_1,y_2;p), \end{aligned}$$

$$\begin{aligned} \hat{w}_{22}(x_1,x_2,y_1,y_2;p) = x_1 & \left[\left((\ln x_2)^2 - (\ln x_1)^2 + x_1 - x_2 + 2\ln\frac{x_2}{x_1} \right) \mathbb{I}_{\{x_1>b\}} \right. \\ & + \left((4 - 2b + 2\ln b)\ln\frac{b}{x_1} + (2-b)(x_1-b) \right. \\ & \left. \left. + (\ln x_2)^2 - (\ln b)^2 - x_2 + b + 2\ln\frac{x_2}{b} \right) \mathbb{I}_{\{x_1\le b\}} \right] \\ & + \frac{x_1 \mathbb{I}_{\{x_1>b\}} + b\mathbb{I}_{\{x_1\le b\}}}{x_2} w_1(x_2,y_1,y_2;p), \end{aligned}$$

TABLE 2
The value and decision points as functions of the priority in the BA *vs.* BOS *game.*

The priority p	$\nu(p)$	$\kappa(p)$	Value of the game
0.00	0.7285	0.2856	−0.2339
0.10	0.7111	0.3084	−0.2068
0.20	0.6922	0.3326	−0.1780
0.30	0.6719	0.3581	−0.1476
0.40	0.6500	0.3832	−0.1156
0.50	0.6268	0.4078	−0.0823
0.5659	0.6109	0.4237	−0.0598
0.60	0.6024	0.4134	−0.0483
0.70	0.5772	0.3843	−0.0167
0.7580	0.5624	0.3679	0
0.80	0.5517	0.3504	0.0114
0.90	0.5265	0.3157	0.0357
1.00	0.5023	0.2881	0.0568

and

$$\begin{aligned}\hat{w}_1(x_2,y_1,y_2;p) = x_2 &\left[(3p-1)\ln\frac{y_1}{x_2} + \frac{3p-1}{2}((\ln y_1)^2 - (\ln x_2)^2) - p(y_1 - x_2)\right] \\ &+ x_2 y_1 \left[3(2p-1)\left(\frac{1}{y_1} - \frac{1}{y_2}\right) + (3p-2)\left(\frac{\ln y_1}{y_1} - \frac{\ln y_2}{y_2}\right)\right. \\ &\left. + (1+p)\ln\frac{y_1}{y_2}\right] + y_1^2\left[(p-1)\left(\frac{1}{y_2} - 1\right)\right. \\ &\left. + (4p-2)\left(\frac{\ln y_2}{y_2} + \frac{1}{y_2} - 1\right) - (2p-1)\ln y_2\right].\end{aligned}$$

Parameter p_1 is asymptotic equivalent of $\tilde{p}_1$. The value of p_1 can be determined as the solution of some equation which will be given later.

Let $d = \lim_{N\to\infty} \frac{r_d}{N} \cong 0.7587$ and $\ell = \lim_{N\to\infty} \frac{r_\ell}{N} \cong 0.4237$. We have that $\nu(p) = \lim_{N\to\infty} \frac{r_\nu(p)}{N}$ is the solution of the equation $\hat{w}_1(\nu,\nu,d;p) = \hat{H}(\nu,2)$ in $[\ell,d]$. Now we can determine p_1 as the solution of the equation $\hat{w}_1(\ell,\nu(p),d;p) = \hat{H}(\ell,1)$ with respect to p in $[0,1]$. Such a solution exists since $\hat{w}_1(\ell,\nu(p),d;p)$ is a nondecreasing function of p and $\hat{H}(\ell,\nu(1),d;1) < \hat{w}_1(\ell,\nu(1),d;1)$. We have $p_1 \cong 0.5659$.

Determine $\kappa(p) = \lim_{N\to\infty} \frac{r_{\kappa(p)}}{N}$. The decision point $\kappa(p)$ is the solution of the equation $\hat{w}(\kappa,\ell,\nu(p);p) = \hat{h}(\kappa,1)\mathbb{I}_{\{p<p_1\}} + \hat{H}(\kappa,1)\mathbb{I}_{\{p\geq p_1\}}$. The asymptotic value function (see also Fig. 2 and Table 2) is

$$\text{(22)}\qquad \hat{v}(x;p) = \lim_{N\to\infty} \tilde{v}(r;p) = \begin{cases} \hat{w}(x,x,x,x;p) & \text{if } d \leq x \leq 1, \\ \hat{w}(x,x,x,d;p) & \text{if } \nu(p) \leq x < d, \\ \hat{w}(x,x,\nu(p),d;p) & \text{if } \ell \leq x < \nu(p), \\ \hat{w}(x,\ell,\nu(p),d;p) & \text{if } \kappa(p) \leq x < \ell, \\ \hat{w}(\kappa(p),\ell,\nu(p),d;p) & \text{if } 0 \leq x < \kappa(p). \end{cases}$$

We can formulate the following theorem.

THEOREM 4.1. *For a large N in the competitive two-person problem of choosing the best vs. the BOS applicant but a better one than the opponent, the asymptotically optimal strategy of the first stop is described by the sets A_t^{ss}, A_t^{fs}, A_t^{sf}, and A_t^{ff}. The second stop is according to ς^* given by* (14) *for Player* 2 *and δ^* given by* (19) *for Player* 1. *The value function of the problem is given by* (20), *the expected value with respect to $\mathbf{P}_{(r,l_r)}$ of* AMC *by* (21), *and its limit by* (22).

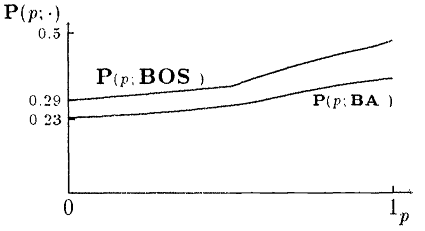


FIG. 3. *The probability of success in the* BOS *vs.* BOS *and* BA *vs.* BOS *games.*

5. Final remarks. The results of §§2 and 3 give an image of the influence of the priority level on the game value. We have naturally realized that the such influence depends also on the definition of the gain function or, in other words, the aims of the players. To compare the role of priority in the two games considered we compare the results obtained. We can observe that the BOS vs. BOS game is fair for $p = 0.5$ (§3) and the BA vs. BOS game is fair for $p \cong 0.7580$. The question is, how much does Player 2 have to reduce his demand concerning the absolute rank of the chosen applicant to recompense the full priority ($p = 1$) of Player 1 looking for the BA?

Finally, let us investigate the probability that Player 1 will be the winner in both examples considered. Let us denote as $\mathbf{P}_N(p; \mathrm{BOS})$ and $\mathbf{P}_N(p; \mathrm{BA})$ the probability of success for Player 1 in the BOS vs. BOS game (§3) and the BA nvs. BOS game (§4), respectively, when the two players adopt the equilibrium strategies and the horizon is N. From the definition of event A_1 and the definition of the equilibrium strategies in the two examples we have

$$\begin{aligned}\mathbf{P}_N(p; \mathrm{BOS}) = &\sum_{j=r_{\beta(p)}+1}^{r_\gamma - 1} \frac{r_{\beta(p)}}{j(j-1)}[\mathbb{I}_{\{p\le 0.5\}} h_1(j,1) + \mathbb{I}_{\{p>0.5\}} h_2(j,1)] \\ &+ \sum_{j=r_\gamma}^{r_{\alpha(p)}-1} \frac{r_{\beta(p)}}{j(j-1)}[ph_1(j,1) + (1-p)h_2(j,1)] \\ &+ \sum_{j=r_{\alpha(p)}}^{r_a - 1} \frac{r_{\beta(p)}(r_{\alpha(p)}-2)}{j(j-1)(j-2)}[\mathbb{I}_{\{p\le 0.5\}} h_1(j,2) \\ &+ \mathbb{I}_{\{p>0.5\}} h_2(j,2) + ph_1(j,1) + (1-p)h_2(j,1)] \\ &+ \sum_{j=r_a}^{N} \sum_{k=1}^{2} \frac{r_{\beta(p)}(r_{\alpha(p)}-2)}{j(j-1)(j-2)}[ph_1(j,k) + (1-p)h_2(j,k)]\end{aligned}$$

and

$$\begin{aligned}\mathbf{P}_N(p; \mathrm{BA}) = &\sum_{j=r_{\kappa(p)}+1}^{r_\ell - 1} \frac{r_{\kappa(p)}}{j(j-1)}[\mathbb{I}_{\{p\le \tilde{p}_1\}} h_{11}(j,1) + \mathbb{I}_{\{p>\tilde{p}_1\}} h_{21}(j,1)] \\ &+ \sum_{j=r_\ell}^{r_{\nu(p)}-1} \frac{r_{\kappa(p)}}{j(j-1)}[ph_{11}(j,1) + (1-p)h_{21}(j,1)]\end{aligned}$$

TABLE 3
The probability of success for Player 1.

Priority	Probability of success	
p	$\mathbf{P}(p;\mathrm{BA})$	$\mathbf{P}(p;\mathrm{BOS})$
0.0	0.2358	0.2900
0.1	0.2402	0.2978
0.2	0.2460	0.3064
0.3	0.2532	0.3152
0.4	0.2626	0.3243
0.5	0.2742	0.3338
0.6	0.2907	0.3679
0.7	0.3122	0.3983
0.8	0.3311	0.4250
0.9	0.3465	0.4483
1.0	0.3584	0.4765

$$+\sum_{j=r_{\nu(p)}}^{r_d-1} \frac{r_{\kappa(p)}(r_{\nu(p)}-2)}{j(j-1)(j-2)}[h_{21}(j,2)+ph_{11}(j,1)+(1-p)h_{21}(j,1)]$$

$$+\sum_{j=r_d}^{N}\sum_{k=1}^{2} \frac{r_{\kappa(p)}(r_{\nu(p)}-2)}{j(j-1)(j-2)}[ph_{11}(j,k)+(1-p)h_{21}(j,k)],$$

where $\tilde{p}_1 \cong 0.5659$ (see Fig. 3).

When $N \to \infty$ such that $\frac{r}{N} \to x$ we get

$$\begin{aligned}
\mathbf{P}(p;\mathrm{BOS}) &= \lim_{N\to\infty} \mathbf{P}_N(p;\mathrm{BOS}) \\
&= \mathbb{I}_{\{p\le 1-p_0\}}\beta(p)\left[(1+b)\ln\frac{b}{\beta(p)}+\beta(p)-b+\ln\frac{\gamma}{b}\right] \\
&\quad + \mathbb{I}_{\{1-p_0<p\le 0.5\}}\beta(p)\ln\frac{\gamma}{\beta(p)} \\
&\quad + \mathbb{I}_{\{0.5<p\le p_0\}}\beta(p)\left[\gamma-\beta(p)+\ln\frac{\beta(p)}{\gamma}+\ln\frac{\beta(p)}{\gamma}\ln(\beta(p)\gamma)\right] \\
&\quad + \mathbb{I}_{\{p>p_0\}}\left[(2-b)(b-\beta(p))+\beta(p)(2b-3-2\ln b)\ln\frac{b}{\beta(p)}\right. \\
&\qquad\qquad \left.+\beta(p)(\gamma-b)+\beta(p)\ln\frac{b}{\gamma}+\beta(p)((\ln b)^2-(\ln\gamma)^2)\right] \\
&\quad + p\beta(p)\ln\frac{\alpha(p)}{\gamma}+(1-p)\beta(p)(\alpha(p)-\gamma)-(1-p)\beta(p)\ln\frac{\alpha(p)}{\gamma} \\
&\qquad + (1-p)\beta(p)\ln\frac{\gamma}{\alpha(p)}\ln(\alpha(p)\gamma) \\
&\quad + \mathbb{I}_{\{p\ge 0.5\}}\left[\alpha(p)\beta(p)\ln\frac{a}{\alpha(p)}+\frac{\beta(p)}{a}(2-3a+2\ln a)(a-\alpha(p))\right. \\
&\qquad\qquad \left.+2(1-p)\frac{\beta(p)}{a}(\alpha(p)\ln a-a\ln\alpha(p))\right] \\
&\quad + \alpha(p)\beta(p)\left[\frac{1}{a}-1+(2p-1)\ln a+2(1-p)-2(1-p)\frac{1}{a}\right. \\
&\qquad\qquad \left.-2(1-p)\frac{\ln a}{a}\right]
\end{aligned}$$

and

$$\begin{aligned}\mathbf{P}(p;\mathrm{BA}) = {}& \mathbb{I}_{\{p>=p_1\}}\left[c-\kappa(p)+\frac{\kappa(p)}{2}(1-(\ln\ell)^2)\right] \\ & \mathbb{I}_{\{p<p_1\}}[\kappa(p)((\ln\kappa(p))^2-(\ln\ell)^2)] \\ & +\kappa(p)\left[p\ln\frac{\nu(p)}{\ell}-\frac{(1-p)}{2}((\ln\nu(p))^2-(\ln\ell)^2)\right] \\ & +\kappa(p)\nu(p)\left[1-2p+\frac{2p-1}{\nu(p)}-\frac{(1-p)}{\nu(p)}\ln\nu(p)\right] \\ & +\kappa(p)\nu(p)\left[1-\frac{1+\ln\nu(p)}{\nu(p)}-p\left(1-\frac{1+\ln d}{d}\right)\right].\end{aligned}$$

The numerical comparison of the probabilities of success is given in Table 3 (see also Fig. 3).

REFERENCES

[1] K. Ano, *Bilateral secretary problem recognizing the maximum or the second maximum of a sequence*, J. Inform. Optim. Sci., 11 (1990), pp. 177–188.

[2] E. Dynkin and A. Yushkevich, *Theorems and Problems on Markov Processes*, Plenum Press, New York, 1969.

[3] R. Eidukjavicjus, *Optimalna ostanovka markovskoj cepi dvumia momentami ostanovki*, Inf. XIX conf. math., Litovsk. Mat. Sb., 19 (1979), pp. 181–183.

[4] E. Enns and E. Ferenstein, *On a multi-person time-sequential game with priorities*, Sequential Anal., 6 (1987), pp. 239–256.

[5] E. Ferenstein, *Two-person non-zero-sum games with priorities*, in Strategies for Sequential Search and Selection in Real Time, Proceedings of the AMS-IMS-SIAM Joint Summer Research Conferences held June 21–27, 1990, University of Massachusetts at Amherst, T. S. Ferguson and S. M. Samuels, eds., Contemporary Mathematics, vol. 125, 1992, American Mathematical Society, Providence, RI, pp. 119–133.

[6] T. Ferguson, *Who solved the secretary problem?*, Statist. Sci., 4 (1989), pp. 282–296.

[7] P. Freeman, *The secretary problem and its extensions: A review*, Internat. Statist. Rev., 51 (1983), pp. 189–206.

[8] J. Gilbert and F. Mosteller, *Recognizing the maximum of a sequence*, J. Amer. Statist. Assoc., 61 (1966), pp. 35–73.

[9] G. Haggstrom, *Optimal sequential procedures when more then one stop is required*, Ann. Math. Statist., 38 (1967), pp. 1618–1626.

[10] R. Luce and H. Raiffa, *Games and Decisions*, John Wiley, New York, 1957.

[11] A. Mucci, *Differential equations and optimal choice problem*, Ann. Statist., 1 (1973), pp. 104–113.

[12] J. Neveu, *Discrete-Parameter Martingales*, North-Holland, Amsterdam, 1975.

[13] T. Radzik and K. Szajowski, *Sequential games with random priority*, Sequential Anal., 9 (1990), pp. 361–377.

[14] G. Ravindran and K. Szajowski, *Non-zero sum game with priority as Dynkin's game*, Math. Japon., 37 (1992), pp. 401–413.

[15] J. Rose, *Twenty years of secretary problems: A survey of developments in the theory of optimal choice*, Management Studies, 1 (1982), pp. 53–64.

[16] M. Sakaguchi, *Sequential games with priority under expected value maximization*, Math. Japon., 36 (1991), pp. 545–562.

[17] A. Shiryaev, *Optimal Stopping Rules*, Springer-Verlag, New York, Heidelberg, Berlin, 1978.

[18] W. Stadje, *An optimal k-stopping problem for the Poisson process*, in Proc. 6th Pannonian Symp. on Math. Stat., Bad Tazmannsdorf, Austria, 1986, Comp. in Mathematical Statistics and Probability Theory vol. B., D. Reidel, Dordrecht, 1987, pp. 231–244.

[19] K. Szajowski, *Optimal choice problem of a-th object*, Mat. Stos., 19 (1982), pp. 51–65. (In Polish.)

[20] ———, *Double stop by two decision makers*, Adv. Appl. Probab., 25 (1993), pp. 438–452.

[21] M. Yasuda, *On a randomized strategy in Neveu's stopping problem*, Stochastic Process. Appl., 21 (1985), pp. 159–166.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 33, No. 5, pp. 1411 1424, September 1995

© 1995 Society for Industrial and Applied Mathematics

006

ON NONLINEAR OPTIMAL CONTROL PROBLEMS WITH STATE CONSTRAINTS*

MONICA MOTTA[†]

Abstract. This paper is concerned with an optimal control problem where the state is constrained to stay either in a smooth open set Ω or in its closure $\overline{\Omega}$. Under a "higher-order" sufficient condition for the viability of Ω and $\overline{\Omega}$, we prove that the optimal cost function v_Ω is the unique continuous constrained solution of the Hamilton–Jacobi–Bellman equation. Furthermore, we show that v_Ω coincides with the optimal cost function $v_{\overline{\Omega}}$ on Ω.

Key words. viscosity solutions, Hamilton–Jacobi–Bellman (HJB) equations, state–space constraints, Lie brackets

AMS subject classifications. 93C10, 49L25

1. Introduction.

This paper studies the value functions of optimal control problems where the state is constrained to stay either in a smooth open set Ω or in its closure $\overline{\Omega}$. We are interested in the continuity properties of the value functions v_Ω and $v_{\overline{\Omega}}$, respectively, in the Hamilton–Jacobi–Bellman (HJB) equation of dynamic programming they satisfy in some weak sense, and in proving that $v_\Omega = v_{\overline{\Omega}}$ on Ω.

These problems were first studied by Soner in [13] in the context of viscosity solutions (see [5], [6], [10]) under the assumption that at each point of $\partial\Omega$ there is a field of the system pointing inward Ω. Clearly this condition ensures the "viability" of the sets Ω and $\overline{\Omega}$ for the system, that is, the existence of controls that do not violate the constraints. Soner proved that $v_{\overline{\Omega}}$ is the unique viscosity solution of the HJB equation with a suitable new boundary condition. This notion of *constrained viscosity solution* was studied by Capuzzo-Dolcetta and P. L. Lions in [4] for more general problems and by Loreti and Tessitore in [12]. Loreti proved in [11], among other things, that $v_\Omega = v_{\overline{\Omega}}$ under Soner's assumption. The goal of this paper is to weaken this assumption and consider a "first-order" sufficient condition for the viability of Ω and $\overline{\Omega}$, that is, a condition involving the derivatives of the fields of the system. This is analogous to the application of "higher-order" controllability conditions to the study of the value function in time-optimal control problems, see, e.g., the survey paper of Stefani [14], Sussmann [15], and Evans and James [7] as well as Bardi and Soravia [1] for the connection with Hamilton–Jacobi equations.

Our condition is that, around the points of $\partial\Omega$ where Soner's assumption is not satisfied, the system is symmetric and there is a Lie bracket between two fields of the system pointing inward Ω.

More precisely, let Ω be an open subset of $\mathbb{R}^n$, let $b(\cdot,\cdot)$ be a continuous function from $\mathbb{R}^n \times U$ into $\mathbb{R}^n$, and let $\sigma(\cdot)$ be a C^1 function from $\mathbb{R}^n$ into $M_{n\times m}$, where $M_{n\times m}$ denotes the vector space of $n \times m$ matrices on $\mathbb{R}$. We consider the trajectories $y_x(\cdot, u)$ which are solutions of the controlled dynamical system

$$\begin{cases} y'(t) = b\big(y(t), u(t)\big), & t > 0, \\ y(0) = x, \end{cases} \tag{1.1}$$

* Received by the editors April 14, 1993; accepted for publication (in revised form) March 1, 1994.

[†] Dipartimento di Matematica Pura e Applicata, Università di Padova, via Belzoni 7, I–35131 Padova, Italy (motta@pdmat1.unipd.it).

which will be assumed to have the symmetric structure

$$\begin{cases} y'(t) = \sigma(y(t))u(t), & t > 0, \\ y(0) = x, \end{cases} \tag{1.2}$$

around certain points of the boundary, and where the control $u(\cdot)$ is measurable and takes values in a compact subset U of $\mathbb{R}^m$ symmetrical with respect to the origin. Let $\mathcal{A}_x^0$ and $\mathcal{A}_x$ be the sets of the controls under which $y_x(t,u)$ lies in Ω or, respectively, in $\overline{\Omega}$ (bar denotes the closure). Given a discounted cost associated to every control u and x in $\mathbb{R}^n$, we consider either $\mathcal{A}_x^0$ or $\mathcal{A}_x$ as the set of admissible controls, so that we obtain, respectively, the optimal value functions

$$v_\Omega(x) = \inf_{u\in\mathcal{A}_x^0} \int_0^{+\infty} e^{-t} f(y(t),u(t))\,dt, \qquad x \in \Omega, \tag{1.3}$$

or

$$v_{\overline{\Omega}}(x) = \inf_{u\in\mathcal{A}_x} \int_0^{+\infty} e^{-t} f(y(t),u(t))\,dt, \qquad x \in \overline{\Omega}, \tag{1.4}$$

by assuming $\mathcal{A}_x^0$ (and hence $\mathcal{A}_x$) $\neq \emptyset$.

The sufficient condition given in [13] to prove the continuity of $v_{\overline{\Omega}}$ follows:

$$\text{(S)} \qquad \forall x \in \partial\Omega \qquad \exists u(x) \in U: \quad b(x,u(x)) \cdot n(x) < 0,$$

where $n(x)$ is the exterior normal vector to Ω at x. Moreover, [13] introduces the definition of constrained viscosity solution of the HJB equation

$$\text{(HJB)} \qquad v + H(x,Dv) = 0 \qquad x \in \overline{\Omega},$$

where the Hamiltonian is given by

$$H(x,p) = \max_{u\in U}\{-b(x,u)\cdot p - f(x,u)\}, \tag{1.5}$$

that is, a subsolution in Ω which is supersolution in $\overline{\Omega}$, and [13] checks that, under condition (S), $v_{\overline{\Omega}}$ is the unique constrained viscosity solution of the HJB equation.

Here we replace (S) with the following assumption:

(A1) $\forall x \in \partial\Omega$ $\min_{u\in U} b(x,u)\cdot n(x) \leq 0$ and in case $\min_{u\in U} b(x,u)\cdot n(x) = 0$, there is a neighborhood of x where the system takes the form (1.2) and

$$\exists u_1(x), u_2(x) \in U \quad \text{satisfying} \quad [\sigma_1,\sigma_2](x)\cdot n(x) < 0, \tag{1.6}$$

where $\sigma_i(x) := \sigma(x)u_i(x)$ $(i = 1,2)$ and $[h,k]$ denotes the Lie bracket between $h,k \in C^1(\mathbb{R}^n,\mathbb{R}^n)$, i.e., $[h,k](x) := k'(x)h(x) - h'(x)k(x)$.

Note that under such an assumption, the set of all points $x \in \partial\Omega$ which fails to verify (S) has always empty interior, as it is observed in Remark 2.2.

Thus by assuming (A1) and some technical hypotheses stated in (A0) in §2, we prove that v_Ω coincides with $v_{\overline{\Omega}}$ in Ω. Furthermore we check that the value function v_Ω is (bounded and) uniformly continuous (briefly, $v_\Omega \in BUC(\Omega)$), so that it has unique continuous extension $\bar{v}$ on $\overline{\Omega}$, and we characterize $\bar{v}$ as the unique continuous constrained viscosity solution of the HJB equation, therefore extending the main result of [13]. Here we need some results in the theory of discontinuous viscosity solution, see Barles and Perthame [3], Ishii [9], and Bardi and Soravia [2]. Finally by means of an example we show that, unlike assumption (S), (A1) does not guarantee the continuity of $v_{\overline{\Omega}}$ on $\partial\Omega$.

The paper is organized as follows. In §2 we lay down the hypotheses. In §3 it is shown that $v_\Omega \in BUC(\Omega)$. and in §4 we check that $v_\Omega = v_{\overline{\Omega}}$ in Ω. In §5 we prove that $\bar{v}$ is the unique constrained solution of the HJB equation, and in §6 we give an example where $v_{\overline{\Omega}}$ is discontinuous on $\partial\Omega$.

2. Statement of the problem. Throughout this paper we will use (A1) and the following assumptions:

(A0) (i) Ω is a connected, open subset of $\mathbb{R}^n$ with a C^2 compact boundary $\partial\Omega$;

(ii) the controls take values in a compact symmetrical subset of $\mathbb{R}^m$, U;

(iii) $\sigma \in C^1(\mathbb{R}^n, M_{n\times m})$ is bounded and Lipschitz continuous;

(iv) $f \in C(\mathbb{R}^n \times U, \mathbb{R})$ and $b \in C(\mathbb{R}^n \times U, \mathbb{R}^n)$ are bounded and Lipschitz continuous in x, uniformly w.r.t. U, i.e.,

$$(2.1)\qquad \begin{array}{ll} |f(x,u) - f(y,u)| \le L(f)|x-y|, & |f(x,u)| \le K(f) \\ |b(x,u) - b(y,u)| \le L(b)|x-y|, & |b(x,u)| \le K(b) \end{array} \qquad \forall x,y \in \mathbb{R}^n,\ \forall u \in U.$$

For any Lipschitz-continuous (possibly in x, uniformly w.r.t. U) and bounded function g, we call $L(g)$ and $K(g)$ the Lipschitz constant and the bound of g, respectively.

Consider $\mathcal{A}$, the set of all measurable maps of $[0,+\infty)$ into U. For any $u \in \mathcal{A}$ and $x \in \overline{\Omega}$, let $y_x(\cdot,u)$ be the solution of (1.1) with initial data $y_x(0,u) = x$. The associated payoff is

$$(2.2)\qquad J(x,u) = \int_0^{+\infty} e^{-t} f\big(y(t),u(t)\big)\,dt.$$

We allow only the controls which leave $y_x(\cdot,u)$ in Ω (or in $\overline{\Omega}$); thus the sets of admissible controls are

$$\mathcal{A}_x^0 := \big\{u \in \mathcal{A} :\ y_x(t,u) \in \Omega\ \forall t \ge 0\big\} \quad \text{for any } x \in \Omega$$

and

$$\mathcal{A}_x := \big\{u \in \mathcal{A} :\ y_x(t,u) \in \overline{\Omega}\ \forall t \ge 0\big\} \quad \text{for any } x \in \overline{\Omega},$$

respectively. We suppose $\mathcal{A}_x^0 \neq \emptyset$, hence both the optimal value functions v_Ω and $v_{\overline{\Omega}}$ defined in §1 are bounded. Setting $J_t(x,u) := \int_0^t e^{-s} f\big(y(s),u(s)\big)\,ds$ the dynamic programming principle (DPP) assumes for every $t > 0$ the form

$$(2.3)\qquad v_\Omega(x) = \inf_{u\in\mathcal{A}_x^0} \big\{J_t(x,u) + e^{-t} v_\Omega(y(t))\big\} \qquad \forall x \in \Omega$$

and

$$(2.4)\qquad v_{\overline{\Omega}}(x) = \inf_{u\in\mathcal{A}_x} \big\{J_t(x,u) + e^{-t} v_{\overline{\Omega}}(y(t))\big\} \qquad \forall x \in \overline{\Omega},$$

respectively.

Remark 2.1. For the sake of simplicity we assume that the boundary $\partial\Omega$ is compact and that both the functions $b(\cdot,U)$ and $\sigma(\cdot)$ are bounded and Lipschitz continuous, but we could obtain the same results by using a local version of the techniques indicated in this paper.

Remark 2.2. The assumption (A1) implies that the set of the points $x \in \partial\Omega$ for which the Soner's condition (S) fails, i.e., where $\min_{u\in U} b(x,u)\cdot n(x) = 0$, has empty interior.

Proof. Assume (A1) and let us suppose that there exists an open subset Λ of $\mathbb{R}^n$ such that $\partial\Lambda := \Lambda \cap \partial\Omega \neq \emptyset$ and for all $x \in \partial\Lambda$, one has $\min_{u\in U} b(x,u)\cdot n(x) = 0$; hence we can assume that the system has the form (1.2) on Λ. Then we find a contradiction in that for any pair $u_1, u_2 \in U$, the Lie bracket $[\sigma_1, \sigma_2](x)$, where $\sigma_i(x) = \sigma(x)u_i$ $(i = 1, 2)$, is tangent to Ω in x. Indeed, since under our assumption $\forall u \in \mathcal{A}$ the corresponding trajectory $y_x(t, u)$ is in $\partial\Omega$ for all time t small enough, by standard properties of the Lie brackets we have $[\sigma_1, \sigma_2](x) \cdot n(x) = 0$ for all $u_1, u_2 \in U$. Hence (1.6) fails to hold, and that contradicts assumption (A1). □

3. Uniform continuity of the value function v_Ω on Ω.

In this section we prove that the value function v_Ω belongs to $BUC(\Omega)$. First, let us note that, if u is an admissible control for $x_0 \in \Omega$ which leaves $y_{x_0}(t, u)$ either in $\overline{\Omega}$ or in Ω $\forall t > 0$, then u is not necessarily admissible at any point x, regardless of how close that point is to x_0. The following lemma shows a way to modify it, keeping $y_x(\cdot, u)$ in Ω at least on the interval $[0, t^*]$ with $t^* > 0$, independent of x_0 and x, and changing the cost proportionally to $|x - x_0|^{1/2}$.

LEMMA 3.1. *Assume* (A0), (A1). *Then there are some* $t^* > 0$ *and* $L > 0$ *such that for any* $x \in \Omega$ *and* $u \in \mathcal{A}$ *there exists some constant* $\alpha_x > 0$ *such that for all* $\alpha \in [0, \alpha_x]$ *one can determine a control* $\bar{u} \in \mathcal{A}$ *satisfying*

(i) $y_x(t, \bar{u}) \in \overline{\Omega}_\alpha$ $\forall t \in [0, t^*]$;

(ii) $|J_{t^*}(x, \bar{u}) - J_{t^*}(x, u)| \le L\sqrt{\tilde{d}_\alpha}$,

where

$$\Omega_\alpha := \{y \in \Omega : \ \mathrm{dist}(y, \partial\Omega) > \alpha\} \tag{3.1}$$

and

$$\tilde{d}_\alpha := \sup\{\mathrm{dist}(y_x(t, u), \Omega_\alpha) : \quad t \in [0, t^*]\}. \tag{3.2}$$

Proof. First of all, let us observe that under the hypotheses on Ω and $\partial\Omega$ in (A0), there exists some $\hat{\alpha} > 0$ such that for every $\alpha \in [0, \hat{\alpha}]$ the set Ω_α is a nonempty, connected open subset of $\mathbb{R}^n$ with C^2 compact boundary $\partial\Omega_\alpha$. Moreover, the signed distances $d(\cdot) := d(\cdot, \partial\Omega)$ and $d_\alpha(\cdot) := d(\cdot, \partial\Omega_\alpha)$ verify the simple relation

$$d_\alpha(y) = d(y) + \alpha \qquad \forall y \in \mathbb{R}^n. \tag{3.3}$$

The proof of (3.3) follows quite easily, e.g., by the ε-neighborhood theorem in §3 of [8].

Furthermore, we note that $\forall x \in \partial\Omega$ there are $\delta(x) > 0$ and $\xi_0(x) > 0$ such that the signed distance $d(\cdot) = d(\cdot, \partial\Omega)$ can be redefined as a C^2 and negative function inside Ω, verifying, by (A1), either

$$b(y, u(x)) \cdot d'(y) \le -\xi_0(x) < 0 \qquad \forall y \in B(x, \delta(x)), \tag{3.4}$$

if $\min_{u \in U} b(x, u) \cdot n(x) < 0$ and this minimum is attained at $u = u(x)$ or

$$\left.\begin{aligned} b(y, u) &= \sigma(y)u, \\ [\sigma_1, \sigma_2](y) \cdot d'(y) &\le -32\xi_0(x) < 0 \end{aligned}\right. \qquad \forall y \in B(x, \delta(x)), \tag{3.5}$$

with $\sigma_i(y) = \sigma(y)u_i(x)$ $(i = 1, 2)$, if $\min_{u \in U} b(x, u) \cdot n(x) = 0$. By the compactness of $\partial\Omega$, there exist $x_1, \dots, x_R \in \partial\Omega$ such that $\{B(x_j, \delta_j/4) \ : \ j = 1, \dots, R\}$ with $\delta_j = \delta(x_j)$ as an open covering of $\partial\Omega$. By setting

$$\delta := \min\{\delta_j/4 : \ j = 1, \dots, R\}, \quad \xi_0 := \min\{\xi_0(x_j) : \ j = 1, \dots, R\},$$
$$I := \left\{j \in \{1, \dots, R\} : \ \min_{u \in U} b(x_j, u) \cdot n(x_j) = 0\right\},$$

one has for all $x \in B(\partial\Omega, \delta) := \{y \in \mathbb{R}^n \ : \ \operatorname{dist}(y, \partial\Omega) < \delta\}$:

$$\begin{aligned} x &\in B(x_j, \delta_j/2) \quad \text{for some } j, \\ y_x(t,u) &\in B(x_j, \delta_j) \qquad \forall t \in [0, \delta/K(b)] \text{ and } \forall u \in \mathcal{A}. \end{aligned} \tag{3.6}$$

Furthermore, if $j \in \{1, \ldots, R\} \setminus I$, one has

$$b\big(y_x(t,u), u(x_j)\big) \cdot d'\big(y_x(t,u)\big) \le -\xi_0 < 0 \qquad \forall t \in [0, \delta/K(b)] \text{ and } \forall u \in \mathcal{A}; \tag{3.7}$$

otherwise, i.e., if $j \in I$,

$$[\sigma_1, \sigma_2]\big(y_x(t,u)\big) \cdot d'\big(y_x(t,u)\big) \le -32\xi_0 < 0 \qquad \forall t \in [0, \delta/K(b)] \text{ and } \forall u \in \mathcal{A}, \tag{3.8}$$

with $\sigma_i(y) = \sigma(y) u_i(x_j)$ (and $i = 1, 2$) holds.

For any $x \in \Omega$ let us set

$$\alpha_x := \frac{|d(x)| \wedge \delta}{2}, \tag{3.9}$$

where $\forall \nu, \mu \in \mathbb{R}$, $\nu \wedge \mu$ denotes the $\min\{\nu, \mu\}$. Note that, by decreasing δ if necessary, we can assume $\delta \le \hat{\alpha}$, hence $\forall \alpha \in [0, \alpha_x]$, (3.3) holds.

Let $x \in \Omega_\delta$. Then $B(x, \delta/2) \subset \Omega_{\delta/2}$, hence by (3.9) and by a standard estimate it follows that for all $\alpha \in [0, \alpha_x]$, $y_x(t,u) \in \overline{\Omega}_\alpha$ $\forall t \in [0, t^*]$ and $\forall u \in \mathcal{A}$ if $t^* = \delta/2K(b)$. Thus both (i) and (ii) hold by choosing $\bar{u} \equiv u$.

Let $x \in \Omega \cap B(\partial\Omega, \delta)$, so that $x \in B(x_j, \delta_j/2)$ for some $j \in \{1, \ldots, R\}$ and either $j \in \{1, \ldots, R\} \setminus I$, or $j \in I$. Since for all $\alpha \in [0, \alpha_x]$ the signed distance $d_\alpha(\cdot) = d(\cdot, \partial\Omega_\alpha)$ satisfies (3.3), in the first case Soner's condition (S) holds on $\partial\Omega_\alpha \cap B(x, \delta)$ $\forall \alpha \in [0, \alpha_x]$, see (3.7). Thus, analogous to Lemma 3.2 of [13], one can verify that there exist some constants $t^* > 0$ and $L > 0$, independent of x and α, such that $\forall u \in \mathcal{A}$, the control $\bar{u}$ defined by

$$\bar{u}(t) := u(t)\chi_{[0,t_0)}(t) + u(x_j)\chi_{[t_0, t_0 + k\tilde{d}_\alpha]}(t) + u(t - k\tilde{d}_\alpha)\chi_{(t_0 + k\tilde{d}_\alpha, +\infty)}(t), \tag{3.10}$$

with $\tilde{d}_\alpha$ in (3.2), and

$$k := \frac{4}{\xi_0}, \qquad t_0 := \inf\{t \in (0, t^*] \ : \ y_x(t,u) \in \partial\Omega_\alpha\},$$
$$\chi_{\mathcal{I}}(t) := \begin{cases} 1 & t \in \mathcal{I} \\ 0 & t \in \mathbb{R} \setminus \mathcal{I} \end{cases} \qquad \forall \mathcal{I} \subset \mathbb{R},$$

satisfies both (i) and (ii), for we can assume $\tilde{d}_\alpha \le 1$ so that $\tilde{d}_\alpha \le \sqrt{\tilde{d}_\alpha}$ $\forall \alpha \in [0, \alpha_x]$. Otherwise, if $j \in I$, for a fixed $t^* \in (0, \delta/K(b)]$, we set

$$\alpha_i(t) := u_1\big(\chi_{[0, t_i/4)}(t) - \chi_{[t_i/2, 3t_i/4)}(t)\big) + u_2\big(\chi_{[t_i/4, t_i/2)}(t) - \chi_{[3t_i/4, t_i]}(t)\big), \tag{3.11}$$

where $t_i \ge 0$ will be chosen later, $u_h = u_h(x_j)$ $(h = 1, 2)$, and we define $T_0 := 0$, $T_k := \sum_{i=1}^k t_i$. Let some $u \in \mathcal{A}$ and $\alpha \in [0, \alpha_x]$ be fixed. We claim that the control

$$\bar{u}(t) := \sum_{i=1}^{N} \alpha_i(t - T_{i-1})\chi_{[T_{i-1}, T_i)}(t) + u(t - T_N)\chi_{[T_N, +\infty)}(t), \tag{3.12}$$

by choosing convenient values of $t^* > 0$, $t_i = t_i(x,\alpha)$ and $N = N(x,\alpha)$, reaches our goal. We set $y_x(\cdot) := y_x(\cdot,u)$ and $\bar{y}_x(\cdot) := y_x(\cdot,\bar{u})$. Then for $t \geq 0$ we get the estimate

$$\begin{aligned}(3.13)\quad d\big(\bar{y}_x(t+T_N)\big) &= d\big(\bar{y}_x(T_N)\big) + \int_0^t d'\big(\bar{y}_x(s+T_N)\big)\cdot\sigma\big(\bar{y}_x(s+T_N)\big)u(s)\,ds \\ &\leq d\big(\bar{y}_x(T_N)\big) + d\big(y_x(t)\big) - d(x) + \big(L(b)+K(b)\big)\big|\bar{y}_x(T_N)-x\big|L(b)^{-1}\big(e^{L(b)t}-1\big),\end{aligned}$$

by using

$$(3.14)\qquad \big|\bar{y}_x(s+T_N) - y_x(s)\big| \leq \big|\bar{y}_x(T_N) - x\big|e^{L(b)s},$$

which follows by Gronwall's lemma. Furthermore, by $\bar{u}$'s definition (3.12), one has for $k = 1,\ldots,N$

$$\bar{y}_x(T_k) = \bar{y}_x(T_{k-1}) + [\sigma_1,\sigma_2]\big(\bar{y}_x(T_{k-1})\big)\frac{t_k^2}{16} + o(t_k^2),$$

which implies

$$(3.15)\qquad \bar{y}_x(T_k) = x + \sum_{i=1}^{k}[\sigma_1,\sigma_2]\big(\bar{y}_x(T_{i-1})\big)\frac{t_i^2}{16} + \sum_{i=1}^{k}o(t_i^2),$$

and therefore

$$(3.16)\qquad \big|\bar{y}_x(T_N) - x\big| \leq C_1\sum_{i=1}^{N}t_i^2 \qquad \text{with } C_1 > 0 \text{ independent on } x \text{ and } \alpha.$$

By $d(\cdot)$'s expansion around x we obtain

$$d\big(\bar{y}_x(T_k)\big) = d(x) + d'(x)\left\{\sum_{i=1}^{k}[\sigma_1,\sigma_2]\big(\bar{y}_x(T_{i-1})\big)\frac{t_i^2}{16}\right\} + \sum_{i=1}^{k}o(t_i^2),$$

which yields

$$(3.17)\qquad d\big(\bar{y}_x(T_k)\big) \leq d(x) - \xi_0\sum_{i=1}^{k}t_i^2 \qquad \text{for } k = 1,\ldots,N,$$

if $t_i \leq T$ (for a convenient $T > 0$ independent of x and α) for all $i = 1,\ldots,N$ and $T_{N-1} \leq \delta/K(b)$, by (3.6), (3.8), and $d'(\cdot)$'s continuity. Applying these results to the estimate (3.13) and recalling the relation (3.3) between $d(\cdot)$ and $d_\alpha(\cdot)$, we get

$$d_\alpha\big(\bar{y}_x(t+T_N)\big) \leq d_\alpha\big(y_x(t)\big) - \xi_0\sum_{i=1}^{N}t_i^2 + C_2\big(e^{L(b)t}-1\big)\sum_{i=1}^{N}t_i^2$$

for all $t \geq 0$ and for some $C_2 > 0$, i.e., by the definition (3.2) of $\tilde{d}_\alpha$

$$(3.18)\qquad d_\alpha\big(\bar{y}_x(t+T_N)\big) \leq \tilde{d}_\alpha\left(1 - \frac{\xi_0}{2}\sum_{i=1}^{N}k_i\right) \qquad \forall t \in [0, t^* - T_N],$$

if we assume $t^* \leq \frac{\delta}{K(b)} \wedge \frac{1}{L(b)}\ln\big(1 + \frac{\xi_0}{2C_2}\big)$ and

$$(3.19)\qquad t_i := \sqrt{\tilde{d}_\alpha k_i} \qquad \text{for } i = 1,\ldots,N,$$

where $k_1,\ldots,k_N$ are to be chosen (generally depending on x and α). As we will show, verifying theses (i) and (ii) is equivalent to checking that there exist N, $k_1,\ldots,k_N$

such that, first, the estimate (3.17) holds for $t_1, \dots, t_N$ given by (3.19), the following relations

$$\sum_{i=1}^{N} k_i = \frac{4}{\xi_0}, \tag{3.20}$$

and

$$\sum_{i=1}^{N} \sqrt{k_i} \leq C_3 \qquad \text{for some } C_3 \text{ independent of } x \text{ (and } \alpha\text{)}, \tag{3.21}$$

hold; second, $\bar{y}_x(t) \in \overline{\Omega}_\alpha$ for all $t \in [0, T_N]$. Note that the estimate (3.17), decreasing t^* if necessary, is verified. It is important to remark that it is not restrictive to assume

$$d'(y_k) \cdot \sigma(y_k) u_1 \leq 0 \quad \text{and} \quad d'(y_k) \cdot \sigma(y_k) u_2 \leq 0, \tag{3.22}$$

where $y_k := \bar{y}_x(T_k)$ and $k = 0, \dots, N$. Indeed by using the identities $[\sigma_1, \sigma_2] = [-\sigma_2, \sigma_1] = [\sigma_2, -\sigma_1] = [-\sigma_2, -\sigma_1]$, we can modify in the definition (3.11) of α_i with $i = k$ the order of u_1, u_2 and their opposites in a convenient way, so that both (3.22) and (3.15) are true. Then we claim that both

$$d(\bar{y}_x(t)) \leq d(x) + C_4 t^2 \qquad \forall t \in [0, t_1], \tag{3.23}$$

and

$$\begin{aligned} d(\bar{y}_x(t)) &\leq d(\bar{y}_x(T_{k-1})) + C_4 (t - T_{k-1})^2 \\ &\leq d(x) - \xi_0 \sum_{i=1}^{k-1} t_i^2 + C_4 t_k^2 \quad \forall t \in [T_{k-1}, T_k], \end{aligned} \tag{3.24}$$

for $k = 2, \dots, N$ hold. In fact,

$$\begin{aligned} d(\bar{y}_x(t)) &= d(x) + \int_0^t d'(\bar{y}_x(s)) \sigma_1(\bar{y}_x(s))\, ds \\ &\leq d(x) + t d'(x) \sigma_1(x) + L(d'\sigma) \int_0^t |\bar{y}_x(s) - x|\, ds \\ &\underset{(3.22)_1}{\leq} d(x) + C_4' t^2 \qquad \forall t \in \left[0, \frac{t_1}{4}\right]; \\ d(\bar{y}_x(t)) &= d\left(\bar{y}_x\left(\frac{t_1}{4}\right)\right) + \int_{t_1/4}^t d'(\bar{y}_x(s)) \sigma_2(\bar{y}_x(s))\, ds \\ &\leq C_4' \left(\frac{t_1}{4}\right)^2 + \left(t - \frac{t_1}{4}\right) d'(x) \sigma_2(x) + C_4' \left(t - \frac{t_1}{4}\right)^2 + d(x) \\ &\underset{(3.22)_2}{\leq} d(x) + C_4' t^2 \qquad \forall t \in \left[\frac{t_1}{4}, \frac{t_1}{2}\right]; \\ \bar{y}_x(t) &= \bar{y}_x\left(\frac{t_1}{2}\right) - \int_{t_1/2}^t \{\sigma_1(x) + \sigma_1'(x)(\bar{y}_x(s) - x)\}\, ds + o\left(\left(t - \frac{t_1}{2}\right)^2\right) \\ &= x + \frac{t_1}{4}(\sigma_1(x) + \sigma_2(x)) + \frac{1}{2}\left(\frac{t_1}{4}\right)^2 [\sigma_2' \sigma_2 + 2\sigma_2' \sigma_1 + \sigma_1 \sigma_2'] + o(t_1^2) \end{aligned}$$

$$- \left(t - \frac{t_1}{2}\right)\sigma_1(x) - \sigma_1'(x)\int_{t_1/2}^{t} (\bar{y}_x(s) - x)\, ds + o\left(\left(t - \frac{t_1}{2}\right)^2\right)$$
$$\forall t \in \left[\frac{t_1}{2}, \frac{3t_1}{4}\right],$$

and expanding $d(\cdot)$ around x we get

$$d(\bar{y}_x(t)) \leq d(x) + d'(x)\left[\left(\frac{3t_1}{4} - t\right)\sigma_1(x) + \frac{t_1}{4}\sigma_2(x)\right] + C_4''\left[\left(\frac{t_1}{2}\right)^2 + \left(t - \frac{t_1}{2}\right)^2\right]$$
$$\underset{(3.22)}{\leq} d(x) + C_4''t^2 \qquad \forall t \in \left[\frac{t_1}{2}, \frac{3t_1}{4}\right].$$

Furthermore, we have

$$\bar{y}_x(t) = \bar{y}_x\left(\frac{3t_1}{4}\right) - \int_0^t \{\sigma_2(x) + \sigma_2'(x)[\bar{y}_x(s) - x]\}\, ds + o\left(\left(t - \frac{3t_1}{4}\right)^2\right)$$
$$= x + (t_1 - t)\sigma_2(x) + \frac{1}{2}\left(\frac{t_1}{4}\right)^2 (2[\sigma_1, \sigma_2] - \sigma_2'\sigma_1) - \sigma_2'(x)\int_0^t [\bar{y}_x(s) - x]\, ds$$
$$+ o(t_1^2) + o\left(\left(t - \frac{3t_1}{4}\right)^2\right) \qquad \forall t \in \left[\frac{3t_1}{4}, t_1\right],$$

which yields, as before, $d(\bar{y}_x(t)) \leq d(x) + C_4''t^2$ for all $t \in [3t_1/4, t_1]$. Now, by replacing x with y_k for $k = 1, \dots, N$ and choosing $C_4 := \max\{C_4', C_4''\}$ we can conclude that (3.23) and (3.24) hold.

By estimate (3.23), to keep $\bar{y}_x(t)$ in $\overline{\Omega}_\alpha$ for all $t \in [0, t_1]$, it suffices to have

$$k_1 := \frac{4}{\xi_0} \wedge \beta, \qquad \text{with } \beta := \frac{|d_\alpha(x)|}{2C_4\tilde{d}_\alpha}. \tag{3.25}$$

If $4/\xi_0 \leq \beta$, choosing $N = 1$ we have (3.20) and (3.21); otherwise, i.e., if $\beta \in (0, 4/\xi_0)$, we set for $r > 1$

$$k_r := \left(\frac{4}{\xi_0} - \sum_{i=1}^{r-1} k_i\right) \wedge \frac{\xi_0}{C_4}\sum_{i=1}^{r-1} k_i. \tag{3.26}$$

By the estimate (3.24), (3.26) implies that $\bar{y}_x(t) \in \overline{\Omega}_\alpha$ for all $t \in [0, T_r]$. As long as $\sum_{i=1}^r k_i < 4/\xi_0$, k_r is defined by $(3.26)_2$, thus

$$\frac{k_r}{k_{r-1}} = \frac{\xi_0}{C_4}\frac{k_1 + \cdots + k_{r-1}}{k_{r-1}} = 1 + \frac{\xi_0}{C_4} =: a,$$

i.e., we get

$$k_r = a^{r-1}\beta, \qquad \text{where } a > 1 \text{ and } r \geq 1. \tag{3.27}$$

The geometric series $\sum_{r=1}^{+\infty} k_r$ diverges, hence there exists some $N(x, \alpha) \in \mathbb{N}$ such that $\sum_{r=1}^{N} k_r \geq 4/\xi_0$; moreover, we can determine an upper bound $\overline{N}$ for N:

$$\sum_{r=1}^{N} a^{r-1}\beta = \beta\frac{a^N - 1}{a - 1} \geq \frac{4}{\xi_0} \quad \text{for } a^N - 1 \geq \frac{4}{\xi_0}(a - 1)\beta^{-1},$$

i.e.,

$$\overline{N} := \left[\log_a\left(1 + \frac{4}{\xi_0}(a-1)\beta^{-1}\right)\right] + 1, \tag{3.28}$$

where $[\ \cdot\ ]$ represents the integer part. The values $k_1, \dots, k_N$ satisfy (3.20) by construction, hence it remains to check that (3.21) is also true. By (3.28) it follows that

$$\begin{aligned}\sum_{r=1}^{N} \sqrt{k_r} = \sqrt{\beta}\sum_{r=1}^{N}(\sqrt{a})^{r-1} &= \sqrt{\beta}\,\frac{a^{N/2}-1}{\sqrt{a}-1}\\ &\le g_a(\beta) := \frac{\sqrt{a}}{\sqrt{a}-1}\left(\sqrt{\beta + \frac{4}{\xi_0}(a-1)} - \frac{\sqrt{\beta}}{\sqrt{a}}\right),\end{aligned}$$

where

$$\sup_{\beta\in(0,4/\xi_0)} g_a(\beta) = g_a(0) = \frac{2\sqrt{a}}{\sqrt{a}-1}\sqrt{\frac{a-1}{\xi_0}}.$$

At this point, the part (i) of the lemma is proved for all $x \in \Omega$, $\alpha \in [0, \alpha_x]$, and $u \in \mathcal{A}$.

Part (ii) follows easily from (3.14), (3.16), and (3.21). ▯

This technical lemma allows us to obtain the following continuity result.

THEOREM 3.1. *Under assumptions* (A0), (A1), *the value function* v_Ω *belongs to* $BUC(\Omega)$.

Proof. Let x, z be in Ω and $|x - z| < r$ (where $r < 1$). For any $\rho > 0$ by the DPP (2.3) and choosing t^* as in Lemma 3.1, there exists a control $u \in \mathcal{A}_z^0$ such that

$$J_{t^*}(z, u) + e^{-t^*} v_\Omega\left[y_z(t^*)\right] \le v_\Omega(z) + \rho,$$

with $y_z(\cdot) = y_z(\cdot, u)$, holds. Since $u \in \mathcal{A}_z^0$, there is a constant $\bar{\alpha} > 0$ such that $y_z(t) \in \Omega_\alpha$ for all $\alpha \in [0, \bar{\alpha}]$. For any x, choose $\alpha_1 := \alpha_x \wedge \bar{\alpha}$ and let $\bar{u}$ be the control defined in Lemma 3.1, let $\bar{y}_x$ be the corresponding trajectory, and recall that $\tilde{d}_{\alpha_1} = \sup\left\{\operatorname{dist}\left(y_x(t), \overline{\Omega}_{\alpha_1}\right) : t \in [0, t^*]\right\}$, with $y_x(\cdot) = y_x(\cdot, u)$. By standard estimates, $\tilde{d}_{\alpha_1} < A_1 r$ for some $A_1 > 0$. Thus we have

$$\left|J_{t^*}(x, \bar{u}) - J_{t^*}(x, u)\right| \le L\sqrt{A_1 r}$$

and the definition of $\bar{u}$ implies

$$\left|y_x(t^*) - \bar{y}_x(t^*)\right| \le A_2\sqrt{r}, \quad A_2 > 0.$$

Therefore we get

$$\left|\bar{y}_x(t^*) - y_z(t^*)\right| \le A_3\sqrt{r}, \quad A_3 > 0, \tag{3.29}$$

and

$$\left|J_{t^*}(x, \bar{u}) - J_{t^*}(z, u)\right| \le L\sqrt{A_1 r} + \left|J_{t^*}(x, u) - J_{t^*}(z, u)\right| \le A_4\sqrt{r}, \quad A_4 > 0. \tag{3.30}$$

Now we set

$$\omega(r) := \sup\left\{\left|v_\Omega(x) - v_\Omega(z)\right| : \quad x, z \in \Omega \text{ and } |x - z| < r\right\}.$$

Combining these relations and using the DPP (2.3) we have

$$\begin{aligned}v_\Omega(x) - v_\Omega(z) &\le J_{t^*}(x, \bar{u}) - J_{t^*}(z, u) + e^{-t^*}\left[v_\Omega\left(\bar{y}_x(t^*)\right) - v_\Omega\left(y_z(t^*)\right)\right] + \rho\\ &\le A_4\sqrt{r} + e^{-t^*}\omega(A_3\sqrt{r}) + \rho \quad \forall\rho > 0,\end{aligned} \tag{3.31}$$

i.e.,

$$\omega(r) \le A_4\sqrt{r} + e^{-t^*}\omega(A_3\sqrt{r}) \quad \text{where } A_3 > 1 \text{ and } A_4 > 0. \tag{3.32}$$

Since ω is nondecreasing on $(0, +\infty)$, the limit $\omega(+0) \in [0, +\infty)$ exists. By (3.32), we have $\omega(0+) \le e^{-t^*}\omega(0+)$ and conclude that $\omega(0+) = 0$. □

4. Relation between the value functions $v_{\overline{\Omega}}$ and v_Ω. In this section we prove that the value functions $v_{\overline{\Omega}}$ and v_Ω coincide on Ω.

THEOREM 4.1. *Assume* (A0) *and* (A1). *Then*

$$v_\Omega(x) = v_{\overline{\Omega}}(x) \qquad \forall x \in \Omega.$$

Proof. Since $\mathcal{A}_x^0 \subset \mathcal{A}_x$, the inequality $v_\Omega(x) \ge v_{\overline{\Omega}}(x)$ holds by definition. Therefore it is enough to show that $v_\Omega(x) \le v_{\overline{\Omega}}(x)$ $\forall x \in \Omega$. Let $x \in \Omega$ and $\varepsilon > 0$ be fixed. Then there exist some $u \in \mathcal{A}_x$ such that

$$\int_0^{+\infty} e^{-t} f\big(y_x(t), u(t)\big)\, dt \le v_{\overline{\Omega}}(x) + \varepsilon,$$

and some $T > 0$ such that

$$\left| \int_T^{+\infty} e^{-t} f\big(\tilde{y}_x(t), \tilde{u}(t)\big)\, dt \right| \le \varepsilon \quad \forall \tilde{u} \in \mathcal{A},$$

if $y_x(\cdot)$ and $\tilde{y}_x(\cdot)$ denote $y_x(\cdot, u)$ and $y_x(\cdot, \tilde{u})$, respectively. Therefore the inequality

$$v_\Omega(x) - v_{\overline{\Omega}}(x) \le \int_0^T e^{-t}\big\{f\big(\tilde{y}_x(t), \tilde{u}(t)\big) - f\big(y_x(t), u(t)\big)\big\}\, dt + 3\varepsilon \tag{4.1}$$

holds for any $\tilde{u} \in \mathcal{A}_x^0$. Now let us consider some $\alpha \in (0, \alpha_x]$ $(\alpha < 1)$, where α_x is given by Lemma 3.1. By this lemma, there is a control $\tilde{u}_1 \in \mathcal{A}$ satisfying

$$y_x(t, \tilde{u}_1) \in \overline{\Omega}_\alpha \qquad \forall t \in [0, t^*],$$

and (since u is in $\mathcal{A}_x$)

$$\big|J_{t^*}(x, \tilde{u}_1) - J_{t^*}(x, u)\big| \le L\sqrt{\alpha}.$$

Now, if $t^* \ge T$ by choosing $\alpha \le \varepsilon^2/L^2$ and using the estimate (4.1), we can conclude that $v_\Omega(x) - v_{\hat{\Omega}}(x) \le 0$ by the arbitrariness of ε. Otherwise, we set

$$x_1 := y_x(t^*), \qquad \tilde{x}_1 := y_x(t^*, \tilde{u}_1),$$

and call u^{t^*} the control $u^{t^*}(t) = u(t + t^*)$ for all $t \ge 0$. Since $\tilde{x}_1 \in \overline{\Omega}_\alpha$ and $u^{t^*} \in \mathcal{A}_{x_1}$ we can again apply Lemma 3.1 with α replaced by $\alpha/2$ (see the definition (3.9) of α_x) to obtain a control $\bar{u}_2$ analogous to $\tilde{u}_1$, but starting from $\tilde{x}_1$ instead of x. Let M be $[T/t^*] + 1$. For $m \in \{1, \dots, M - 1\}$ we set recursively

$$x_m := y_x(mt^*), \qquad \tilde{x}_m := y_x(mt^*, \tilde{u}_m),$$

where any $\bar{u}_m$ for $m \ge 2$ is determined on the basis of Lemma 3.1 applied to $\tilde{x}_{m-1} \in \overline{\Omega}_{-\alpha/2^{m-2}}$ with α replaced by $\alpha/2^{m-1}$ and $u^{(m-1)t^*} \in \mathcal{A}_{x_{m-1}}$; and $\tilde{u}_m$ is defined by

$$\tilde{u}_m(t) := \tilde{u}_{m-1}(t)\chi_{[0,(m-1)t^*]}(t) + \bar{u}_m(t - (m-1)t^*)\chi_{((m-1)t^*, +\infty)}(t).$$

We claim that for any $m \in \{1, \dots, M - 1\}$

$$|\tilde{x}_m - x_m| \le a\alpha^{2^{-m}} \quad \text{for some } a > 0, \tag{4.2}$$

and

$$(4.3) \qquad \left|J_{(m+1)t^*}(x, \tilde{u}_{m+1}) - J_{(m+1)t^*}(x, u)\right| \leq L_1 \alpha^{2^{-(m+1)}} \quad (L_1 \geq L),$$

hold. For $m = 1$ they follow by Lemma 3.1 and standard estimates, under which we use in particular

$$\operatorname{dist}\big(y_{\tilde{x}_1}(t, u^{t^*}), \Omega\big) \leq \left|y_{\tilde{x}_1}(t, u^{t^*}) - y_{x_1}(t, u^{t^*})\right| \leq L'|\tilde{x}_1 - x_1|,$$
$$\operatorname{dist}\big(y_{\tilde{x}_1}(t, u^{t^*}), \Omega_{\alpha/2}\big) \leq \operatorname{dist}\big(y_{\tilde{x}_1}(t, u^{t^*}), \Omega\big) + \alpha/2, \qquad \alpha \leq \sqrt{\alpha}.$$

By finite induction we can easily verify (4.2) and (4.3) for all m. Now, for any fixed ε it suffices to choose some $\alpha = \alpha_\varepsilon \leq (\varepsilon/L_1)^{2^M}$ and to set $\tilde{u} \equiv \tilde{u}_M$ on $[0, T]$ to have $v_\Omega(x) - v_{\overline{\Omega}}(x) \leq 4\varepsilon$ by inequality (4.1); by the arbitrariness of ε, that concludes the proof. □

5. Constrained viscosity solutions and a comparison result.

We begin recalling the definition of viscosity solution (see [3], [9]) and a comparison theorem in [9]. First of all, for any bounded function $u : E \longrightarrow \mathbb{R}$, $E \subset \mathbb{R}^k$ we define

$$(5.1) \qquad \begin{aligned} u^*(x) &:= \lim_{r \searrow 0} \sup\{u(y) :\ y \in E,\ |x - y| \leq r\} \\ u_*(x) &:= \lim_{r \searrow 0} \inf\{u(y) :\ y \in E,\ |x - y| \leq r\} \end{aligned} \qquad \text{for all } x \in \overline{E}.$$

Throughout this paper we consider the Hamiltonian H defined by (1.5).

DEFINITION 5.1. *We say u is a viscosity subsolution (supersolution) of the Hamilton-Jacobi equation*

$$(5.2) \qquad u(x) + H\big(x, Du(x)\big) = 0$$

on E, if for all $\phi \in C^1(E)$ such that $u^* - \phi$ $(u_* - \phi)$ has a local maximum (minimum) at $x \in E$, we get

$$u^*(x) + H\big(x, \nabla\phi(x)\big) \leq 0 \quad \big(u_*(x) + H\big(x, \nabla\phi(x)\big) \geq 0\big).$$

If u is both subsolution and supersolution, then u is called a viscosity solution. When Ω is an open set and u is subsolution on Ω and supersolution on $\overline{\Omega}$, we say u is a constrained viscosity solution on $\overline{\Omega}$.

THEOREM 5.1. *Let $v_i : \overline{\Omega} \longrightarrow \mathbb{R}$ where $i = 1, 2$ be bounded, v_1 be upper semicontinuous and continuous at each point of $\partial\Omega$, and v_2 be lower semicontinuous. Assume v_1 and v_2 are, respectively, viscosity subsolution in Ω and supersolution in $\overline{\Omega}$ of* (5.2). *Then $v_1 \leq v_2$ in $\overline{\Omega}$.*

Proof. Theorem 5.1 follows as a corollary of Theorem 2.1 in [9] or from Theorem 1.1 in [2] in the unbounded case. □

In a way similar to the one indicated in Theorem 2.1 of [13], but using also Barles and Perthame's and Ishii's arguments on the discontinuous solutions in [3] and in [9], respectively, it is not hard to check the following proposition.

PROPOSITION 5.1. *The value functions $v_{\overline{\Omega}}$ and v_Ω are constrained viscosity solutions of the HJB equation on $\overline{\Omega}$.*

The last result says that under the assumptions of §2, v_Ω is the unique continuous constrained solution of (5.2).

COROLLARY 5.1. *Assume* (A0), (A1). *Then, the unique continuous extension $\bar{v}$ of v_Ω to $\overline{\Omega}$ is the unique continuous constrained solution of the HJB equation* (5.2).

Moreover,

$$\bar{v}(x) = (v_{\overline{\Omega}})_*(x) \quad \textit{for all } x \in \overline{\Omega}. \tag{5.3}$$

Proof. The existence of the unique continuous extension $\bar{v}$ of v_Ω follows by Theorem 3.1, and moreover by Theorem 4.1 in §4 the restriction of $v_{\overline{\Omega}}$ to Ω coincides with the value function v_Ω. Hence the inequality

$$(v_{\overline{\Omega}})_* \leq \bar{v} \leq (v_{\overline{\Omega}})^* \quad \text{on } \partial\Omega$$

is always true, while the relation

$$\bar{v} \leq (v_{\overline{\Omega}})_* \quad \text{on } \partial\Omega$$

follows from the comparison Theorem 5.1 applied to $v_1 = \bar{v}$ and $v_2 = (v_{\overline{\Omega}})_*$. Thus we have verified that (5.3) holds. The uniqueness of the continuous constrained viscosity solution is a consequence of Proposition 5.1 and Theorem 5.1. □

6. An example of discontinuous $v_{\overline{\Omega}}$. Under hypotheses (A0), (A1) in the previous sections, we proved the (uniform) continuity of the value function $v_{\overline{\Omega}}$ only on the open set Ω. In fact, unlike assumption (S), (A1) does not ensure the continuity of $v_{\overline{\Omega}}$ on the whole $\overline{\Omega}$, as shown by the following example.

Consider the control system

$$\begin{cases} y_1'(t) = 2u_1, \\ y_2'(t) = 2(y_1 + y_2)u_2, \end{cases} \tag{6.1}$$

where the state (y_1, y_2) is constrained to stay in the closure of the set Ω defined by

$$\Omega := \{(y_1, y_2) \in \mathbb{R}^2 \ : \ y_1^2 < y_2\}, \tag{6.2}$$

while the control $u = (u_1, u_2)$ takes values on the compact symmetrical w.r.t. 0 subset U of $\mathbb{R}^2$ given by

$$U := \{(1,0), (-1,0), (0,1), (0,-1)\}. \tag{6.3}$$

Let $\psi \in C^\infty(\mathbb{R}^+)$ be a nonincreasing function such that $\psi(r) \equiv 1$ if $0 \leq r \leq 1$ and $\psi(r) \equiv 0$ if $r \geq 2$, and consider

$$J(x,u) := \int_0^{+\infty} e^{-t}\psi(y_{2_x}(t,u))\,dt \tag{6.4}$$

as the associated payoff to be minimized. Thus the value function $v_{\overline{\Omega}}$ is defined by

$$v_{\overline{\Omega}}(x) := \inf_{u \in \mathcal{A}_x} J(x,u) \qquad \forall x = (x_1, x_2) \in \overline{\Omega}, \tag{6.5}$$

where $\mathcal{A}_x$ denotes the set of the admissible controls u such that $y_x(t,u) \in \overline{\Omega}$ $\forall t \in [0, +\infty)$. Note that the admissible vector fields are all tangent only at $(0,0)$ and here there is a Lie bracket pointing inward Ω. Indeed, by setting $u_1 := (1,0)$, $u_2 := (0,1)$, one has $[\sigma_1, \sigma_2](0,0) = (0,4)$, where $\sigma_1(y_1,y_2) = (2,0)$ and $\sigma_2(y_1,y_2) = (0, 2y_1 + 2y_2)$. We shall prove that $v_{\overline{\Omega}}$ is discontinuous at $(0,0)$ by showing that for some $T > 0$: $v_{\overline{\Omega}}(0,\delta) \leq 1 + e^{-T} < 1 = v_{\overline{\Omega}}(0,0)$ $\forall \delta > 0$. To this end, in a similar way to that used in Lemma 3.1, see (3.11), (3.12), for any $\delta > 0$ we define a control $u_\delta \in \mathcal{A}_{(0,\delta)}$ as follows:

$$u_\delta(t) := \sum_{i=1}^{N} \alpha_i(t - T_{i-1})\chi_{[T_{i-1},T_i)}(t) + u_2(t - T_N)\chi_{[T_N,+\infty)}(t), \tag{6.6}$$

where

$$\alpha_i(t) := u_1\big(\chi_{[0,t_i)}(t) - \chi_{[2t_i,3t_i)}(t)\big) + u_2\big(\chi_{[t_i,2t_i)}(t) - \chi_{[3t_i,4t_i]}(t)\big),$$

$$T_0 := 0, \qquad T_i := 4\sum_{j=1}^{i} t_j,$$

and

$$\begin{cases} t_1^2 := \dfrac{\delta}{4}, \\ t_k^2 := \left(\dfrac{1}{4} - \displaystyle\sum_{i=1}^{k-1} t_i^2\right) \wedge \dfrac{1}{2}\displaystyle\sum_{i=1}^{k-1} t_i^2 \quad \forall k > 1, \end{cases} \tag{6.7}$$

while N represents the minimum integer such that the y_2-component of the trajectory $y_{(0,\delta)}(\cdot, u_\delta)$ satisfies $y_2(T_N) \geq 1/2$. It is straightforward to check that at T_k $(k \geq 0)$ one has $y_{(0,\delta)}(T_k, u_\delta) \equiv (0, y_k)$, where y_k is given by

$$y_0 = \delta, \qquad y_k = y_{k-1} + 2t_k(1 - e^{-2t_k}) \quad (k > 0), \tag{6.8}$$

which is equivalent to

$$y_k = \delta + 4\sum_{i=1}^{k} t_i^2 + \sum_{i=1}^{k} o(t_i^2) \quad \text{with } o(t_i^2) := 2t_i(1 - e^{-2t_i}) - 4t_i^2 \quad \forall i, k > 0.$$

Hence it follows that

$$y_k \geq \delta + 2\sum_{i=1}^{k} t_i^2 \qquad \forall t_i \leq \frac{1}{2}.$$

Furthermore as long as $\sum_{i=1}^{k} t_i^2 < 1/4$, t_k^2 is given by

$$t_k^2 = \left(\frac{3}{2}\right)^{k-1} \frac{\delta}{4},$$

we can determine an upper bound $\bar{N}$ for N:

$$\bar{N} := \left[\log_{(3/2)}\left(1 + \frac{1}{2\delta}\right)\right] + 1,$$

where $[\ \cdot\ ]$ represents the integer part. Hence a time sufficient to reach the position $(0, 1/2)$ is given by

$$4\sum_{i=1}^{N} t_i^2 \leq \frac{\sqrt{\delta}}{2}\sum_{i=1}^{N}\left(\sqrt{\frac{3}{2}}\right)^{i-1} = \frac{\sqrt{\delta}}{2}\,\frac{(\sqrt{3/2})^{N-1} - 1}{(\sqrt{3/2}) - 1} \leq \sqrt{3}(2 + \sqrt{6}),$$

and by definition (6.6) of u_δ it follows that

$$v_{\overline{\Omega}}(0, \delta) \leq J\big((0, \delta), u_\delta\big) \leq 1 - e^{-T} \quad \text{if, e.g., } T := \sqrt{3}(2 + \sqrt{6}) + \ln 2 \ \ \forall \delta > 0.$$

Acknowledgments. I thank M. Bardi for helpful discussions and good advice and H. M. Soner for a useful suggestion about the construction of an example of discontinuous $v_{\overline{\Omega}}$. Thanks are also due to the referees for the suggestion of a simplification in the proof of Theorem 3.1.

REFERENCES

[1] M. Bardi and P. Soravia, *Time-optimal control, Lie brackets, and Hamilton–Jacobi equations*, preprint.

[2] ———, *A comparison result for Hamilton–Jacobi equations and applications to some differential games lacking controllability*, Funkcial. Ekvac. 37 (1994), pp. 19–43.

[3] G. Barles and B. Perthame, *Discontinuous solutions of deterministic optimal stopping time problems*, RAIRO Math. Methods Numer. Anal., 21 (1987), pp. 557–579.

[4] I. Capuzzo-Dolcetta and P. L. Lions, *Hamilton–Jacobi equations and state constrained problems*, Trans. Amer. Math. Soc., 318 (1990), pp. 643–668.

[5] M. G. Crandall and P. L. Lions, *Viscosity solutions of Hamilton–Jacobi equations*, Trans. Amer. Math. Soc., 277 (1983) pp. 1–42.

[6] M. G. Crandall, L. C. Evans, and P. L. Lions, *Some properties of viscosity solutions of Hamilton–Jacobi equation*, Trans. Amer. Math. Soc., 262 (1984), pp. 487–502.

[7] L. C. Evans and M. R. James, *The Hamilton–Jacobi–Bellman equation for time-optimal control*, SIAM J. Control Optim., 27 (1989), pp. 1477–1489.

[8] V. Guillemin and A. Pollak, *Differential topology*, Prentice–Hall, Inc., Englewood Cliffs, NJ, 1974, pp. 69–73.

[9] H. Ishii, *A boundary value problem of the Dirichlet type for Hamilton–Jacobi equations*, Ann. Scuola Norm. Sup. Pisa Cl. Sci., 16 (1989), pp. 105–135.

[10] P. L. Lions, *Generalized Solutions of Hamilton–Jacobi Equations*, Pitman, London, 1982.

[11] P. Loreti, *Some properties of constrained viscosity solutions of Hamilton–Jacobi–Bellman equations*, SIAM J. Control Optim., 25 (1987), pp. 1244–1252.

[12] P. Loreti and E. Tessitore, *Approximation and regularity results on constrained viscosity solutions of Hamilton–Jacobi–Bellman equations*, J. Math. Systems, Estim. Control, to appear.

[13] H. M. Soner, *Optimal control with state-space constraints*, SIAM J. Control Optim., 24 (1986), pp. 551-561, 1110–1122.

[14] G. Stefani, *Regularity properties of the minimum-time map*, in Nonlinear Systems, C. I. Byrnes and A. Kurzhansky, eds., Birkhäuser, Boston, pp. 270–282.

[15] H. J. Sussmann, *Optimal control*, in Three decades of mathematical system theory, Lecture Notes in Control and Information Science 135, H. Nijmeijer and J. M. Schumacher, eds., Springer-Verlag, Berlin, New York, 1989, pp. 409–425.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 33, No. 5, pp. 1425–1442, September 1995

© 1995 Society for Industrial and Applied Mathematics

007

THE LINEAR-QUADRATIC CONTROL PROBLEM REVISITED*

TOMASZ R. BIELECKI[†]

Abstract. A long-run, average-cost, stochastic, linear-quadratic control problem that incorporates different time scales is considered. The system dynamics and the cost functional are modeled with the help of a locally square-integrable semimartingale process with independent increments and the corresponding predictable quadratic variation process. The solution of the control problem is given in terms of the solution of certain system of algebraic and differential Riccati equations. The model considered here embodies as particular cases the "traditional" continuous-time and discrete-time linear quadratic control problems, and is applicable, for example, to certain hybrid control problems that cannot be treated using existing control methods.

Key words. linear-quadratic control, square-integrable semimartingales, various time scales, hybrid control

AMS subject classifications. 60H30, 93E20

1. Introduction.

In recent years there has been growing interest in developing a unified approach to control and identification problems for both discrete and continuous-time scales. In Middleton and Goodwin (1990), the unified approach to control and estimation is presented via the so-called "generalized transform." In spite of its many advantages, the method is not capable of handling the problems that "live" in continuous and discrete time simultaneously or stochastic control problems involving the continuous-time scale, for example. This paper provides a way of looking at some of these problems via the stochastic calculus for locally square-integrable semimartingales.

In this paper we consider a long-run, average-cost, stochastic, linear-quadratic control problem that incorporates different time scales. The system dynamics and the cost functional are modeled with the help of a locally square-integrable semimartingale process with independent increments and the corresponding predictable quadratic variation process. The situation considered here is not, of course, "the most general" one. But it is general enough to produce as particular cases the traditional continuous-time (e.g., Davis, 1977) and discrete-time (e.g., Hall and Heyde, 1980) linear-quadratic, stochastic control problems with the average cost per unit of time criterion. The results obtained in the paper follow from an application of the powerful general theory of random processes (Dellacherie and Meyer, 1975, 1980, 1983; Jacod, 1979; Jacod and Shiryayev, 1987; Lipster and Shiryayev, 1989; Protter, 1990, among others). We emphasize that the asymptotic results obtained here are essentially due to the strong law of large numbers type property for semimartingales (see Lipster and Shiryayev, 1989, for example). The L^2-ergodic type results for martingales, discussed by Sundar (1989), may be useful in the study of the control problem with the expected long-run average cost, which is not included here.

The solution of the control problem considered in this paper is given in terms of the solution to the system of algebraic and differential Riccati equations (3.1). Theorem 3.1 concerning the existence and uniqueness of the solution for system (3.1) is

* Received by the editors July 9, 1992; accepted for publication (in revised form) March 2, 1994.

[†] Department of Mathematics, Statistics, and Computer Science, University of Illinois at Chicago, Chicago, Illinois 60680.

interesting in itself. At the very least it interprets the relationship between the algebraic Riccati equations corresponding to continuous and discrete time, as indicated in Remark 3.2 and in §6. The "classical" relationship between continuous- and discrete-time Riccati equations resulting from time discretization is reconfirmed by limiting analysis of equations (3.1) (see §6).

We were inspired to consider a control problem incorporating different time scales by some work on the semimartingale regression problem (Christopeit, 1986; LeBreton and Musiela, 1988), where the time scales are modeled in terms of the predictable quadratic variation process of a semimartingale. Control problems involving continuous-time semimartingale dynamics were considered before in Foldes (1990), for example. To the best of our knowledge, linear quadratic (LQ) control problems incorporating both continuous- and discrete-time scales in the system dynamics have not been considered in the literature before.

Although we consider here only the ergodic linear-quadratic control problem, the modeling methodology presented in this paper is applicable to a wider spectrum of control problems. As a direct control application we see an application of our methodology to a class of hybrid control problems which are attracting more and more interest (see, e.g., Elliot and Sworder, 1992). A simple example of a hybrid control problem that can be treated by methods presented in this paper is given in §7. In this paper we treat neither a finite-time horizon problem, nor an infinite-time horizon with a discounted cost criterion. These are for future research.

The paper is organized as follows. In §2 we describe the noise process. Section 3 introduces a system of differential-algebraic Riccati equations that plays a central role in characterization of the optimal controls (as expected). The system of those equations reduces to the well-known algebraic Riccati equations corresponding to the continuous-time or discrete-time linear-quadratic control problems under appropriate parametrization. Section 4 formulates a semimartingale driven linear-quadratic control problem and provides a solution to it. In §5 we point out how our control problem relates to some other problems considered before in the literature. Section 6 contains three limiting results. One of them reconfirms the classical relationship between continuous and discrete Riccati equations resulting from time discretization. Moreover the result indicates that our approach allows for a "partial" time discretization, that is, time discretization with respect to only some of the components of the state vector. The other two limiting results analyze the effect on the control system of vanishing discrete components ($k_3 \to 0$) and continuous components ($k_1 \to 0$), respectively. In §7 we provide a simple but illustrative example of a hybrid control problem and solve it by our method. A few final remarks are formulated in §8.

Much of the notation used in the paper is taken from Jacod and Shiryayev (1987). "T" denotes the transposition of a matrix.

2. The noise process.

In this section we shall describe the noise process $Z \cong \{Z_t, t \geq 0\}$ that will be appearing in the dynamics equation of the control model. We begin with the following assumption about Z.

Assumption A1. Z is an n-dimensional locally square-integrable semimartingale (Jacod and Shiryayev (1987), Def. II.2.27) and a process with independent increments. The underlying stochastic basis is $(\Omega, \mathcal{F}, \mathbf{F}, P)$ and it is supposed to satisfy the usual conditions.

Let J denote the set of fixed times of discontinuity of Z, that is, $J := \{t \geq 0 : P(\Delta Z_t \neq 0) > 0\}$, where $\Delta Z_t := Z_t - Z_{t-}$ is the jump of Z at time t, $(\Delta Z_0 = 0)$. As usual $\{B, C, \nu\}$ will denote a triplet of predictable characteristics of Z with regard to some truncation function h. According to Theorem II.4.15 of Jacod and Shiryayev (1987) we also have, under A1, that the characteristics of Z are deterministic processes and $J = \{t \geq 0 : \nu(\{t\} \times R^n) > 0\}$. We will denote the stochastically continuous and stochastically discontinuous components of Z by $\bar{Z}$, and $\bar{\bar{Z}}$ respectively. This means that

$$\bar{\bar{Z}}_t = \sum_{\substack{0<s\leq t \\ s\in J}} \Delta Z_s \text{ and } \bar{Z}_t = Z_t - \bar{\bar{Z}}_t,\ t \geq 0.$$

From Proposition II.1.16 of Jacod and Shiryayev (1987) we know that J is countable. Denote elements of J by j_n, $n \in I$, where I is a countable index set. Also let $\epsilon_n := \bar{\bar{Z}}_{j_n}$, $n \in I$, so that $\{\epsilon_n\}_{n\in I}$ is the embedded random sequence. Because of the control problem we will treat in §4 we introduce the following two assumptions.

Assumption A2. (a) $I = N^* := \{1, 2, 3, \dots\}$; (b) $j_n = \epsilon n$, $\epsilon > 0$, $n \in N^*$.

Assumption A3. Both $\bar{Z}$ and $\{\epsilon_n\}_{n\in N^*}$ have stationary increments.

Assumption A2(b) is not essential for the control-theoretic considerations to follow. It will be used to simplify the presentation.

We will keep the usual notation for the measure of jumps of $Z : \mu^z$. From the above stated assumptions and the results of Jacod and Shiryayev (1987), Chapters I and II, we infer the following.

PROPOSITION 2.1. *Assume* A1–A3. *Then*

(i) *the canonical decomposition of Z has the form $Z = Z_0 + N_1 + N_2 + N_3 + A$, where $N_1 := Z^c$ is the continuous martingale part of Z, $N_2 := (z1_{J^c} * (\mu^z - \nu)$ is the stochastically continuous jump-martingale part of Z, $N_3 := (z1_J) * (\mu^z - \nu)$ is the stochastically discontinuous jump-martingale part of Z, and $A = B + (z - h(z)) * \nu$ is a deterministic process.*

(ii) *the characteristics of Z are $B_t = bt + (h1_J) * \nu$, $C_t = ct$ and $\nu(\cdot, dt, dz) = dtK_2(dz)1_{J^c}(t) + K_3(dz)1_J(t)$, where $b \in R^n$, $c \in L(R^n, R^n)$ and $c \geq 0$, K_2 and K_3 are positive measures on R^n satisfying $K_i(\{0\}) = 0$, and $k_i := \int |z|^2 K_i(dz) < +\infty$, $i = 2, 3$.*

(iii) *N_1, N_2, and N_3 are independent and their predictable quadratic variation processes are given by*

$$\begin{aligned}
\langle N_1^i, N_1^j \rangle &= C^{ij}, \\
\langle N_2^i, N_2^j \rangle &= (z^i z^j)1_{J^c} * \nu, \\
\langle N_3^i, N_3^j \rangle &= (z^i z^j)1_J * \nu - \sum_{s\leq\cdot} \int z^i \nu(\{s\} \times dz) \int z^j \nu(\{s\} \times dz)
\end{aligned}$$

for $i, j = 1, 2, \dots, n$.

Proof. (i) The result follows from (2.30) and (2.39), Chapter II of Jacod and Shiryayev (1987).

(ii) The result follows from (2.14), (4.16), and the result analogous to Corollary 4.19 of Jacod and Shiryayev (1987) applied to $\bar{Z}$ and $\bar{\bar{Z}}$, respectively.

(iii) This part of the proposition follows from (2.31) and (4.16) of Jacod and Shiryayev (1987), Chapter II. □

As usual, we let $\langle N_1 \rangle :=$ trace C, $\langle N_2 \rangle :=$ trace$(zz^T)1_{J^c} * \nu$ and $\langle N_3 \rangle :=$ trace$(zz^T)1_J * \nu$ denote the scalar predictable quadratic variation processes of N_1, N_2, and N_3, respectively.

Remark 2.1. From now on we will assume (without loss of generality) that $A \equiv 0$

and $Z_0 \equiv 0$.

COROLLARY 2.1. *Under conditions of Proposition* 2.1 *we have*

$$\begin{aligned}\langle N_1\rangle_t &= t \cdot \text{ trace } c,\\ \langle N_2\rangle_t &= t \cdot k_2 \text{ for } t \geq 0, \text{ and}\\ \langle N_3\rangle_t &= n \cdot k_3,\ t \in]\epsilon n, \epsilon(n+1)],\ n \geq 0,\\ \langle N_3\rangle_0 &= 0.\end{aligned}$$

Remark 2.2. In fact, N_1 is a Wiener process.

3. The Riccati equations.

In this section we let $A, B \in L(R^n, R^n)$, $E \in L(R^m, R^n)$, and $F \in L(R^k, R^n)$. Also let $Q_1, Q_2 \in L(R^n, R^n)$, $R_1 \in L(R^m, R^m)$, $R_2 \in L(R^k, R^k)$, and $Q_1, Q_2 \geq 0$, $R_1, R_2 > 0$.

In Definition 3.1 below we recall the concepts of Hurwitz and Schur stability of a matrix, which we shall respectively call c- and d-stability with reference to "continuous" and "discrete" time.

DEFINITION 3.1. *A quadratic matrix M is called d-stable iff its spectrum is contained in an open unit disk. A quadratic matrix N is called c-stable iff its spectrum is contained in the complex left open half-plane.*

DEFINITION 3.2.

(a) *A pair (B, F) is called d-stabilizable iff there exists $H \in L(R^n, R^k)$ so that $B + FH$ is d-stable. A pair (B, F) is said to be d^T-stabilizable iff there exists $H \in L(R^k, R^n)$ so that $B + HF$ is d-stable.*

(b) *A four-tuple (A, E, B, F) is called cd-stabilizable iff there exist $H_1 \in L(R^n, R^m)$ and $H_2 \in L(R^n, R^k)$ so that $\mathcal{A}(H_1, H_2)$ is d-stable, where*

$$\mathcal{A}(H_1, H_2) := e^{A+EH_1} \cdot (B + FH_2).$$

DEFINITION 3.3. *A four-tuple (A, Q_1, B, Q_2) is called cd-detectable iff $(Be^A, \sqrt{B^T e^{A^T} Q_1 e^A B + Q_2})$ is d^T-stabilizable.*

Remark 3.1. Note that if $(e^A, \sqrt{e^{A^T} Q_1 e^A})$ is d^T-stabilizable then $(e^A, \sqrt{Q_1})$, is d^T-stabilizable and consequently $(\sqrt{Q_1}, A)$ is c-detectable, which means that there is a matrix H such that $A^T + \sqrt{Q_1^T}\, H$ is d^T-stable.

Proof. This follows from the fact that $\text{Ker}(\sqrt{M}) = \text{Ker}(M)$ for any symmetric, nonnegative semidefinite matrix M and from Proposition 3.1 in Wonham (1979). □

In what follows we will require more notation. Let $\epsilon > 0$. Let $P_t : [0, \epsilon] \to L_+(R^n, R^n)$ be a continuous function, where "+" denotes nonnegative semidefiniteness. Next define

$$\begin{aligned}\mathcal{P}(\epsilon) &:= \{P_t, t \in [0, \epsilon]\},\\ L_t(\epsilon) &:= -R_1^{-1} E^T P_{\epsilon - t}, \quad L_1(\mathcal{P}(\epsilon), \epsilon) := \int_0^\epsilon L_t(\epsilon)dt,\\ \mathcal{D}_1(\mathcal{P}(\epsilon), \epsilon) &:= \epsilon A + EL_1(\mathcal{P}(\epsilon), \epsilon), \quad \mathcal{D}_{1,t}(\epsilon) := A + EL_t(\epsilon),\\ L_2(\mathcal{P}(\epsilon), \epsilon) &:= -(F^T P_\epsilon F + R_2)^{-1} F^T P_\epsilon B,\\ \mathcal{D}_2(\mathcal{P}(\epsilon), \epsilon) &:= B + FL_2(\mathcal{P}(\epsilon), \epsilon),\\ \mathcal{S}(\mathcal{P}(\epsilon), \epsilon) &:= \int_0^\epsilon e^{\int_s^\epsilon \mathcal{D}^T_{1,\epsilon - t}(\epsilon)dt} [Q_1^T + L^T_{\epsilon - s}(\epsilon) R_1 L_{\epsilon - s}(\epsilon)] e^{\int_s^\epsilon \mathcal{D}^T_{1,\epsilon - t}(\epsilon)dt} ds,\\ \mathcal{B}_{(\mathcal{P}(\epsilon),\epsilon)} &:= \mathcal{D}_2^T(\mathcal{P}(\epsilon), \epsilon)\mathcal{S}(\mathcal{P}(\epsilon), \epsilon)\mathcal{D}_2(\mathcal{P}(\epsilon), \epsilon)) + Q_2 + L_2^T(\mathcal{P}(\epsilon), \epsilon) R_2 L_2(\mathcal{P}(\epsilon), \epsilon),\end{aligned}$$

and

$$\mathcal{A}_{(\mathcal{P}(\epsilon),\epsilon)} : L(R^n, R^n) \to L(R^n, R^n)$$

given by

$$\mathcal{A}_{(\mathcal{P}(\epsilon),\epsilon)}(K) := \mathcal{D}_2^T(\mathcal{P}(\epsilon),\epsilon)e^{\mathcal{D}_1^T(\mathcal{P}(\epsilon),\epsilon)}Ke^{\mathcal{D}_1(\mathcal{P}(\epsilon),\epsilon)}\mathcal{D}_2(\mathcal{P}(\epsilon),\epsilon).$$

Consider the following system of Riccati equations, which we will call a *cd*-Riccati equation:

$$\text{(3.1)} \qquad \begin{cases} \mathcal{A}_{(\mathcal{P}(\epsilon),\epsilon)}(R) + \mathcal{B}_{(\mathcal{P}(\epsilon),\epsilon)} = R, \\ \dot{P}_t = Q_1 + A^T P_t + P_t A - P_t E R_1^{-1} E^T P_t, \\ P_0 = R, \ t \in [0,\epsilon]. \end{cases}$$

Observe that

$$\text{(3.2)} \qquad P_\epsilon = e^{\mathcal{D}_1^T(\mathcal{P}(\epsilon),\epsilon)}Re^{\mathcal{D}_1(\mathcal{P}(\epsilon),\epsilon)} + \mathcal{S}(\mathcal{P}(\epsilon),\epsilon).$$

Therefore the first equation in (3.1) can equivalently be written as

$$\text{(3.3)} \quad (B+FL_2(\mathcal{P}(\epsilon),\epsilon))^T P_\epsilon(B+FL_2(\mathcal{P}(\epsilon),\epsilon))+Q_2+L_2^T(\mathcal{P}(\epsilon),\epsilon)R_2L_2(\mathcal{P}(\epsilon),\epsilon) = R.$$

Remark 3.2.

(a) If we assume that $B = I$, or $B = -I$, $F = 0$, $Q_2 = 0$, and $P_t =$ const, $t \in [0,\epsilon]$, then (3.1) reduces to the following algebraic Riccati equation;

$$\text{(c-ARE)} \qquad \begin{aligned} 0 &= Q_1 + A^T R + RA - RER_1^{-1}E^T R, \\ P_t &= R, \ t \in [0,\epsilon]. \end{aligned}$$

We call the above equation c-ARE because it is related to the continuous-time linear-quadratic control problem (see Davis (1977), p. 185).

(b) If we assume that $A = 0$, $E = 0$, $Q_1 = 0$, and $\epsilon = 1$ then (3.1) reduces to the following algebraic Riccati equation,

$$\text{(d-ARE)} \qquad \begin{aligned} R &= B^T[R - RF(F^T RF + R_2)^{-1}F^T R]B + Q_2, \\ P_t &= R, \ t \in [0,1]. \end{aligned}$$

We call the above equation d-ARE since it is related to the discrete-time linear-quadratic control problem (see Bertsekas (1976), p. 355).

In Theorem 3.1 below we shall consider equations (3.1) for $\epsilon = 1$ only. The result is true for any $\epsilon > 0$, as can be easily deduced from the proof of the theorem. We will use a simplified notation by omitting $\epsilon = 1$ from the above definitions. So, for example, we will write $\mathcal{P}$ instead of $\mathcal{P}(1)$, L_1 instead of $L_1(\mathcal{P}(1),1)$, $\mathcal{A}_\mathcal{P}$ instead $\mathcal{A}_{(\mathcal{P}(1),1)}$, etc.

We note that Theorem 12.2 of Wonham (1979) and Proposition on page 75 of Bertsekas (1977) are special cases of Theorem 3.1.

THEOREM 3.1. *Let* $\epsilon = 1$. *Assume that* (A, E, B, F) *is cd-stabilizable and* (A, Q_1, B, Q_2) *is cd-detectable. Then there exists the unique solution* $(\bar{R}, \bar{\mathcal{P}})$ *to* (3.1) *such that* $\bar{R} \geq 0$, $\bar{P}_t \geq 0$ *for* $t \in [0,1]$, *and* $\mathcal{A}(\bar{L}_1, \bar{L}_2)$ *is d-stable, where* $\bar{L}_i := L_i(\bar{\mathcal{P}})$, $i = 1, 2$.

Proof. See Appendix 1.

4. Linear-quadratic stochastic control problem. We begin with introducing the dynamics of the controlled process first:

$$dx_t = (\widetilde{A}x_{t-} + \widetilde{E}v_{t-})d\langle M\rangle_t + (\widetilde{B}x_{t-} + \widetilde{F}u_{t-})d\langle N\rangle_t$$
$$+ dZ_t,\ x_0 = x,\ t \geq 0, \tag{4.1}$$

where $M = N_1 + N_2$, $N = N_3$. The admissible control processes $u. := \{u_t, t \geq 0\}$ and $v. := \{v_t, t \geq 0\}$ are supposed to satisfy the following conditions:

- They are non-anticipating w.r.t. Z,
- There exists a weak semimartingale solution to (4.1) in the sense of Jacod (1979) Chap. XIV,
- $\lim_{t\to\infty} \frac{\|x_t\|^2}{t} = 0$, a.s.,
- $\overline{\lim}_{t\to\infty} \frac{1}{t}\int_0^t(\|x_s\|^2 + \|u_s\|^2 + \|v_s\|^2)ds < +\infty$, a.s.

The class of admissible controls is denoted by $\mathcal{U}_{ad}$. The cost functional will be given in terms of $(T \geq 0)$

$$\begin{aligned}\mathbf{C}_T(v., u., x) = \int_{0^+}^T &[\langle\widetilde{Q}_1 x_{t-}, x_{t-}\rangle \\ &+ \langle\widetilde{R}_1 v_{t-}, v_{t-}\rangle]d\langle M\rangle_t \\ &+ \int_{0^+}^T [\langle\widetilde{Q}_2 x_{t-}, x_{t-}\rangle \\ &\qquad + \langle\widetilde{R_2} u_{t-}, u_{t-}\rangle]d\langle N\rangle_t.\end{aligned}$$

We want to show the existence and characterization of optimal controls, that is, admissible controls $u^0.$ and $v^0.$ such that for all $v., u. \in \mathcal{U}_{ad}$ and $x \in R^n$ it holds that

$$\mathbf{C}(v^0., u^0., x) \leq \mathbf{C}(v., u., x),$$

where

$$\mathbf{C}(v., u., x) := \overline{\lim_{T\to\infty}} \frac{1}{T}\mathbf{C}_T(v., u., x).$$

In the above description of the control problem we have supposed $\widetilde{A}, \widetilde{B} \in L(R^n, R^n)$, $\widetilde{E} \in L(R^m, R^n)$, $\widetilde{F} \in L(R^k, R^n)$, $\widetilde{Q}_1$ and $\widetilde{Q}_2$ are in $L(R^n, R^n)$, and $\widetilde{Q}_1, \widetilde{Q}_2 \geq 0$, $\widetilde{R}_1 \in L(R^m, R^m)$, $\widetilde{R}_2 \in L(R^k, R^k)$, and $\widetilde{R}_1, \widetilde{R} > 0$. Throughout this section we let $k_1 := k_2 + \text{trace } c$, $A := k_1\widetilde{A}$, $E := k_1\widetilde{E}$, $B := k_3\widetilde{B} + I$, $F := k_3\widetilde{F}$, $Q_1 := k_1\widetilde{Q}_1$, $Q_2 := k_3\widetilde{Q}_2$, $R_1 := k_1\widetilde{R}_1$, and $R_2 := k_3\widetilde{R}_2$. The following assumptions will be used.

Assumption A4. (A, E, B, F) is *cd*-stabilizable.

Assumption A5. (A, Q_1, B, Q_2) is *cd*-detectable.

Fix $\epsilon > 0$. Let $(\bar{\bar{R}}, \bar{\bar{P}})$ denote the solution to (3.1) with t substituted with $k_3 t$ and $\dot{P}_t$ changed to $k_3\dot{P}_{k_3 t}$, $t \in [0, \epsilon]$. Define $\Pi_t : [0, \infty) \to L(R^n, R^n)$ by

$$\Pi_{n+s} = \bar{\bar{P}}_{(\epsilon-s)k_3},\ s \in [0, \epsilon),\ n = 0, 1, 2, \ldots .$$

Remark 4.1. Note that

$$\begin{aligned}\dot{\Pi}_t &= -k_3\dot{\bar{\bar{P}}}_{(\epsilon n-t)k_3} = -Q_1 - A^T\bar{\bar{P}}_{(\epsilon n-t)k_3} - \bar{\bar{P}}_{(\epsilon n-t)k_3}A \\ &\quad + \bar{\bar{P}}_{(\epsilon n-t)k_3}ER_1^{-1}E^T\bar{\bar{P}}_{(\epsilon n-t)}k_3 \\ &= -Q_1 - A^T\Pi_t - \Pi_t A + \Pi_t E R_1^{-1}E^T\Pi_t\end{aligned}$$

for $t \in [\epsilon n, \epsilon(n+1))$, $n = 0, 1, 2, \ldots$.

Also let $\widetilde{P} := \epsilon^{-1}\int_0^1 \Pi_t dt = \epsilon^{-1}\int_0^1 \bar{\bar{P}}_{tk_3} dt$. Define controls $(u^0., v^0.)$ by

$$v_t^0 := \bar{\Lambda}_t x_t,\ u_t^0 := \Lambda_2 x_t, \tag{4.2}$$

where

$$\bar{\Lambda}_t := -R_1^{-1}E^T\Pi_t,\ \Lambda_2 := -(F^T\Pi_0 F + R_2)^{-1}F^T\Pi_0 B$$

for $t \geq 0$.

We will need one more assumption.

Assumption A6. $\int |z|^4 K_2(dz) < +\infty$.

THEOREM 4.1. *Suppose assumptions* A1–A6 *are satisfied. Then we have the following:*

(a) *The definitions* (4.2) *are correct: there exists a unique, strong, semimartingale solution to* (4.1) *with* $(v^0., u^0.)$ *in place of* $(v., u.)$,

(b) *The controls* $(v^0., u^0.)$ *are optimal,*

(c) $\mathbf{C}(v^0., u^0., x) = \epsilon^{-1}$*trace*$(c + k_2 I)\widetilde{P} + \epsilon^{-1}$*trace* $k_3\Pi_0$ *for all* $x \in R^n$.

Proof. (a). It is enough to note that $\langle M\rangle$ and $\langle N\rangle$ are special semimartingales and apply Theorem V.3.7 of Protter (1990).

(b). We will use the standard comparison method.

Step 1. Let us first observe the following:

- For all $x \in R^n$ and $t \geq 0$

$$\min_{v\in R^m}[x^T\Pi_t E v + v^T E^T \Pi_t x + v^T R_1 v] = x^T\Pi_t E R_1^{-1}E^T\Pi_t x = -x^T\Pi_t E\bar{\Lambda}_t x,$$

where the minimum is realized by

$$^0v_t := \bar{\Lambda}_t x;$$

- For each $x \in R^n$

$$\begin{aligned}\min_{u\in R^k}&[u^T F^T\Pi_0 Bx + x^T B^T\Pi_0 F u + u^T F^T \Pi_0 F u + u^T R_2 u]\\ &= x^T B^T \Pi_0 F(F^T\Pi_0 F + R_2)^{-1}F^T\Pi_0 Bx\\ &= x^T\Lambda_2^T(F^T\Pi_0 F + R_2)\Lambda_2 x\end{aligned}$$

and the minimum is realized by

$$^0u := \Lambda_2 x.$$

Step 2. Now let $(v., u.) \in \mathcal{U}_{ad}$ be an arbitrary pair of admissible controls. From Lemma A.3.1 of Appendix 3 it follows that $(v^0., u^0.) \in \mathcal{U}_{ad}$. Consider the function $V : [0,\infty) \times R^n \to R$ given by

$$V(t,x) := x^T\Pi_t x.$$

Upon application of Ito's rule for semimartingales (Jacod and Shiryayev (1987), Thm. I.4.57) to V we obtain, for $t \geq 0$,

(4.3)

$$\begin{aligned}x_t^T\Pi_t x_t - x^T\Pi_0 x = \int_{0+}^t &[x_{s-}^T\Pi_{s-}\widetilde{A}x_{s-}\\ &+ x_{s-}^T\Pi_{s-}\widetilde{E}v_{s-} + x_{s-}^T\widetilde{A}^T\Pi_{s-}x_{s-} + v_{s-}^T\widetilde{E}^T\Pi_{s-}x_{s-}]d\langle M\rangle_s\end{aligned}$$

$$
\begin{aligned}
&+\int_{0+}^{t}[x_{s-}^T\Pi_{s-}\widetilde{B}x_{s-}+x_{s-}^T\Pi_{s-}\widetilde{F}u_{s-}+x_{s-}^T\widetilde{B}^T\Pi_{s-}x_{s-}\\
&\qquad +u_{s-}^T\widetilde{F}^T\Pi_{s-}x_s]d\langle N\rangle_s\\
&+2\int_{0+}^{t}x_{s-}^T\Pi_{s-}dZ_s+\int_{0+}^{t}x_{s-}^Td\Pi_s x_{s-}\\
&+\int_{0+}^{t}\text{trace }\Pi_{s-}c ds\\
&+\sum_{\substack{0<s\le t\\ s\in J^c}}\{x_s^T\Pi_s x_s-x_{s-}^T\Pi_{s-}x_{s-}-\Delta x_s^T\Pi_{s-}x_{s-}\\
&\qquad -x_{s-}^T\Pi_{s-}\Delta x_s-x_{s-}^T\Delta\Pi_s x_s\}\\
&+\sum_{0<\epsilon n\le t}\{(Bx_{\epsilon n-}+Fu_{\epsilon n-}+\Delta N_{\epsilon n})^T\Pi_{\epsilon n}(Bx_{\epsilon n-}+Fu_{\epsilon n-}+\Delta N_{\epsilon n})\\
&\qquad -x_{\epsilon n-}^T\Pi_{\epsilon n-}x_{\epsilon n-}-(Bx_{\epsilon n-}+Fu_{\epsilon n-})^T\Pi_{\epsilon n-}x_{\epsilon n-}\\
&\qquad -x_{\epsilon n-}^T\Pi_{\epsilon n-}(Bx_{\epsilon n-}+Fu_{\epsilon n-})-x_{\epsilon n-}^T\Delta\Pi_{\epsilon n}x_{\epsilon n-}\}\\
&=-\int_{0+}^{t}[x_{s-}^T\widetilde{Q}_1x_{s-}-v_{s-}^T\widetilde{R}_1v_{s-}]d\langle M\rangle_s\\
&+\int_{0+}^{t}[x_{s-}^T\Pi_{s-}\widetilde{E}v_{s-}+v_{s-}^T\widetilde{E}^T\Pi_{s-}x_{s-}+v_{s-}^T\widetilde{R}_1v_{s-}\\
&\qquad +x_{s-}^T\Pi_{s-}\widetilde{E}\widetilde{R}_1^{-1}\widetilde{E}^T\Pi_{s-}x_{s-}]d\langle M\rangle_s\\
&+\sum_{\substack{0<s\le t\\ s\in J^C}}\{x_s^T\Pi_s x_s-x_{s-}^T\Pi_{s-}x_{s-}-\Delta x_s^T\Pi_{s-}x_{s-}\\
&\qquad -x_{s-}^T\Pi_{s-}\Delta x_s\}+\int_{0+}^{t}\text{trace }\Pi_{s-}c\, ds\\
&-\int_{0+}^{t}[x_{s-}^T\widetilde{Q}_2x_{s-}+u_{s-}^T\widetilde{R}_2u_{s-}]d\langle N\rangle_s\\
&+\int_{0+}^{t}[x_{s-}^TB^T\Pi_sFu_{s-}+u_{s-}^TF^T\Pi_sBx_{s-}+u_{s-}^TF^T\Pi_sFu_{s-}\\
&\qquad +u_{s-}^TR_2u_s-x_{s-}^T\Lambda_2^T(F^T\Pi_sF+R_2)\Lambda_2x_s]\frac{d\langle N\rangle_s}{k^3}\\
&+2\int_{0+}^{t}[x_{s-}^TB^T\Pi_s+u_{s-}^TF^T\Pi_s]dN_s+2\int_{0+}^{t}x_{s-}^T\Pi_{s-}dZ_s\\
&+\int_{0+}^{t}dN_s^T\Pi_s dN_s\\
&=\sum_{i=1}^{9}I_t^i,\quad \text{a.s.}
\end{aligned}
$$

Note that from Step 1 it follows that $I_t^2\ge 0$ and $I_t^6\ge 0$, $t\ge 0$. As in Theorem 3.6.1 of Lipster and Shiryayev (1989) we have

$$I_t^3=(z^T\Pi z)1_{J^c}*(\mu^z-\nu)_t+(z^T\Pi z)1_{J^c}*\nu_t$$

and note that in view of A6 the process $\zeta_t := (z^T\Pi z)1_{J^c}*(\mu^z-\nu)_t$ is a locally square-integrable martingale with predictable quadratic variation process

$$\langle\zeta\rangle_t = \int_{0+}^t \int_{R^n} (z^T\Pi_s z)^2 1_{J^c} ds K_2(dz).$$

Taking the above remarks into account we obtain from (4.2) the following, $t \geq 0$,

$$\begin{aligned} x_t^T\Pi_t x_t - x^T\Pi_0 x + \mathbf{C}_t(v., u., x) \\ \geq (z^T\Pi z)1_{J^c} * \nu_t + \int_{0+}^t \text{trace } \Pi_{s-} c \, ds \\ + \int_{0+}^t dN_s^T\Pi_s dN_s + \zeta_t + \rho_t + \xi_t, \text{ a.s.,} \end{aligned} \tag{4.4}$$

where

$$\rho_t := 2\int_{0+}^t [x_{s-}^T B^T\Pi_s + u_{s-}^T F^T\Pi_s] dN_s$$

and $\xi_t := 2\int_{0+}^t x_{s-}^T\Pi_{s-} dz_s$ are locally square-integrable martingales. Since $(\Pi_t)_{t\geq 0}$ is periodic we have $\lim_{t\to\infty} \frac{1}{t}(z^T\Pi z)1_{J^c} * \nu_t = \epsilon^{-1}k_2 \text{trace } \widetilde{P}$ and $\lim_{t\to\infty} \frac{1}{t}\int_{0+}^t \text{trace } \Pi_{s-} c ds = \epsilon^{-1}$ trace $\widetilde{P}c$. Applying ergodic theorem to $\varphi_t := \int_{0+}^t dN_s^T\Pi_s dN_s$ we get $\lim_{t\to\infty}\frac{1}{t}\varphi_t = \epsilon^{-1}k_3$ trace Π_0, a.s. Also, it follows from the results of Lipster and Shiryayev (1989), §2.6, that $\lim_{t\to\infty}\frac{1}{t}\zeta_t = \lim_{t\to\infty}\frac{1}{t}\rho_t = \lim_{t\to\infty}\frac{1}{t}\xi_t = 0$, a.s. Therefore from (4.4) we conclude that

$$\mathbf{C}(v., u., x) \geq \epsilon^{-1}\text{trace } (c + k_2 I)\widetilde{P} + \epsilon^{-1}k\Pi_0, \text{ a.s.} \tag{4.5}$$

Using considerations analogous to the ones above, it is straightforward to show that

$$\mathbf{C}(v^0., u^0., x) = \epsilon^{-1}\text{trace } (c + k_2 I)\widetilde{P} + \epsilon^{-1}k\Pi, \text{ a.s.} \tag{4.6}$$

This concludes the proof of (b).

(c). This result follows from (4.5) and (4.6). ||

5. Some special cases. In this section we will shortly demonstrate that Theorem 4.1 encompasses solutions to some "classical" stochastic linear-quadratic control problems.

Case 1 (continuous time system driven by Wiener process). Using our notation this case corresponds to

$$k_2 = 0, \ k_3 = 0.$$

For a problem of this type see, for example, Davis (1977).

Case 2 (continuous time system driven by Wiener and Poisson processes). This corresponds to

$$k_3 = 0, \ N_2 \text{ equivalent to a Poisson process.}$$

Note that in case of a Poisson process Assumption A6 is automatically satisfied. For a problem of this type see, for example, Wonham (1970).

Case 3 (continuous time system driven by a Poisson process). This case corresponds to

$$c = 0 \text{ and } k_3 = 0, \ N_2 \text{ equivalent to a Poisson process.}$$

For a more general model of this type (including the multiplicative noise components) see, for example, Li and Blankenship (1986).

Case 4 (discrete time system driven by a sequence of independent random variables). In our terminology this case corresponds to

$$c = 0, \quad k_2 = 0.$$

For a problem of this type see, for example, Hall and Heyde (1980).

6. Three limiting results. Let us consider equations (3.1) with $A, B, E, F, Q_1, Q_2, R_1, R_2$ as in §4. We also require that k_3 is changed to ϵk_3 in the definitions of B, F, Q_2, and R_2, that time index t is substituted with $k_3 t$, and that $\dot{P}_t$ is changed to $k_3 \dot{P}_{k_3 t}$.

In this section we shall analyze the behavior of equations (3.1) in the present setting when (i) ϵ tends to 0, (ii) k_3 tends to zero, and (iii) k_1 tends to zero. Note that the first case corresponds to "increasing frequency of the discrete time component." A "classical" prototype of it has been considered before in the context of approximating of a continuous-time linear-quadratic problem with a sequence of discrete-time linear-quadratic problems (see Whittle, 1983, Ex. 1, p. 209, for example). The second case corresponds to vanishing of the discrete-time component, and the third case corresponds to vanishing of the continuous-time component of the system.

Case i. Assume A4 and A5. Also assume that $(A + k_3\widetilde{B}, [E \ F])$ and $(A^T + k_3\widetilde{B}^T, \sqrt{Q_1 + Q_2})$ are c-stabilizable pairs. Denote by $(R(\epsilon, k_1, k_3), \mathcal{P}(\epsilon, k_1, k_3))$ the solution to (3.1). Then, using (3.1)–(3.3) and some algebra, it can be shown that

$$\lim_{\epsilon \to 0} (R(\epsilon, k_1, k), \mathcal{P}(\epsilon, k_1, k)) = (P(0, k_1, k_3), P(0, k_1, k_3)) \ , \tag{6.1}$$

where $P(0, k_1, k_3)$ is the solution to

$$\begin{aligned} Q_1 + Q_2 + (A + k_3\widetilde{B})^T P + P(A + k_3\widetilde{B}) \\ -P[E \ F] \begin{pmatrix} R_1^{-1} & 0 \\ 0 & R_2^{-1} \end{pmatrix} [E \ F]^T P = 0 \ . \end{aligned} \tag{6.2}$$

Example (partial time discretization). Here $n = 2$, $m = k = k_1 = k_3 = 1$. We also let

$$\widetilde{A} = \begin{pmatrix} a & b \\ 0 & 0 \end{pmatrix}, \ \widetilde{E} = \begin{pmatrix} e \\ 0 \end{pmatrix}, \ \widetilde{B} = \begin{pmatrix} 0 & 0 \\ c & d \end{pmatrix}, \ \widetilde{F} = \begin{pmatrix} 0 \\ f \end{pmatrix},$$

$$\widetilde{Q}_1 = \begin{pmatrix} q_1 & 0 \\ 0 & 0 \end{pmatrix}, \ \widetilde{Q}_2 = \begin{pmatrix} 0 & 0 \\ 0 & q_2 \end{pmatrix},$$

$$\widetilde{R}_1 = r_1, \ \widetilde{R}_2 = r_2 \ .$$

This parametrization corresponds, for example, to a partial time discretization, with time step ϵ, of the following control problem (here we are using notation $x(t)$ and $u(t)$ instead of x_t and u_t):

(6.3) Minimize

$$\overline{\lim_{T \to \infty}} \, T^{-1} \int_0^T \left\{ (x_1(t), x_2(t)) \begin{pmatrix} q_1 & 0 \\ 0 & q_2 \end{pmatrix} \begin{pmatrix} x_1(t) \\ x_2(t) \end{pmatrix} + (u_1(t), u_2(t)) \begin{pmatrix} r_1 & 0 \\ 0 & r_2 \end{pmatrix} \begin{pmatrix} u_1(t) \\ u_2(t) \end{pmatrix} \right\} dt$$

subject to

$$\begin{aligned} dx_1(t) &= (ax_1(t) + bx_2(t))dt + eu_1(t)dt + dw_1(t), \\ dx_2(t) &= (cx_1(t) + dx_2(t))dt + fu_2(t)dt + dw_2(t), \end{aligned}$$

where w_1 and w_2 are standard one-dimensional Brownian motions.

In this case "partial time discretization" means time discretization with respect to the second state component $x_2(t)$. The partially time-discretized problem is

Minimize

$$\overline{\lim_{T\to\infty}}\, T^{-1}\int_0^T \left(q_1 x_{1,\epsilon}^2(t) + r_1 u_1(t)^2\right) dt + \epsilon^{-1}\,\overline{\lim_{N\to\infty}}\, N^{-1}\sum_{n=1}^{N}\left(\epsilon q_2 x_{2,\epsilon}^2(n) + \epsilon r_2 u_{2,\epsilon}^2(n)\right)$$

subject to

$$\begin{aligned} dx_{1,\epsilon}(t) &= (ax_{1,\epsilon}(t) + bx_{2,\epsilon}(n))dt + eu_1(t)dt + dw_1(t),\\ &\quad t \in [\epsilon n, \epsilon(n+1)),\\ x_{2,\epsilon}(n+1) &= \epsilon c x_{1,\epsilon}(\epsilon n) + (\epsilon+1)dx_{2,\epsilon}(n) + \epsilon u_{2,\epsilon}(n) + (w_2(\epsilon(n+1)) - w_2(\epsilon n))\\ &\quad n = 0, 1, 2, \ldots . \end{aligned}$$

Here, the limiting equation (6.2) coincides with the algebraic Riccati equation corresponding to the original problem (6.3).

We believe that our methodology will allow for time discretization of continuous-time control problems using various time steps for various components of state vector, if necessary.

Case ii. Assume A4 and A5. Then

$$\lim_{k_3\to 0}(R(\epsilon,k_1,k_3), \mathcal{P}(\epsilon,k_1,k_3)) = (P(\epsilon,k_1,0), \mathcal{P}(\epsilon,k_1,0)) \,, \tag{6.4}$$

where $\mathcal{P}(\epsilon,k_1,0) = \{P_t = P(\epsilon,k_1,0), t\in[0,\epsilon]\}$ and $P(\epsilon,k_1,0)$ is the solution of

$$0 = Q_2 + A^T P + PA - PER^{-1}E^T P \,. \tag{6.5}$$

Case iii. Assume A4 and A5. Then

$$\lim_{k_1\to 0}(R(\epsilon,k_1,k_3), \mathcal{P}(\epsilon,k_1,k_3)) = (P(\epsilon,0,k_3), \mathcal{P}(\epsilon,0,k_3)) \,, \tag{6.6}$$

where $\mathcal{P}(\epsilon,0,k_3) = \{P_t = P(\epsilon,0,k_3), t\in[0,\epsilon]\}$ and $P(\epsilon,0,k_3)$ is the solution to

$$P = B^T[P - PF(F^T PF + R_2)^{-1}F^T P]B + Q_2 \,. \tag{6.7}$$

7. A simple hybrid control problem.

Consider the following special form of the control problem considered in §4.

System dynamics

$$\begin{aligned} dx_t &= y_n dt + dw_t,\ t\in[n,n+1),\\ y_{n+1} &= y_n + u_n + e_n,\ n = 0,1,2,\ldots\\ x_0 &= x,\ y_0 = y, \end{aligned}$$

where $x_t, y_n \in R^1$, $(w_t)_{t\geq 0}$ is a standard Brownian motion in R^1, and $(e_n)_{n=0}^{\infty}$ is an independently and identically distributed (i.i.d.) sequence of Gaussian random variables with mean zero and variance one.

Cost functional

$$\overline{C}(\overline{u}_\cdot,(x,y)) = \overline{\lim_{T\to\infty}}\, T^{-1}\int_0^T (x_t^2 + \overline{u}_t^2)\, dt \,,$$

where $\overline{u}_t = u_n$, $t\in[n,n+1)$.

The point here is that a continuous-time subsystem corresponding to x_t is controlled via a discrete-time subsystem corresponding to y_n. Using results of §4 we compute optimal controls

$$u_n^o = -\frac{1+\sqrt{13}}{2+\sqrt{13}} y_n, \; n = 0, 1, 2, \ldots,$$

and the optimal cost

$$\overline{C}(\overline{u}^o, (x, y)) = \frac{29 + 3\sqrt{13}}{18} \; .$$

8. Concluding remarks. We refer to the control problem considered in §4 as to the "backward problem." The "forward problem" for which the cost functional is given in terms of

$$\left\{ \int_0^t [x_s^t \widetilde{Q}_1 x_s + v_s^t \widetilde{R}_1 v_s] d\langle M \rangle_s + \int_0^t [x_s^t \widetilde{Q}_s x_s + u_s^t \widetilde{R}_2 u_s] d\langle N \rangle_s \right\}$$

can be studied in the similar way as the "backward problem."

Our formulation of the linear-quadratic stochastic control problem does not allow for a direct consideration of a deterministic linear-quadratic control problem, one of the reasons being that the time scales in (4.1) would vanish for $k_1 = 0$ and/or $k_3 = 0$. An obvious reparametrization will allow for including a deterministic situation in the model (4.1) as well. We have not done that in order to keep the calculations easy. Note that equations (3.1) are serving both deterministic and stochastic situations, as in the "classical" case.

It is still an open question under what nontrivial conditions on the parameters there exists a stationary distribution for $(x_t^0)_{t \geq 0}$. We have some preliminary results for the noncontrolled case corresponding to the one considered by Zabczyk (1983).

In a subsequent paper we shall consider implications of the approach taken here for control and identification of general (multiple time scales) ARMA models represented via a certain integral transform that is given in terms of the predictable quadratic variation process of the driving semimartingale noise.

Acknowledgments. I would like to thank the referees for their valuable remarks, which helped me improve the first version of the manuscript.

Appendix 1. In this appendix we prove Theorem 3.1. We will need the following three technical results, which are counterparts of Theorem 3.6 ii) and Lemmas 12.1 and 12.2 of Wonham (1979).

LEMMA A.1.1. *If $Q \geq 0$ and B is d-stable then the equation*

$$B^T R B + Q = R$$

has a unique solution R and $R \geq 0$.

Proof. $R = \sum_{k=0}^{\infty} (B^T)^k Q \, B^k$. □

LEMMA A.1.2 (d-Liapunov criterion). *Suppose $R \geq 0$, $Q \geq 0$, $(B, \sqrt{Q})$ is d^T-stabilizable and $B^T R B + Q = R$. Then B is d-stable.*

Proof. We have

$$R = (B^T)^k R B^k + \sum_{i=0}^{k-1} (B^T)^i Q B^i, \quad k \geq 0.$$

Assume B is not d-stable and let λ be an eigenvalue of B with $|\lambda| \geq 1$, and X the corresponding eigenvector. We have $X^T RX = |\lambda|^{2k} X^T RX + \sum_{i=0}^{k-1} |\lambda|^{2i} \|\sqrt{Q}\, X\|^2$. This means that $|\lambda| = 1$ and $\sqrt{Q}X = 0$. Let $\widetilde{K}$ be a matrix such that $B + \widetilde{K}\sqrt{Q}$ is d-stable. We see that λ and X are respectively an eigenvalue and corresponding eigenvectors of $B + \widetilde{K}\sqrt{Q}$. This is a contradiction. □

LEMMA A.1.3. *Assume that the four-tuple (A, Q_1, B, Q_2) is cd-detectable. Let $G : [0,1] \to L(R^n, R^n)$ be continuous. Let $G := \int_0^1 G_s ds$ and $M > 0$, $N > 0$, $\Phi \geq 0$, L be arbitrary matrices of appropriate dimensions. Then the pair*

$$\Bigg((B+L)e^{(A+G)},$$
$$\sqrt{Q_2 + L^T NL + \Phi + (B+L)^T \left[\int_0^1 e^{\int_t^1 (A+G_{1-s})^T ds}(Q_1 + G_{1-t}^T M G_{1-t}) e^{\int_0^1 (A+G_{1-s})ds} dt\right]}$$
$$\sqrt{\cdot (B+L)}\Bigg)$$

is d^T-stabilizable.

Proof. It will be shown in Lemma A.2.1 of Appendix 2 that

$$\begin{aligned} \mathrm{Ker} \int_0^1 & e^{\int_0^t (A+G_s)^T ds}[Q_1 + G_t^T M G_t] e^{\int_0^t (A+G_s) ds} dt \\ & \subset \mathrm{Ker}\, e^{A^T} \int_0^1 (Q_1 + G_t^T M G_t) dt e^A. \end{aligned}$$

Therefore we have the following chain of inclusions,

$$\begin{aligned} &\mathrm{Ker}\sqrt{Q_2 + L^T NL + \Phi} \\ &\sqrt{+(B+L)^T \left[\int_0^1 e^{\int_t^1 (A+G_{1-s})^T ds}(Q_1 + G_{1-t}^T M G_{1-t}) e^{\int_t^1 (A+G_{1-s})ds} dt\right](B+L)} \\ &\subset \ \mathrm{Ker}\sqrt{Q_2 + L^T NL + B^T e^{A^T} Q_1 e^A B + (B+L)^T e^{A^T} \int_0^1 G_t^T M G_t dt e^A (B+L)} \\ &\subset \ \mathrm{Ker}\sqrt{Q_2 + B^T e^{A^T} Q_1 e^A B} \cap \ \mathrm{Ker}(-L) \cap \ \mathrm{Ker}(e^A B - e^{A+G} B) \\ &\subset \ \mathrm{Ker}(\sqrt{Q_2 + B^T e^{A^T} Q_1 e^A B} - e^{A+G} L + e^A B - e^{A+G} B). \end{aligned}$$

Let K be such that $e^A B + K\sqrt{Q_2 + B^T e^{A^T} Q_1 e^A B}$ is d-stable. In view of the above inclusions we conclude that there exists matrix $\widetilde{K}$ such that

$$\begin{aligned} &e^{A+G}(B+L) \\ &+ \widetilde{K}\sqrt{Q_2 + L^T NL + \Phi} \\ &\sqrt{+(B+L)^T \left[\int_0^1 e^{\int_t^1 (A+G_{1-s})^T ds}(Q_1 + G_{1-t}^T M G_{1-t}) e^{\int_t^1 (A+G_{1-s} ds} dt\right](B+L)} \\ &= e^A B + K\sqrt{Q_2 + B^T e^{A^T} Q_1 e^A B}. \qquad \square \end{aligned}$$

Proof of Theorem 3.1.

Step 1. Let $R \geq 0$. Consider the control problem

$$J(x_0, R) = \min_{u,v.} J(x_0, u, v., R)$$

where

$$J(x_0, u, v., R) := x_0^T Q_2 x_0 + u^T R_2 u + y_1^T R y_1 + \int_0^1 (y_s^T Q_1 y_s + v_s^T R_1 v_s) ds$$

subject to

$$\begin{cases} x_1 &= Bx_0 + Fu, \\ \dot{y}_t &= Ay_t + Ev_t, \\ y_0 &= x_1, \ t \in [0,1]. \end{cases}$$

It can be easily verified that

$$J(x_0, R) = x_0^T [B^T P_1 B + Q_2 - B^T P_1 F (F^T P_1 F + R_2)^{-1} F^T P_1 B] x_0,$$

and the optimal controls are

$$\begin{aligned} v_t^* &:= L_t(\mathcal{P}) y_t, \ t \in [0,1], \\ u^* &:= L_2(\mathcal{P}) x_0, \end{aligned}$$

where $\mathcal{P}$ corresponds to R in the sense that $\mathcal{P} = P_t, t \in [0,1]$ solves the differential equation in (3.1) with $P_0 = R$.

Step 2. (a) Choose $L_0^1 := \{\widetilde{L}_t^1, \ t \in [0,1]\}$ and L_2^1 so that $\mathcal{A}(L_1^1, L_2^1)$ is d-stable. Here $L_1^1 := \int_0^1 \widetilde{L}_t^1 dt$.

(b) Having chosen $(L_0^1, L_2^1), \dots, (L_0^k, L_2^k)$, obtain $R^k \geq 0$ from (use Lemma A.1.1)

$$\mathcal{A}^T(L_1^k, L_2^k) R^k \mathcal{A}(L_1^k, L_2^k) + \int_0^1 e^{\int_s^1 (A + E\widetilde{L}_{1-t}^k)^T dt} \cdot [Q_1 + (\widetilde{L}_{1-s}^k)^T R_1 \widetilde{L}_{1-s}^k] e^{\int_s^1 (A + E\widetilde{L}_{1-t}^k) dt} ds + Q_2 + (L_2^k)^T R_2 L_2^k = R^k.$$

In the above $L_1^k := \int_0^1 \widetilde{L}_t^k dt$. Note that with regard to the control problem of Step 1 we have $x_0^T R^k x_0 = J(x_0, u^k, v^k., R^k)$, where

$$u^k := L_2^k x_0, \ v_t^k := \widetilde{L}_t^k y_t, \ t \in [0,1].$$

(c) Obtain $\mathcal{P}^{k+1} := \{P_t^{k+1}, t \in [0,1]\}$ from

$$\text{(A.1.1)} \qquad \begin{cases} \dot{P}_1^{k+1} &= Q_1 + A^T P_t^{k+1} + P_t^{k+1} A - P_t^{k+1} E R_1^{-1} E^T P_t^{k+1}, \\ P_0^{k+1} &= R^k, \quad t \in [0,1] \end{cases}$$

and define

$$\begin{aligned} \widetilde{L}_t^{k+1} &:= -R_1^{-1} E^T P_{t-1}^{k+1}, \ t \in [0,1], \\ L_2^{k+1} &:= -(F^T P_1^{k+1} F + R_2)^{-1} F^T P_1^{k+1} B. \end{aligned}$$

Remark A.1.1. In case $B = I$, $Q_2 = 0$, $F = 0$, and $P_t =$ const, $t \in [0,1]$, consider (A.1.1) as

$$0 = Q_1 + (A + EL_1^k)^T R^k + R^k (A + EL_1^k) + (L_1^k)^T R_1 L_1^k$$

and $\widetilde{L}_s^k = L_1, \ s \in [0,1]$.

Note that with regard to the control problem of Step 1 we have

$$x_0^T(B^TP_1^{k+1}B+Q_2-B^TP_1^{k+1}F(F^TP_1^{k+1}F+R_2)^{-1}F^TP_1^{k+1}B)x_0 = J(x_0,R^k) \leq x_0^TR^kx_0.$$

After letting $\mathcal{K}^i := B^TP_1^iB + Q_2 - B^TP_1^iF(F^TP_1^iF + R_2)^{-1}F^TP_1^iB$ we then have

$$\mathcal{K}^{k+1} \leq R^k.$$

Now note that

$$\mathcal{A}_{\mathcal{P}^{k+1}}(R^k) + \mathcal{B}_{\mathcal{P}^{k+1}} = \mathcal{K}^{k+1} - R^k + R^k.$$

Therefore, in view of Lemmas A.1.2 and A.1.3 we conclude that $\mathcal{A}(L_1^{k+1}, L_2^{k+1})$ is d-stable. Obtain R^{k+1} from

$$\mathcal{A}_{\mathcal{P}^{k+1}}(R^{k+1}) + \mathcal{B}_{\mathcal{P}^{k+1}} = R^{k+1}$$

and note that

$$\mathcal{A}^T(L_1^{k+1}, L_2^{k+1})(R^{k+1} - R^k)\mathcal{A}(L_1^{k+1}, L_2^{k+1}) + \mathcal{K}^{k+1} - R^k = R^{k+1} - R^k.$$

Henceforth we have

$$0 \leq R^{k+1} \leq R^k.$$

Thus there exist limits

$$0 \leq \bar{R} = \lim_{k\to\infty} R^k,$$
$$0 \leq \bar{P}_t = \lim_{k\to\infty} P_t^k, \ t \in [0,1],$$

and the pair $(\bar{R}, \bar{\mathcal{P}})$ satisfies (3.1), where $\bar{\mathcal{P}} := \{\bar{P}_t,\ t \in [0,1]\}$. Again by Lemmas A.1.2 and A.1.3, we conclude that $\mathcal{A}(\bar{L}_1, \bar{L}_2)$ is d-stable.

Uniqueness of $(\bar{R}, \bar{\mathcal{P}})$ follows from the following argument. Let $(\widetilde{R}, \widetilde{\mathcal{P}})$ be another solution to (3.1) such that $\widetilde{R} \geq 0$ and $\widetilde{\mathcal{P}}_t \geq 0$, $t \in [0,1]$. In view of the control problem analogous to the one considered in Step 1 we have

$$\mathcal{A}_{\widetilde{\mathcal{P}}}(\bar{R} + \mathcal{B}_{\widetilde{\mathcal{P}}}) \leq \bar{R}.$$

Therefore it holds that

$$\mathcal{A}_{\widetilde{\mathcal{P}}}(\bar{R} - \widetilde{R}) \leq \bar{R} - \widetilde{R},$$

and since $\mathcal{A}(\widetilde{L}_1, \widetilde{L}_2)$ is d-stable (Lemmas A.1.2 and A.1.3 again) we obtain that $\bar{R} - \widetilde{R} \geq 0$. Similarly we get that $\mathcal{A}_{\bar{\mathcal{P}}}(\widetilde{R} - \bar{R}) \leq \widetilde{R} - \bar{R}$ and therefore $\widetilde{R} - \bar{R} \geq 0$. □

Appendix 2.

LEMMA A.2.1. *In the notation of Lemma* A.1.3 *it holds that*

$$\operatorname{Ker} \int_0^1 e^{\int_0^t (A+G_s)^T ds}[Q_1 + G_t^T M G_t] e^{\int_0^t (A+G_s)^T ds} dt$$
$$\subset \operatorname{Ker} e^{A^T} \int_0^1 (Q_1 + G_t^T M G_t) dt \ e^A.$$

Proof. Consider the dynamic system

$$\dot{x}_t = (A + G_t)x_t,$$
$$x_0 = 0, \ t \in [0,1],$$

and the functional

$$I = \int_0^1 x_s^T(Q_1 + G_s^T M G_s)x_s ds.$$

Note that $I = 0$ implies $G_s x_s = 0$ for almost all $s \in [0,1]$ and therefore

$$0 = x_0^T \int_0^1 e^{tA^T}(Q_1 + G_t^T M G_t)e^{tA}dt\ x_0.$$

This implies that

$$x_0^T e^{tA^T} Q_1(t^k A^k)x_0 = x_0^T e^{tA^T} G_t M G_t(t^k A^k)x_0 = 0$$

for $k \geq 0$ and almost all $t \in [0,1]$. Thus we obtain

$$0 = x_0^T e^{A^T} \int_0^1 (Q_1 + G_t^T M G_t)dte^A x_0. \qquad \square$$

Appendix 3.

LEMMA A.3.1. *Under the assumptions of Theorem* 4.1 *the pair of controls* $(v^0., u^0.)$ *defined in* (4.2) *is admissible.*

Proof. Let $(x_t^0)_{t\geq 0}$ be the unique, strong, semimartingale solution to (4.1), under $(v^0., u^0.)$, that exists according to Theorem 4.1(a). It remains to show

$$\text{(A.3.1)} \qquad \overline{\lim_{t\to\infty}} \frac{1}{t}\int_0^t \|x_s^0\|^2 ds < +\infty, \text{ a.s.}$$

and

$$\text{(A.3.2)} \qquad \lim_{t\to\infty} \frac{1}{t}\|x_t^0\|^2 = 0, \text{ a.s.}$$

It is enough to consider the stochastic sequence $\{y_n\}_{n\geq 1}$ where $y_n := x_{n-}^0$, $n \geq 1$. Note that

$$\begin{aligned} y_{n+1} &= \mathcal{A}(\Lambda_1, \Lambda_2)y_n + e_n, \\ y_1 &= x_{1-}^0,\ n \geq 1, \end{aligned}$$

where $\Lambda_1 := \int^1 \bar{\Lambda}_t dt$ and

$$e_n := \int_n^{n+1} e^{\int_s^{n+1}(A+E\bar{\Lambda}_t)dt} dM_s + e^{A+E\Lambda_1}\Delta N_n,\ n \geq 1.$$

Since $\mathcal{A}(\Lambda_1, \Lambda_2)$ is d-stable (Theorem 3.1) we conclude, upon applying the strong law of large numbers for martingales (Lipster and Shiryayev (1989), Thm. 2.6.1), that

$$\lim_{n\to\infty} \frac{\|y_n\|^2}{n} = 0, \text{ a.s.},$$

which in turn implies (A.3.2).

Since $\mathcal{A}(\Lambda_1, \Lambda_2)$ is d-stable then there exists a positive definite matrix G such that

$$\mathcal{A}^T(\Lambda_1, \Lambda_2)G\mathcal{A}(\Lambda_1, \Lambda_2) + 2I = G$$

(compare Lemma A.1.1). Next note that, for $n \geq 1$,

$$\begin{aligned} y_{n+1}^T G y_{n+1} &= y_n^T \mathcal{A}^T(\Lambda_1, \Lambda_2) G \mathcal{A}(\Lambda_1, \Lambda_2) y_n \\ &\quad + 2 y_n^T \mathcal{A}^T(\Lambda_1, \Lambda_2) G e_n + e_n^T G e_n \\ &\leq y_n^T G y_n - \|y_n\|^2 + 2 y_n^T \mathcal{A}^T(\Lambda_1, \Lambda_2) G e_n \\ &\quad + e_n^T G e_n, \quad \text{a.s.} \end{aligned}$$

Therefore we have

$$\begin{aligned} y_{n+1}^T G y_{n+1} + \sum_{k=1}^n \|y_k\|^2 &\leq y_1^T G y_1 \\ &\quad + 2 \sum_{k=1}^n y_k^T \mathcal{A}(\Lambda_1, \Lambda_2) G e_k \\ &\quad + \sum_{k=1}^n e_n^T G e_n, \quad \text{a.s.} \end{aligned}$$

Invoking the law of large numbers for martingales again, we finally obtain

$$\overline{\lim_{n \to \infty}} \frac{1}{n} \sum_{k=1}^n \|y_k\|^2 < +\infty, \quad \text{a.s.},$$

from which (A.3.1) follows. □

References

D. Bertsekas (1976), *Dynamic Programming and Stochastic Control,* Academic Press, New York.

N. Christopeit (1986), *Quasi-least-squares estimation in semimartingale regression models,* Stochastics Stochastics Rep., 16, pp. 255–278.

M. H. A. Davis (1977), *Linear estimation and stochastic control,* Chapman and Hall, London.

C. Dellacherie and P. Meyer (1975), *Probabilités et Potentiel,* I, Herman, Paris.

——— (1980), *Probabilités et Potentiel,* II, Herman, Paris.

——— (1983), *Probabilités et Potentiel,* III, Herman, Paris.

R.J. Elliott and D.D. Sworder (1992), *Control of a hybrid conditionally gaussian process,* J. Optim. Theory Appl., 74, pp. 75–85.

L. Foldes (1990), *Conditions for optimality in the infinite-horizon portfolio-cum-saving problem with semimartingale investments,* Stochastics Stochastics Rep., 29, pp. 133–170.

P. Hall and C. C. Heyde (1980), *Martingale Limit Theory and Its Applications,* Academic Press, New York.

J. Jacod (1979), *Calcul Stochastique et Problèmes des Martingales,* Lecture Notes in Mathematics, 714, Springer, New York.

J. Jacod and A. N. Shriyayer (1987), *Limit Theorems for Stochastic Processes,* Springer, New York.

A. LeBreton and M. Musiela (1988), *Laws of Large Numbers for Semimartingales,* preprint.

R. Sh. Lipster and A. N. Shiryayev (1989), *Theory of Martingales,* Kluwer Academic Publishers, Dordrecht.

C. W. Li and G. L. Blankenship (1986), *Almost sure stability of linear stochastic systems with Poisson process coefficients,* SIAM J. Appl. Math., 46, pp. 875–911.

R. H. Middleton and G. C. Goodwin (1990), *Digital Control and Estimation: A Unified Approach,* Prentice-Hall, Englewood Cliffs, NJ.

P. Protter (1990), *Stochastic Integration and Differential Equations: A New Approach,* Springer, New York.

P. Sundar (1989), *Ergodic Solutions of Stochastic Differential Equations,* Stochastics Stochastics Rep., 28, pp. 65-83.

P. WHITTLE (1983), *Optimization Over Time*, Vol. II, Wiley, Chichester.

M. W. WONHAM (1970), *Random Differential Equations in Control Theory,* in Probabilistic Methods in Applied Mathematics, vol. 2, A. T. Bharucha-Reid, ed., pp. 132–212, Academic Press, New York.

——— (1979), *Linear Multivariable Control: A Geometric Approach,* 2nd ed., Springer, New York.

Y. ZABCZYK (1983), *Stationary distributions for linear equations driven by general noise,* Bull. Acad. Pol. Sci., 31, pp. 197–209.

SIAM J. CONTROL OPTIM.
Vol. 33, No. 5, pp. 1443 1488, September 1995

© 1995 Society for Industrial and Applied Mathematics

008

TOOLS FOR SEMIGLOBAL STABILIZATION BY PARTIAL STATE AND OUTPUT FEEDBACK*

ANDREW TEEL† AND LAURENT PRALY‡

Abstract. We develop tools for uniform semiglobal stabilization by partial state and output feedback. We show, by means of examples, that these tools are useful for solving a variety of problems. One application is a general result on semiglobal output feedback stabilizability when global state feedback stabilizability is achievable by a control function that is uniformly completely observable. We provide more general results on semiglobal output feedback stabilization as well. Globally minimum phase input–output linearizable systems are considered as a special case. Throughout our discussion we demonstrate the usefulness of considering local convergence separate from boundedness of solutions. For the former we employ a sufficient small gain condition guaranteeing convergence. For the latter we rely on Lyapunov techniques.

Key words. semiglobal (practical) stabilizability, uniform complete observability, dynamic output feedback, high gain control, nonlinear small gain

AMS subject classifications. 93D15, 93D09

Notation.

- A function is said to be smooth if it is in C^r, i.e., r times continuously differentiable, for some integer $r \geq 1$.
- $d(t)$ is a time-varying signal contained in a compact set $D \subset \mathbb{R}^d$. It will be appropriate to denote $d(t)$ and its time derivatives $\dot{d}(t)$, $\ddot{d}(t), \ldots$ by the same symbol d, i.e., $d = (d, \dot{d}, \ddot{d}, \ldots)$. Since this aggregated d is still assumed to lie in a compact set, in some cases we shall implicitly introduce the strong requirement that the external disturbance is smooth.
- $\dot{V}_{(0)}$ denotes the function $\frac{\partial V}{\partial x}(x) f(x,d) : \mathbb{R}^l \times D \to \mathbb{R}$ and the subscript (0) refers to equation number (0) of the differential equation

$$\dot{x} = f(x, d(t)). \tag{0}$$

- $|\cdot|$ denotes the Euclidean norm.
- $||\cdot||_{t_0}$ denotes ess–$\sup_{t_0 \leq t \leq \infty} |\cdot|$.
- A function $\gamma : \mathbb{R}_{\geq 0} \to \mathbb{R}_{\geq 0}$ is said to be *of class-K* if it is continuous, strictly increasing, and satisfies $\gamma(0) = 0$. It is *of class-K_∞* if in addition $\gamma(s) \to \infty$ as $s \to \infty$.
- A function $\beta : \mathbb{R}_{\geq 0} \times \mathbb{R}_{\geq 0} \to \mathbb{R}_{\geq 0}$ is said to be of *class-KL* if, for each fixed $t \in \mathbb{R}_{\geq 0}$, the function $\beta(\cdot, t)$ is of class-K and for each fixed $s \in \mathbb{R}_{\geq 0}$ the function $\beta(s, \cdot)$ is decreasing and

$$\lim_{t \to \infty} \beta(s, t) = 0. \tag{1}$$

- A function $f : \mho \to \mathbb{R}_{\geq 0}$, where $\mho$ is an open set of $\mathbb{R}^p$, is said to be *proper on* $\mho$ if the preimage of a compact subset of $\mathbb{R}_{\geq 0}$ is a compact subset of $\mho$.
- A function $f : \mho \to \mathbb{R}_{\geq 0}$ is said to be *positive* (*negative*) *definite on* $\mho'$, a subset of $\mho$, if $f(x)$ is strictly positive (negative) for all x in $\mho'$.

* Received by the editors December 9, 1992; accepted for publication (in revised form) April 1, 1994.

† Department of Electrical Engineering, University of Minnesota, 4-174 EE/CS Building, 200 Union Street S.E., Minneapolis, Minnesota 55455.

‡ Centre d'Automatique et Systèmes, Ecole des Mines de Paris, 35 rue St. Honoré, 77305 Fontainebleau Cédex, France.

- A solution $x(t)$ of an ordinary differential equation is said to be *captured by a set* Γ if $x(t)$ is defined on $[0,+\infty)$ and there exists t_o such that $x(t) \in \Gamma$ for all t in $[t_o, +\infty)$.

1. **A motivating problem and some results.** We are interested in the semiglobal stabilization[1] problem, as it is stated in [3], for example. In subsequent sections, the following four tools for solving semiglobal stabilization problems will be presented: two "backstepping" tools, a robust observer, and a local nonlinear small gain theorem. The usefulness of these tools will be illustrated by examples throughout the paper. Initially, to give the reader a sense for what can be proved with these tools, we will state some general nonlinear output feedback stabilization results which will be proved in a stronger form and with full details in later sections.

We start by considering the output feedback stabilization problem for nonlinear systems in the general form

$$\left\{ \begin{array}{rcl} \dot{z} &=& A(z,u), \\ y &=& C(z) . \end{array} \right. \tag{2}$$

We will make use of the following properties.

DEFINITION 1 (stabilizability). *An equilibrium point $z = 0$ of a dynamical system*

$$\dot{z} = A(z,u) \tag{3}$$

with A a smooth function, z in $\mathbb{R}^n$, and u in $\mathbb{R}$ is said to be globally (respectively, locally exponentially and globally) stabilizable if there exists a smooth function $\bar{u}$ such that $z = 0$ is a globally asymptotically (respectively, locally exponentially and globally asymptotically) stable equilibrium of

$$\dot{z} = A(z, \bar{u}(z)) . \tag{4}$$

DEFINITION 2 (uniform complete observability). *A function $\bar{u}(z)$ is said to be uniformly completely observable (UCO) with respect to the dynamical system (2) if there exist two integers n_y and n_u and a C^1 function Ψ such that, for each solution of*

$$\left\{ \begin{array}{rcl} \dot{z} &=& A(z,u_0), \\ \dot{u}_0 &=& u_1, \\ &\vdots& \\ \dot{u}_{n_u} &=& v, \end{array} \right. \tag{5}$$

we have, for all t where the solution makes sense,

$$\bar{u}(z(t)) = \Psi(y(t), \dots, y^{(n_y)}(t), u_0(t), \dots, u_{n_u}(t)), \tag{6}$$

where $y^{(i)}(t)$ denotes the ith time derivative of y at time t.[2]

Achieving global stabilization by output feedback can be impossible for very simple systems that are globally stabilizable by state feedback even when each component

[1] See Definition 3. Depending on the authors, this type of stabilization is also called "potentially global", "on compacta," or "widely local."

[2] If u is not present in (6) we let $n_u = -1$.

of the state is uniformly completely observable. For example, it was shown in [28] that there is no continuous, finite-dimensional dynamic output feedback to globally stabilize the equilibrium point $z = 0$ of the system

$$(7) \qquad \begin{cases} \dot{z}_1 &= z_2, \\ \dot{z}_2 &= z_2^n + u, \\ y &= z_1, \end{cases}$$

with $n \geq 3$. This is true even though the system is globally feedback linearizable and the state is related to the output by $z_1 = y$, $z_2 = \dot{y}$. For this reason we restrict our attention to the semiglobal stabilization problem.

DEFINITION 3 (semiglobal stabilizability). [3] *The equilibrium $z = 0$ of the system* (2) *is said to be semiglobally stabilizable by dynamic state* (*respectively, output*) *feedback if, for each compact set $\mathcal{K}_l$, a neighborhood of* 0, *there exists a locally Lipschitz dynamic state* (*respectively, output*) *feedback $u = \bar{u}(z, \zeta)$, $\dot{\zeta} = \theta(z, \zeta)$* (*respectively, $u = \bar{u}(y, \zeta)$, $\dot{\zeta} = \theta(y, \zeta)$*) *and a compact set $\mathcal{K}_{\zeta l}$ such that the equilibrium $(z, \zeta) = (0, 0)$ is asymptotically stable, with basin of attraction containing $\mathcal{K}_l \times \mathcal{K}_{\zeta l}$.*

It was shown in [40] that, when each component of the state vector z is UCO, global stabilizability by state feedback implies semiglobal stabilizability by output feedback. An implication of the state being UCO is that any globally stabilizing function $\bar{u}(z)$ is UCO. One might hope that this weaker assumption, existence of a UCO globally stabilizing state feedback, would yield semiglobal stabilizability by output feedback as well. Unfortunately, some difficulties appear in this case when attempting to establish local asymptotic stability. To guarantee this local property, we will impose extra local requirements on the system (4). A sufficient condition, generalizations of which are discussed in §5, is local exponential stability.

THEOREM 1.1. *If the equilibrium point $z = 0$ of the system* (2) *is locally exponentially and globally stabilizable by a UCO and C^2 state feedback, then it is semiglobally stabilizable by dynamic output feedback.*

Otherwise, since the only obstruction is local, we can still achieve semiglobal practical stabilization as summarized in the next definition and theorem.

DEFINITION 4 (semiglobal practical stabilizability). *A point $z = 0$* (*not necessarily an equilibrium*) *is said to be semiglobally practically stabilizable by dynamic state* (*respectively, output*) *feedback if, for each pair of compact sets $(\mathcal{K}_s, \mathcal{K}_l)$, neighborhoods of* $(0, 0)$ *with $\mathcal{K}_s \subset \mathcal{K}_l$, there exists a locally Lipschitz dynamic state* (*respectively, output*) *feedback $u = \bar{u}(z, \zeta)$, $\dot{\zeta} = \theta(z, \zeta)$* (*respectively, $u = \bar{u}(y, \zeta)$, $\dot{\zeta} = \theta(y, \zeta)$*) *and a pair of compact sets $(\mathcal{K}_{\zeta s}, \mathcal{K}_{\zeta l})$ such that all the solutions of the closed-loop system, with initial condition in $\mathcal{K}_l \times \mathcal{K}_{\zeta l}$, are captured by the set $\mathcal{K}_s \times \mathcal{K}_{\zeta s}$.*

THEOREM 1.2. *If the equilibrium point $z = 0$ of the system* (2) *is globally stabilizable by a UCO and C^2 state feedback, then it is semiglobally practically stabilizable by dynamic output feedback.*

The technique for proving these theorems is to exhibit a feedback controller based on the given state feedback controller $\bar{u}$, implemented dynamically using estimates of a sufficient number of derivatives of y provided by an observer and a sufficient number of derivatives of u provided by a suitable dynamic extension. The idea of implementing $\bar{u}$ through dynamic extension comes from the work of Tornambè [43]. That such a dynamically extended state feedback controller can be constructed while retaining semiglobal (practical) stabilizability will be shown using the iterating tool of Lemma

[3] It follows from this definition that a family of feedback laws is involved. This family is indexed by $\mathcal{K}_l$.

2.3. Further, we will show that combining the dynamically extended controller and an appropriate observer still yields semiglobal (practical) stability. Our robust observer tool, Lemma 2.4, will provide the technical result for showing this can be done.

Although different from the technicalities of the proof, the intuition behind our analysis follows from considering the closed-loop behavior as having two phases. During the first phase while we are trying to find, in finite time, an exact estimate of the derivatives of y, we acknowledge that this dynamically extended, estimated state feedback makes no sense. Still we must make sure that there is no finite escape time. To solve this problem, we use the a priori information that the actual control $\bar{u}(z)$ is in a known compact set to disregard any estimation $\widehat{\bar{u}(z)}$ which would lie outside this set. Mathematically there are many ways to reject these bad estimates but a very simple and efficient way is to saturate the estimated control as proposed in [11]. Then, using the worst saturated control, we can estimate the smallest time period T which will be needed by the system to go from its initial compact set to some larger compact set on which our estimated state semiglobally stabilizing feedback is valid. This time period is the period within which we should get our exact estimate of the derivatives of y. (See [18, Rem. 5].)

In phase two, if the estimates of the derivatives of y were correct, we could apply our dynamically extended state feedback. Unfortunately we are not always able to get an exact estimate of the derivatives of y. This is due to the possible presence of unobserved states and possible uncertainty in A. However, we can obtain an arbitrarily good approximate estimate. If we have designed our (dynamic) state feedback control using Lyapunov methods we have a measure of the stability robustness achieved by our feedback controller via the derivative of the Lyapunov function. We consequently build our approximate observer to account for this robustness margin. This strategy in fact has been exactly applied in [13] but in discrete time. There, no finite escape time is possible and exact estimation in finite time is assumed.

Finally note that Theorems 1.1 and 1.2 as well as the other results on semiglobal stabilization to come are presented here only as existence results. Nevertheless, the practical significance of the dynamic output feedback we shall exhibit has been investigated in the context of robotics applications in [1] and [2].

The remainder of the paper is organized as follows. In §2 we present tools for semiglobal practical stabilization by state and output feedback. Several applications of these tools are presented including, in §3, the proof of Theorem 1.2 and generalizations. In §4 we present a small gain theorem for local asymptotic stability analysis. This tool along with the tools of §2 are used, in §5, to prove Theorem 1.1 and generalizations. In §6 several corollaries for minimum phase input–output linearizable nonlinear systems are presented. A nonminimum phase example is also discussed.

2. Tools for semiglobal practical stabilization.

As exhibited by the statement of our two theorems, we have found it is very useful to decouple the local convergence analysis in robust semiglobal stabilization problems from the analysis regarding the boundedness of solutions. In this section we are concerned only with the problem of uniform semiglobal practical stabilization by partial state and output feedback. We defer study of local convergence until §4. We will present tools that will be used to construct an output feedback for proving Theorem 1.2. However, these tools have their own interest. As illustrated by examples throughout this section, they can be used to address a wide variety of control problems.

These tools provide conditions under which solutions starting in some compact set are captured by a "smaller" one. They consider systems with an interconnection structure and a state decomposed, accordingly, into two parts, say (z, x), where the $\dot{x}$ equation contains a large gain, say K. The effect of this large gain K is to introduce an exponential dichotomy between the x component and the z component. This implies the existence of a center-stable manifold which can be described by $x = H(z, K)$. It follows that the motion of the solutions can be decomposed into two stages: *convergence* to this manifold and *sliding* along this manifold. This decomposition has been a standard tool used to prove early semiglobal stabilization results (see [36] and [7] for example). Here instead, like in [4], we completely ignore this decomposition and use a Lyapunov argument showing the decrease of an energy function outside a neighborhood of the origin. More precisely what is implicitly used here is the fact that as $K \to \infty$, the manifold tends to the set $\{(z, x) : x = 0\}$. So a Lyapunov function, which is simply the sum of the energy functions of x and z separately, should be sufficient and indeed it is. The availability of a Lyapunov function is extremely useful. It makes explicit the ultimate bound on trajectories as well as the domain of attraction without the formalism of invariant manifolds. It will also allow us to use our tools consecutively.

We will present two closely related "backstepping" tools, to borrow the terminology of [17]. This will be followed by an observer tool useful for analysis when the parameter K comes from a high gain observer. These tools are based on the following technical lemma, inspired by a similar result in [4].

LEMMA 2.1. *Let S be a compact set in a product space $\mathbb{R}^m \times \mathbb{R}^n$, and denote by S_z and S_x its respective projections (i.e., $S \subset S_z \times S_x$). Let $\chi(z)$ be a continuous real function on S_z which is positive definite on the projection of the set $\{(z, x) : x = 0\} \cap S$. Let $\psi(x)$ be a continuous real function on S_x which is positive definite on $S_x \backslash \{0\}$. Let $\varphi(z, x, d)$ be a continuous real function on $S \times D$ which satisfies*

$$\varphi(z, x, d) = 0 \qquad \forall (z, x, d) \in (\{(z, x) : x = 0\} \cap S) \times D. \tag{8}$$

Let κ be a function of class-K_∞. Under these conditions, there exists a positive real number K_* such that, for all $K \geq K_*$,

$$-\chi(z) - \kappa(K)\psi(x) + \varphi(z, x, d) < 0 \qquad \forall (z, x, d) \in S \times D. \tag{9}$$

Proof. For purposes of contradiction, assume the result is false. This implies that, for each n, there exists a point (z_n, x_n, d_n) in $S \times D$ such that

$$-\chi(z_n) - \kappa(n)\psi(x_n) + \varphi(z_n, x_n, d_n) \geq 0. \tag{10}$$

Consequently, since κ is class-K_∞ and $\psi \geq 0$ we have, for each $m \geq 1$ and for all $n \geq m$,

$$-\chi(z_n) - \kappa(m)\psi(x_n) + \varphi(z_n, x_n, d_n) \geq 0. \tag{11}$$

Now, since $S \times D$ is compact, the (sub)sequence (z_n, x_n, d_n) converges to a point (z_*, x_*, d_*) in $S \times D$. By continuity, this point satisfies

$$-\chi(z_*) - \kappa(m)\psi(x_*) + \varphi(z_*, x_*, d_*) \geq 0 \tag{12}$$

for all $m \geq 1$. Then, if $x_* = 0$, (8), (12), and the properties of ψ imply $-\chi(z_*) \geq 0$ which is not possible since χ is strictly positive on the projection of the set $\{(z, x) :$

$x = 0\} \cap S$. On the other hand, if $|x_*| \neq 0$ then $\psi(x_*) > 0$ and there exists an $m_* \geq 1$ such that, for all $m \geq m_*$,

$$-\chi(z_*) - \kappa(m)\psi(x_*) + \varphi(z_*, x_*, d_*) < 0 \tag{13}$$

since $\kappa(\cdot)$ is of class-K_∞. This contradicts (12), however, and completes the proof. □

Throughout the remainder of this section we use the following assumption.

Assumption ULP (*uniform Lyapunov property*)[4]. For the system

$$\dot{z} = h(z, 0, d(t)), \tag{14}$$

there exists an open set $\mho_1$ in $\mathbb{R}^m$, a nonnegative real number $\vartheta < 1$, a real number $c \geq 1$, and a C^1 function $V : \mho_1 \to \mathbb{R}_{\geq 0}$ such that the set $\{z : V(z) \leq c + 1\}$ is a compact subset of $\mho_1$, and we have

$$\dot{V}_{(14)} \leq -\Phi_1(z), \tag{15}$$

where $\Phi_1(z)$ is continuous on $\mho_1$ and positive definite on the set $\{z : \vartheta < V(z) \leq c+1\}$.

Remark 2.1. In the absence of $d(t)$, if the equilibrium $z = 0$ of the system

$$\dot{z} = h(z, 0) \tag{16}$$

is locally asymptotically stable with domain of attraction $\mho_1$, the converse Lyapunov theorem [22, Thm. 7] provides a smooth Lyapunov function satisfying Assumption ULP. Further, ϑ can be chosen to be equal to zero and c can be chosen to be arbitrarily large.

We now present our backstepping tools. The first lemma shows how one can semiglobally practically stabilize from a disturbed first derivative of the control instead of the control itself. The second lemma allows, in one step, the designer to semiglobally practically stabilize from a jth disturbed derivative of the control when the perturbations have a special form.

LEMMA 2.2 (semiglobal backstepping I). *Consider the C^1 nonlinear control system*

$$\left\{ \begin{array}{rcl} \dot{z} &=& h(z, x, d(t)), \\ \dot{x} &=& f(z, x, d(t)) + g(z, x, d(t))u, \end{array} \right. \tag{17}$$

where $x \in \mathbb{R}$, $z \in \mathbb{R}^m$, the sign of $g(z, x, d)$ is constant, and the magnitude of g is bounded away from zero by a strictly positive real number b

$$|g(z, x, d)| \geq b \qquad \forall (z, x, d) \in \mathbb{R}^m \times \mathbb{R} \times D. \tag{18}$$

Suppose Assumption ULP is satisfied. Given $\mu \geq 1$, we define the function

$$W(z, x) = c\frac{V(z)}{c + 1 - V(z)} + \mu\frac{x^2}{\mu + 1 - x^2} \tag{19}$$

and the set

$$\mho_2 = \{z : V(z) < c + 1\} \times \{x : x^2 < \mu + 1\}. \tag{20}$$

[4] The number "1," here and in the following, is arbitrary and could be replaced by any strictly positive real number.

Under these conditions, $W(z,x) : \mho_2 \to \mathbb{R}_{\geq 0}$ is proper on $\mho_2$. Further, if

$$u = -K\mathrm{sgn}(g)x \tag{21}$$

then, for each strictly positive real number ρ, there exists a positive real number K_*, such that, for each $K \geq K_*$, W satisfies

$$\dot{W}_{(17)} \leq -\Phi_2(z,x), \tag{22}$$

where $\Phi_2(z,x)$ is continuous on $\mho_2$ and positive definite on the set $\{(z,x) : \vartheta + \rho \leq W(z,x) \leq c^2 + \mu^2 + 1\}$.

Proof of Lemma 2.2. With $u = -K\mathrm{sgn}(g)x$ the closed loop system is

$$\left\{ \begin{array}{rcl} \dot{z} & = & h(z,x,d(t)), \\ \dot{x} & = & f(z,x,d(t)) - K\mathrm{sgn}(g)g(z,x,d(t))x. \end{array} \right. \tag{23}$$

For sake of generality we replace x^2 in (19) by $U(x)$. Now assume $W(z,x) \leq c^2+\mu^2+1$. This implies

$$V(z) \leq (c+1)\frac{c^2+\mu^2+1}{c^2+\mu^2+1+c}, \qquad U(x) \leq (\mu+1)\frac{c^2+\mu^2+1}{c^2+\mu^2+1+\mu} . \tag{24}$$

Now, we have

$$\dot{W}_{(23)} = \frac{c(c+1)}{(c+1-V)^2}\dot{V}_{(23)} + \frac{\mu(\mu+1)}{(\mu+1-U)^2}\dot{U}_{(23)}. \tag{25}$$

From (24), we get, when $W(z,x) \leq c^2 + \mu^2 + 1$,

$$\begin{array}{c} \frac{c}{c+1} \leq \frac{c(c+1)}{(c+1-V)^2} \leq \frac{(c^2+\mu^2+1+c)^2}{c(c+1)}, \\ \frac{\mu}{\mu+1} \leq \frac{\mu(\mu+1)}{(\mu+1-U)^2} \leq \frac{(c^2+\mu^2+1+\mu)^2}{\mu(\mu+1)}. \end{array} \tag{26}$$

Then, let us define

$$\left\{ \begin{array}{rcl} \chi(z) & = & \frac{c(c+1)}{2(c+1-V(z))^2}\Phi_1(z), \\ \psi(x) & = & \frac{\mu}{\mu+1}bx^2, \\ \varphi(z,x,d) & = & \frac{(c^2+\mu^2+1+c)^2}{c(c+1)} \left|\frac{\partial V}{\partial z}(z)[h(z,x,d) - h(z,0,d)]\right| \\ & & + \frac{(c^2+\mu^2+1+\mu)^2}{\mu(\mu+1)} 2\left|xf(z,x,d)\right|, \\ \kappa(K) & = & K, \end{array} \right. \tag{27}$$

and consider the left-hand side of (9) in Lemma 2.1. We pick an arbitrarily small but strictly positive real number ρ and define a set S by

$$S = \{(z,x) : \vartheta + \rho \leq W(z,x) \leq c^2 + \mu^2 + 1\}. \tag{28}$$

The set S is compact from (24) and Assumption ULP. Also, from (24) the projections of S satisfy

$$S_z \subset \{z : V(z) < c+1\}, \qquad S_x \subset \{x : x^2 < \mu + 1\}. \tag{29}$$

Consequently $\chi(z)$ is continuous on S_z and $\psi(x)$ is continuous on S_x and positive definite on $S_x \backslash \{0\}$. Further, (24) also implies that $\varphi(z,x,d)$ is continuous on $S \times D$. From (27), it follows that

$$\varphi(z,x,d) = 0 \qquad \forall (z,x,d) \in (\{(z,x) : x = 0\} \cap S) \times D. \tag{30}$$

Finally, to see that $\chi(z)$ is positive definite on the projection of the set $\{(z,x) : x = 0\} \cap S$, we note

$$\{x = 0\ ,\ \vartheta + \rho \le W(z,x)\} \quad \Longrightarrow \quad \vartheta + \rho \le c\frac{V(z)}{c+1-V(z)}. \tag{31}$$

Further, for $0 \le \vartheta \le 1$,

$$\vartheta + \rho \le c\frac{V(z)}{c+1-V(z)} \quad \Longrightarrow \quad \vartheta < V(z) \qquad \forall \rho > 0. \tag{32}$$

Then, from Assumption ULP, $\chi(z)$ is positive definite on the projection of the set $\{(z,x) : x = 0\} \cap S$. This demonstrates that the conditions of Lemma 2.1 are satisfied. It follows that there exists a positive real number K_* such that, for all $K \ge K_*$, (22) is satisfied with

$$\Phi_2(z,x) = \frac{c(c+1)}{2(c+1-V)^2}\Phi_1(z) + \frac{\mu}{\mu+1}Kbx^2. \tag{33}$$

Also, since φ is positive, it follows, from (27) and (9), that this function Φ_2 is positive definite on $\{(z,x) : \vartheta + \rho \le W(z,x) \le c^2 + \mu^2 + 1\}$ for all $K \ge K_*$. □

Example 2.1. The first application of Lemma 2.2 is a result for the C^1 control system:

$$\left\{ \begin{array}{rcl} \dot z &=& A(z,\zeta), \\ \dot \zeta &=& F(z,\zeta,d(t)) + G(z,\zeta,d(t))u, \end{array} \right. \tag{34}$$

$z \in \mathbb{R}^m$, $\zeta \in \mathbb{R}$, where the sign of $G(z,\zeta,d(t))$ is constant and the magnitude of G is bounded away from zero. Specifically, *if the equilibrium point $z = 0$ is semiglobally stabilizable by C^ℓ ($\ell \ge 2$) state feedback, with ζ as control, then $(z,\zeta) = (0,0)$ is semiglobally practically stabilizable by C^ℓ state feedback.*

This statement is to be added to the many results known on the stabilization via a disturbed derivative of the input ([6], [9], [12], [44]). Its proof follows.

Let $\bar u(z)$ represent the control law we get once the compact set $\mathcal{K}_{zl}$ of the semiglobal stabilizability property for the z subsystem is chosen. Define $x = \zeta - \bar u(z)$. Then we have

$$\left\{ \begin{array}{rcl} \dot z &=& A(z,\bar u(z) + x) \\ &\doteq& h(z,x) \\ \dot x &=& F(z,x+\bar u(z),d(t)) + G(z,x+\bar u(z),d(t))u - \frac{\partial \bar u}{\partial z}(z)A(z,\bar u(z)+x) \\ &\doteq& f(z,x,d(t)) + g(z,x,d(t))u. \end{array} \right. \tag{35}$$

From [22, Thm. 7], Assumption ULP is satisfied with $\vartheta = 0$ and a positive definite function V such that $\mathcal{K}_{zl}$ is contained in the set $\{z : V(z) \le c\}$ for some real number $c \ge 1$.

With $\mathcal{K}_{\zeta l} \subset \mathbb{R}$, a chosen compact set, we choose μ to satisfy

$$\mu \ge \max\left\{1\ ,\ \max_{\{z\in\mathcal{K}_{zl},\zeta\in\mathcal{K}_{\zeta l}\}}\{\zeta - \bar u(z)\}\right\}\ . \tag{36}$$

Similarly, let $\mathcal{K}_s$, a neighborhood of $(0,0)$, be the compact set we want the solutions of the closed-loop system to be captured by. We choose ρ to satisfy

$$0 < \rho \le \min\left\{1\,,\,\frac{1}{2}\inf_{(z,\zeta)\notin\mathcal{K}_s}\left\{\max\{V(z),(\zeta-\bar{u}(z))^2\}\right\}\right\}\ . \tag{37}$$

With these choices, the function $W(z,x)$ defined in (19) satisfies

$$W(z,\zeta-\bar{u}(z)) \le \rho \implies V(z) < 2\rho\ ,\ (\zeta-\bar{u}(z))^2 < 2\rho\ , \tag{38}$$

$$\implies (z,\zeta)\in\mathcal{K}_s\ . \tag{39}$$

Now, from Lemma 2.2, if u is chosen to be of the form

$$u = -K\mathrm{sgn}(g)x = -K\mathrm{sgn}(G)[\zeta-\bar{u}(z)], \tag{40}$$

then there exists a positive real number K_* such that, for each $K \ge K_*$, (22) holds with $\Phi_2(z,x)$ positive definite on the set $\{(z,x) : \rho \le W(z,x) \le c^2+\mu^2+1\}$. We conclude that, for each initial condition $(z(0),\zeta(0))$ in $\mathcal{K}_{zl}\times\mathcal{K}_{\zeta l}$, the corresponding solution of (34), (40) is captured by the set $\{(z,\zeta) : W(z,\zeta-\bar{u}(z)) \le \rho\}$ and therefore by $\mathcal{K}_s$. Since this holds for any compact sets $\mathcal{K}_{zl}$, $\mathcal{K}_{\zeta l}$, and $\mathcal{K}_s$, the semiglobal practical stabilizability result follows.

Now, if $y = C(z)$ is an output function, the discussion above and the very special structure of (40) yields the following result.

If the equilibrium point $z = 0$ is semiglobally stabilizable by C^ℓ ($\ell \ge 2$) and UCO state feedback, with ζ as control, and ζ is UCO, then the point $(z,\zeta) = (0,0)$ is semiglobally practically stabilizable by C^ℓ and UCO state feedback.

Example 2.2 (almost disturbance decoupling). A solution to the almost disturbance decoupling problem as described in [25] can be obtained by repeated application of Lemma 2.2 for systems that can be put in the following form:

$$\left\{\begin{array}{rcl} \dot{z} & = & h(z,x_1),\\ \dot{x}_1 & = & x_2+f_1(z,x_1,d(t)),\\ \dot{x}_2 & = & x_3+f_2(z,x_1,x_2,d(t)),\\ & \vdots & \\ \dot{x}_{r-1} & = & x_r+f_{r-1}(z,x_1,\dots,x_{r-1},d(t)),\\ \dot{x}_r & = & f_r(z,x_1,\dots,x_r,d(t))+g(z,x_1,\dots,x_r,d(t))u, \end{array}\right. \tag{41}$$

where the equilibrium point $z = 0$ of $\dot{z} = h(z,0)$ is globally asymptotically stable and where the sign of g is constant and the magnitude of g is bounded away from zero. This is illustrated by the following example (compare with (7)):

$$\left\{\begin{array}{rcl} \dot{x}_1 & = & x_2+d_1(t),\\ \dot{x}_2 & = & x_2^3d_2(t)+d_3(t)+u,\\ y & = & x_1, \end{array}\right. \tag{42}$$

where $d_1(t), d_2(t), d_3(t)$ are unknown bounded disturbances. The problem is to achieve $|x_1(t)| \le \wp \le 1$ asymptotically from arbitrarily large domains of attraction. Without loss of generality we assume $|d_i(t)| \le 1$.

Assume the initial conditions satisfy $|x_i(0)|^2 \le c$. We first consider the intermediate subsystem

$$\dot{x}_1 = u_1 + d_1(t). \tag{43}$$

If we choose the control $u_1 = -K_1 x_1$, and the Lyapunov function candidate $W_1(x_1) = x_1^2$, then for the intermediate closed-loop system

$$\dot{x}_1 = -K_1 x_1 + d_1(t) \tag{44}$$

we have

$$\dot{W}_{1_{(44)}} \leq -2x_1[K_1 x_1 - d], \tag{45}$$

which is negative definite on the set $\{x : \frac{1}{(K_1-1)^2} \leq W_1(x_1)\} \times \{|d| \leq 1\}$. We then choose

$$K_1 = 1 + \frac{2}{\wp} \tag{46}$$

so that $\dot{W}_{1_{(44)}}$ is negative definite on the set $\{x : \frac{\wp^2}{4} \leq W_1(x_1)\} \times \{|d| \leq 1\}$. We now make the coordinate change $\zeta = x_2 + K_1 x_1$ to get the system

$$\left\{ \begin{array}{rcl} \dot{x}_1 & = & -K_1 x_1 + \zeta + d_1(t), \\ \dot{\zeta} & = & u + (\zeta - K_1 x_1)^3 d_2(t) + K_1(\zeta - K_1 x_1 + d_1(t)) + d_3(t). \end{array} \right. \tag{47}$$

By applying Lemma 2.2 with $\vartheta = \rho = \frac{\wp^2}{4}$, we get the final control

$$u = -K_2 \zeta = -K_2 x_2 - K_2 K_1 x_1 \tag{48}$$

and a Lyapunov function candidate

$$W_2(x_1, \zeta) = \frac{\mu_1 W_1(x_1)}{\mu_1 + 1 - W_1(x_1)} + \frac{\mu_2 \zeta^2}{\mu_2 + 1 - \zeta^2}, \tag{49}$$

where $\mu_1 = c$ and μ_2 is so that the initial value of ζ satisfies $\zeta^2 \leq \mu_2$, i.e., $\mu_2 = (1 + K_1)^2 c$. We then have that the initial condition satisfies $W_2(x_1, \zeta) \leq \mu_1^2 + \mu_2^2$. Also, we know that, for K_2 large enough (see [41] for an explicit expression), the time derivative of W_2 is negative definite on the compact set

$$\Gamma_2 = \left\{ (x_1, \zeta) : \frac{\wp^2}{2} \leq W_2(x_1, \zeta) \leq \mu_1^2 + \mu_2^2 + 1 \right\}. \tag{50}$$

Therefore, the solutions, with $|x_i(0)|^2 \leq c$, are captured by the set $\{(x_1, \zeta) : W_2(x_1, \zeta) \leq \frac{\wp^2}{2}\}$, contained in the set $\{(x_1, \zeta) : |x_1| \leq \wp\}$.

It is important to note that, with our controller (48), we do not have the vanishing regions of attraction phenomenon as described in [21] and [25]. Indeed, in these papers, the same type of high gain controller is proposed but with the implicit constraint that $K_2 = K_1$. Here, instead, our iterative design leads to gains such that the ratio K_2/K_1 tends to infinity as K_1 tends to $+\infty$. However, although x_1 and ζ can be made ultimately arbitrarily small, x_2, called the peaking component, remains of unity magnitude as long as d_1 is present. For a discussion of the peaking phenomenon in feedback systems, see [36] and the references therein.

Finally, we remark that, if $\dot{d}_1(t)$ has a known bound (see our notation section), by applying our forthcoming robust observer tool, Lemma 2.4, the almost disturbance decoupling problem for the system (42) can be solved semiglobally by output feedback. (See [41].)

It is possible to handle a block of integrators in one step, instead of iterating the application of Lemma 2.2, when the system has the structure described in the following lemma.

LEMMA 2.3 (semiglobal backstepping II). *Consider the C^1 nonlinear control system*

$$(51)\qquad \left\{\begin{array}{rcl} \dot z &=& h(z,x_1,d(t)),\\ \dot x_1 &=& x_2+f_1(z,x_1,d(t)),\\ \dot x_2 &=& x_3+f_2(z,x_1,d(t)),\\ &\vdots& \\ \dot x_{j-1} &=& x_j+f_{j-1}(z,x_1,d(t)),\\ \dot x_j &=& u+f_j(z,x_1,d(t)), \end{array}\right.$$

where $x=(x_1,\dots,x_j)^T\in\mathbb{R}^j$, $z\in\mathbb{R}^m$. Suppose Assumption ULP is satisfied. Let the polynomial

$$(52)\qquad p(s)=s^j+a_js^{j-1}+\cdots+a_1$$

be Hurwitz and let A_c be the companion form matrix corresponding to $p(s)$. Also let P_c solve the matrix equation $A_c^TP_c+P_cA_c=-I$. For $K\geq 1$ to be specified, define the variables

$$(53)\qquad \xi_i=\frac{x_i}{K^{i-1}},\qquad i=1,\dots,j.$$

Then given $\mu\geq 1$, define the function

$$(54)\qquad W(z,\xi)=c\frac{V}{c+1-V}+\mu\frac{\xi^TP_c\xi}{\mu+1-\xi^TP_c\xi}$$

and the set

$$(55)\qquad \mho_2=\{z:V(z)<c+1\}\times\{\xi:\xi^TP_c\xi<\mu+1\}.$$

Under these conditions, $W(z,\xi):\mho_2\to\mathbb{R}_{\geq 0}$ is proper on $\mho_2$. Also, if

$$(56)\qquad u=-K^j(a_1\xi_1+\cdots+a_j\xi_j)\ ,$$

then, for each strictly positive real number ρ, there exists a positive real number $K_*\geq 1$ such that, for all $K\geq K_*$, W satisfies

$$(57)\qquad \dot W_{(17)}\leq-\Phi_2(z,\xi),$$

where $\Phi_2(z,\xi)$ is continuous on $\mho_2$ and positive definite on the set $\{(z,\xi):\vartheta+\rho\leq W(z,\xi)\leq c^2+\mu^2+1\}$.

Proof of Lemma 2.3. With the control (56) and the coordinates ξ defined in (53), the closed-loop system becomes

$$(58)\qquad \left\{\begin{array}{rcl} \dot z &=& h(z,\xi_1,d(t)),\\ \dot\xi &=& KA_c\xi+F_K(z,\xi_1,d(t)), \end{array}\right.$$

where

$$(59)\qquad F_K(z,\xi_1,d)=\begin{pmatrix} f_1(z,\xi_1,d)\\ \frac{1}{K}f_2(z,\xi_1,d)\\ \vdots\\ \frac{1}{K^{j-1}}f_j(z,\xi_1,d) \end{pmatrix}.$$

From here, if we replace $\xi^T P_c \xi$ in (54) by $U(\xi)$, then we can follow the proof of Lemma 2.2 with the only modifications being that, in (27) and (33), x^2 is replaced by $\frac{1}{2}\xi^T\xi$, xf by $\xi^T P_c F_K$, and $b = 1$. The fact that $F_K(z, \xi_1, d)$ depends on K is immaterial because, for $K \geq 1$, $\xi^T P_c F_K$ can be bounded by a continuous function which is independent of K. □

Remark 2.2. Lemma 2.3 is used in the same manner as Lemma 2.2 in Example 2.1. One difference is that the free parameter μ is chosen so that the initial conditions of x satisfy $\xi^T P_c \xi \leq \mu$ with ξ defined as in (53). The parameter μ thus appears to depend on K. However, for $K \geq 1$, we have

$$x^T P_c x \leq \mu \qquad \Longrightarrow \qquad \xi^T P_c \xi \leq \mu \frac{\lambda_{\max}\{P_c\}}{\lambda_{\min}\{P_c\}}, \tag{60}$$

where the left-hand side can be achieved independent of K. Nevertheless, the inequality (57) will not guarantee that x ultimately becomes small but only that (z, ξ) ultimately becomes small. As mentioned in Example 2.2, the coordinates x are called peaking coordinates.

Example 2.3 (observer canonical form). We have used Lemma 2.3 as a tool in [42] to design a semiglobally stabilizing output feedback for the following class of systems:

$$\left\{\begin{array}{rcl} \dot{z} & = & h(z, x_1), \\ \dot{x}_1 & = & x_2 + f_1(z, x_1), \\ & \vdots & \\ \dot{x}_r & = & u + f_r(z, x_1), \\ y & = & x_1 \end{array}\right. \tag{61}$$

under a global minimum phase assumption (the point $z = 0$ of the system $\dot{z} = h(z, 0)$ is globally asymptotically stable) and a small gain-based assumption which guarantees local convergence. Here, h and f_i are C^1 and u in $\mathbb{R}$. The special form of (61) permits a technique for output feedback stabilization different from the one mentioned at the end of §1 and used in the proof of Theorem 1.2. Here, on the contrary, we design the observer first, then we define the controller in such a way that the stability it provides is robust to the estimation errors. Our algorithm is inspired by the global results in [17], [26], [27], and [29]. We begin by building the dynamic compensator

$$\left\{\begin{array}{rcl} \dot{\hat{x}}_1 & = & \hat{x}_2 + \ell_1(x_1 - \hat{x}_1), \\ & \vdots & \\ \dot{\hat{x}}_r & = & u + \ell_r(x_1 - \hat{x}_1), \end{array}\right. \tag{62}$$

where the coefficients ℓ_i are the coefficients of a Hurwitz polynomial. If we define $e_i = x_i - \hat{x}_i$ we get the error dynamics

$$\dot{e} = A_o e + F(z, x_1). \tag{63}$$

We choose to consider the dynamics

$$\dot{e} = A_o e + F(z, 0) \tag{64}$$

as an augmentation of the zero dynamics $\dot{z} = h(z, 0)$ so that the equilibrium point $(z, e) = (0, 0)$ of the augmented system

$$\left\{\begin{array}{rcl} \dot{z} & = & h(z, 0), \\ \dot{e} & = & A_o e + F(z, 0) \end{array}\right. \tag{65}$$

is globally asymptotically stable. This follows from the cascade structure and that the state e is input-to-state stable with respect to the input z. (See [35] or Lemma 4.1.) Now we consider the complete system

$$(66)\qquad \left\{\begin{array}{rcl} \dot{z} &=& h(z,x_1), \\ \dot{e} &=& A_o e + F(z,x_1), \\ \hline \dot{x}_1 &=& \hat{x}_2 + e_2 + f_1(z,x_1), \\ \dot{\hat{x}}_2 &=& \hat{x}_3 + \ell_2 e_1, \\ &\vdots& \\ \dot{\hat{x}}_r &=& u + \ell_r e_1. \end{array}\right.$$

It is in the form (51), and we can apply Lemma 2.3 to construct a controller depending only on $x_1, \hat{x}_2, \ldots, \hat{x}_r$ and achieving bounded trajectories from a given compact set of initial conditions. Local exponential stability of the point $z = 0$ of the system $\dot{z} = h(z,0)$ is a sufficient condition to guarantee convergence to the equilibrium $(z, e, x_1, \hat{x}_2, \ldots, \hat{x}_r) = (0, \ldots, 0)$. This condition can be relaxed by using the tools of §4.

When the output feedback stabilization problem is approached from the point of view discussed in §1, a linear high gain observer is introduced to get approximations of the derivatives of the output. The high gain parameter is tuned according to size of the compact set of initial conditions and the stability robustness that would be achieved by the state feedback controller. However, the linear high gain observer introduces possibly very large values of the state estimate over a short period of time. As already noted, this means that during this short period of time, the state estimate makes no sense and should be disregarded. This was achieved in [11] by saturating the control when the estimates had a value which was known a priori to be unreachable within this period of time by the actual state. The success of this modification was demonstrated by using a singular perturbation approach. However, the result seemed to require a form of nonlocal exponential stability [11, Assump. 2]. Even the more general interconnection conditions of [30] on which this assumption is based are too restrictive for the problem of boundedness (only) of solutions from compact sets. These assumptions mix the local and nonlocal analysis while weaker assumptions can be imposed if these aspects are handled separately. The next lemma demonstrates this.

LEMMA 2.4 (robust observer [11]). *Consider the C^1 nonlinear system*

$$(67)\qquad \left\{\begin{array}{rcl} \dot{z} &=& h(z,e,d(t)), \\ \dot{e} &=& LA_o e + p(z,e,d(t)), \end{array}\right.$$

where $z \in \mathbb{R}^m$, $e \in \mathbb{R}^n$, and L is a strictly positive real number. Suppose Assumption ULP is satisfied and let

$$(68)\qquad \Gamma = \{z : V(z) \leq c+1\} \ .$$

Also assume the matrix A_o is Hurwitz and there exist positive real numbers ν_1 and ν_2 and a bounded continuous function γ with $\gamma(0) = 0$ satisfying

$$(69)\qquad \left.\begin{array}{rcl} |h(z,e,d) - h(z,0,d)| &\leq& \gamma(|e|) \\ |p(z,e,d)| &\leq& \nu_1 + \nu_2|e| \end{array}\right\} \quad \forall (z,e,d) \in \Gamma \times \mathbb{R}^n \times D.$$

Let $\mu(L)$ be a class-K_∞ function satisfying

$$\liminf_{L\to\infty} \frac{L}{\mu^4(L)} \to \infty \ . \tag{70}$$

Let P_o solve the matrix equation $A_o^T P_o + P_o A_o = -I$ and, finally, define the function

$$W(z,e) = c\frac{V(z)}{c+1-V(z)} + \mu(L)\frac{\ln(1+e^T P_o e)}{\mu(L)+1-\ln(1+e^T P_o e)} \tag{71}$$

and the set

$$\mho_2 = \{z : V(z) < c+1\} \times \{e : \ln(1+e^T P_o e) < \mu(L)+1\} \ . \tag{72}$$

Under these conditions, for each strictly positive real number L, the function $W(z,e) : \mho_2 \to \mathbb{R}_{\geq 0}$ is proper on $\mho_2$. Also, for each strictly positive real number ρ, there exists a positive real number L_* such that, for all $L \geq L_*$, W satisfies

$$\dot{W}_{(67)} \leq -\Phi_2(z,e), \tag{73}$$

where $\Phi_2(z,e)$ is continuous on $\mho_2$ and positive definite on the set $\{(z,e) : \vartheta + \rho \leq W(z,e) \leq c^2 + \mu^2(L) + 1\}$.

Remark 2.3. The motivation for allowing μ to depend on L, in contrast to the previous two lemmas, is to allow the initial conditions of e to possibly depend on L. If the initial conditions of e can be bounded independent of L, then

1. the bounds in (69) are not needed,
2. μ can be chosen independent of L and the function $\ln(1+e^T P_o e)$ in (71) can be replaced by $e^T P_o e$.

Examples 2.4 and 2.5 demonstrate situations where the initial condition of the observation error can be bounded independent of L.

The motivation for the choice of the function ln is that for our problem, as will be seen later in the proof of Theorem 1.2, we wish to allow initial conditions of e to be of order L^{n_y}. This requires that we choose a Lyapunov function $U(e)$ and a function $\mu(L)$ satisfying the limit (70) and such that, given a strictly positive real number λ_1, we have

$$|e| \leq \lambda_1 L^{n_y} \quad \Longrightarrow \quad U(e) \leq \mu(L) \ . \tag{74}$$

For instance, if we choose $\mu(L) = \ln(1 + \lambda_2 L^{2(n_y)})$, with λ_2 any strictly positive real number, then the limit (70) is satisfied since we have

$$\lim_{s\to\infty} \frac{s}{\ln(1+\lambda_2 s^{\alpha_1})^{\alpha_2}} = \infty \qquad \forall \lambda_2, \alpha_1, \alpha_2 > 0 \ . \tag{75}$$

Then, with the appropriate choice of λ_2, (74) is satisfied if we choose $U(e) = \ln(1 + e^T P_o e)$. The choice of ln in turn requires the special form of the bounds imposed in (69).

With this remark, we see that if we disregard the issue of ultimate convergence, we recover the result of [11, Thm. 2].

Proof of Lemma 2.4. We follow the lines of the proof of Lemma 2.2. We begin by replacing $\ln(1+e^T P_o e)$ in (71) by $U(e)$. Assume that $W(z,e) \leq c^2 + \mu^2(L) + 1$. From

(24) and the definition of Γ in (68), this implies, for any L, that z is in Γ. Hence from Assumption ULP and the bounds in (69) we can write

(76)

$$\left.\begin{array}{lcl} \dot{V}_{(67)} & \leq & -\Phi_1(z) + \nu_3\gamma(|e|) \\ \dot{U}_{(67)} & \leq & \frac{1}{1+e^T P_o e}\left[-L|e|^2 + 2\lambda_{\max}\{P_o\}|e|\,(\nu_2|e| + \nu_1)\right] \end{array}\right\} \forall (z,e,d) \in \Gamma \times \mathbb{R}^n \times D,$$

where ν_3 is a positive real number which bounds $\frac{\partial V}{\partial z}$ on the set Γ. Such a bound exists because V is C^1 and Γ is compact. Then, from (71), (25), and (76) we can write

$$\begin{aligned} \dot{W}_{(67)} \leq & \tfrac{c(c+1)}{(c+1-V)^2}\left\{-\Phi_1(z) + \nu_3\gamma(|e|)\right\} \\ & + \tfrac{\mu(\mu+1)}{(\mu+1-U)^2}\tfrac{1}{1+e^T P_o e}\left[-L|e|^2 + 2\lambda_{\max}\{P_o\}|e|\,(\nu_2|e| + \nu_1)\right]. \end{aligned} \tag{77}$$

Now fix L_{*_1} so that $\mu^2(L_{*_1}) = c^2 + c + 1$. Such an L_{*_1} exists because $\mu(L)$ is of class-K_∞. Then, using the bounds from (24) and (26), using $c \geq 1$ from Assumption ULP, and choosing $L \geq L_{*_1}$ we have

$$\frac{1}{2} \leq \frac{c(c+1)}{(c+1-V)^2} \leq 2\mu^4. \tag{78}$$

Thus we can rewrite (77) as

$$\begin{aligned} \dot{W}_{(67)} \leq & \tfrac{c(c+1)}{(c+1-V)^2}\Big\{-\Phi_1(z) + \nu_3\gamma(|e|) \\ & + \frac{\mu(\mu+1)}{(\mu+1-U)^2}\frac{1}{1+e^T P_o e}\left[-\frac{L}{2\mu^4}|e|^2 + 4\lambda_{\max}\{P_o\}|e|(\nu_2|e| + \nu_1)\right]\Big\} \end{aligned} \tag{79}$$

Since $(c(c+1))/((c+1-V)^2)$ is positive and bounded away from zero on Γ, it suffices to consider the expression

(80)

$$-\Phi_1(z) + \nu_3\gamma(|e|) + \tfrac{\mu(L)(\mu(L)+1)}{(\mu(L)+1-U(e))^2}\tfrac{1}{1+e^T P_o e}\left[-\tfrac{L}{2\mu(L)^4}|e|^2 + 4\lambda_{\max}\{P_o\}|e|\,(\nu_2|e| + \nu_1)\right].$$

We are interested in evaluating this expression on the set

$$\Lambda_L \doteq \{(z,e) : \vartheta + \rho \leq W(z,e) \leq c^2 + \mu^2(L) + 1\}. \tag{81}$$

We do so by considering the two sets

$$\Lambda_1 \doteq \{(z,e) : V(z) \leq c+1\,,\ 1 < U(e) < \mu(L) + 1\}, \tag{82}$$

$$\begin{aligned} \Lambda_0 \doteq & \{(z,e) : V(z) \leq c+1\,,\ U(e) \leq 1\} \\ & \cap \left\{(z,e) : \vartheta + \rho \leq \frac{cV(z)}{c+1-V(z)} + U(e)\right\} \end{aligned} \tag{83}$$

and by observing that Λ_L is contained in $\Lambda_1 \cup \Lambda_0$, since we have

$$\{U(e) \leq 1\,,\ \vartheta + \rho \leq W(z,e)\} \implies \vartheta + \rho \leq \frac{cV(z)}{c+1-V(z)} + U(e)\,. \tag{84}$$

In the set Λ_1, observe that the limit (70) holds, z is contained in a compact set independent of L, the function $\gamma(|e|)$ is bounded, and $\frac{\mu(\mu+1)}{(\mu+1-U(e))^2}$ is bounded away from zero from (26). We do not use the upper bound on $\frac{\mu(\mu+1)}{(\mu+1-U(e))^2}$ from (26) which depends on L. Finally note that the function $\frac{|e|^2}{1+e^TP_oe}$ is positive and bounded away from 0 on Λ_1. Thus, by examination of expression (80), it follows that there exists a positive real number L_{*_2} such that, for each $L \geq L_{*_2}$, the function $\dot{W}_{(67)}$ can be upper bounded by a function of (z, e) which is negative definite on Λ_1.

In the set Λ_0, to check that $\dot{W}_{(67)}$ is negative for all $(z, e, d) \in \Lambda_0 \times D$, we apply Lemma 2.1 to the expression (80). We remark that, for $L \geq L_{*_1}$, we have

$$(85) \qquad \min_{0\leq U\leq 1}\left\{\frac{\mu(\mu+1)}{(\mu+1-U)^2}\right\} = \frac{\mu}{\mu+1}, \qquad \max_{0\leq U\leq 1}\left\{\frac{\mu(\mu+1)}{(\mu+1-U)^2}\right\} \leq 2.$$

It follows that to know the sign of the expression (80), we can look at (9) by taking

$$(86) \qquad \left\{\begin{array}{rcl} x &=& e\,, \\ K &=& L\,, \\ \chi(z) &=& \frac{1}{2}\Phi_1(z)\,, \\ \psi(e) &=& \frac{1}{2}\frac{e^Te}{1+e^TP_oe}\,, \\ \varphi(z,e) &=& \nu_3\gamma(|e|) + 8\lambda_{\max}\{P_o\}|e|\,(\nu_2|e|+\nu_1) \end{array}\right.$$

and $\kappa(\cdot)$ any class-K_∞ function satisfying

$$(87) \qquad \kappa(L) \leq \frac{L}{2\mu^3(L)(\mu(L)+1)}\,.$$

Such a function exists because $L/(2\mu^3(L)(\mu(L)+1)) > 0$ for $L > 0$ and (70) holds. The set S in Lemma 2.1 is given by Λ_0. It is independent of L and compact. The respective projections satisfy

$$(88) \qquad S_z \subset \{z : V(z) \leq c+1\} = \Gamma, \qquad S_e \subset \{e : U(e) \leq 1\}\,.$$

Then, from (86) and the properties of Φ_1, $\chi(z)$ is continuous on S_z and $\psi(e)$ is continuous on S_e. Clearly, $\psi(e)$ is positive definite on $S_e\backslash\{0\}$. Also, from the continuity of γ and the fact that $\gamma(0) = 0$, $\varphi(z, e)$ is continuous on S and

$$(89) \qquad \varphi(z,e) = 0 \qquad \forall (z,e) \in \{(z,e) : e = 0\} \cap S.$$

To see that $\chi(z)$ is positive definite on the projection of the set $\{(z, e) : e = 0\} \cap S$, we have, with $0 \leq \vartheta \leq 1$ and $\rho > 0$,

$$(90) \qquad \left\{e = 0\,,\ \vartheta + \rho \leq \frac{cV}{c+1-V} + U(e)\right\} \quad \Longrightarrow \quad \vartheta < V(z).$$

So from Assumption ULP, $\chi(z)$ is positive definite on the projection of the set $\{(z, e) : e = 0\} \cap S$. It follows that there exists a positive real number L_{*_3} such that, for each $L \geq L_{*_3}$, the expression (80) can be upper bounded on S by the function

$$-\tfrac{1}{2}\Phi_1(z) - \tfrac{1}{2}\kappa(L)\frac{e^Te}{1+e^TP_oe},$$

which is negative definite on Λ_0 since φ is positive. We then take $L_* = \max\{L_{*_1}, L_{*_2}, L_{*_3}\}$. □

Example 2.4 (mechanical systems). We consider the multi-input nonlinear system

$$\left\{\begin{array}{lcl} \dot{q} & = & r \\ \dot{r} & = & f(q,r) + g(q,r)u, \end{array}\right. \tag{91}$$

where $q \in \mathbb{R}^n$, $r \in \mathbb{R}^n$, $u \in \mathbb{R}^m$ is the input and f and g are C^1. This system could represent a robot model, for example. We assume the existence of a (dynamic) compensator

$$\left\{\begin{array}{lcl} \dot{v} & = & C(q,r,u), \\ u & = & \bar{u}(q,r,v), \end{array}\right. \tag{92}$$

with $v \in \mathbb{R}^l$ such that the closed-loop system

$$\left\{\begin{array}{lcl} \dot{q} & = & r, \\ \dot{r} & = & f(q,r) + g(q,r)\bar{u}(q,r,v), \\ \dot{v} & = & C(q,r,\bar{u}(q,r,v)), \end{array}\right. \tag{93}$$

which we rewrite, with $z = (q^T, r^T, v^T)^T$, as

$$\dot{z} = h(z,0) \ , \tag{94}$$

satisfies Assumption ULP for some neighborhood $\mathcal{U}_1$ and some function V, proper on $\mathcal{U}_1$, with $\vartheta = 0$ and c arbitrarily large. Assumption ULP is satisfied if, for example, the equilibrium $(q,r,v) = (0,0,0)$ is made (locally) asymptotically stable by the compensator (92). To implement the compensator (92) without measurement of r we build the observer

$$\left\{\begin{array}{lcl} \dot{\hat{q}} & = & \hat{r} + L\ell_1(q - \hat{q}), \\ \dot{\hat{r}} & = & L^2\ell_2(q - \hat{q}), \end{array}\right. \tag{95}$$

where L is an adjustable parameter and ℓ_1, ℓ_2 are coefficients of a Hurwitz polynomial. We implement the compensator

$$\left\{\begin{array}{lcl} \dot{v} & = & C(q,\Delta(\hat{r}),v,u), \\ u & = & \bar{u}(q,\Delta(\hat{r}),v), \end{array}\right. \tag{96}$$

where

$$\Delta(\hat{r}) = \hat{r}\min\left\{1, \frac{r_{\max}}{|\hat{r}|}\right\} \tag{97}$$

and $r_{\max}$ is the maximum value of $|r|$ on the set $\Gamma = \{(q,r,v) = z : V(z) \leq c+1\}$, where $V(z)$ and c come from Assumption ULP. This idea for the modification of the compensator is based on the idea in [11]. We choose to saturate the state $\hat{r}$ rather than the entire control u and compensator C because the state r has physical significance and thus determining $r_{\max}$ in the region of interest should be quite natural. Compare with equations (116) and (139) in the proof of Theorem 1.2. If we define the error state

$$e_q \doteq L(q - \hat{q}) \ , \quad e_r \doteq r - \hat{r}, \tag{98}$$

we have the error dynamics

$$\left\{\begin{array}{lcl} \dot{e}_q & = & Le_r - L\ell_1 e_q, \\ \dot{e}_r & = & -L\ell_2 e_q + f(q,r) + g(q,r)\bar{u}(q,\Delta(r - e_r),v) \end{array}\right. \tag{99}$$

and we can apply Lemma 2.4. The bounds in (69) can be readily checked and follow from the introduction of Δ in the compensator (96). Consequently, by choosing c large enough, the modified compensator (96) together with the observer (95) can be used to yield bounded trajectories from the compact set of initial conditions $\mathcal{K}_l \times \mathcal{K}_{(\hat{q},\hat{r})\,l} \subset \mathbb{R}^{2n+l} \times \mathbb{R}^{2n}$, where $\mathcal{K}_l$ is any compact subset of $\mathcal{U}_1$.

As pointed out in Remark 2.3, the bounds in (69) are required because the initial conditions of e grow with L. Specifically, $e_q = L(q-\hat{q})$. However, observe that it may be reasonable to initialize the value of $\hat{q}$ such that $\hat{q}(0) = q(0)$ since q is measured. In this case, the initial condition of e is $(e_q(0) = 0, e_r(0) = r(0) - \hat{r}(0))$ which is independent of L. As mentioned in Remark 2.3, in this case the bounds in (69), and hence the function Δ in (96), are not needed. Nevertheless, if this initialization cannot be done exactly, then the function Δ should be retained.

It would also be possible to build a reduced-order observer for this system. Consider the state $s = r - Lq$. We have

$$\dot{s} = f(q,r) + g(q,r)u - Lr \tag{100}$$

$$= f(q,r) + g(q,r)u - Ls - L^2 q. \tag{101}$$

If we build the observer

$$\left\{ \begin{array}{rcl} \dot{\hat{s}} & = & -L\hat{s} - L^2 q, \\ \hat{r} & = & \hat{s} + Lq, \end{array} \right. \tag{102}$$

then for the error $e_r = r - \hat{r}$, we have

$$\dot{e}_r = f(q,r) + g(q,r)u - Le_r. \tag{103}$$

If we don't specify the initial value of $\hat{s}$, then we choose the modified compensator in (96). If $\hat{s}(0)$ is chosen so that $\hat{s}(0) = -Lq(0)$ then $e_r(0) = r(0)$ and the function Δ is not needed. Let us also remark that the linear operator $q \mapsto \hat{r}$ defined by (102) is output strictly passive. This important property has been exploited in [5].

In all cases, if the original compensator (92) is locally exponentially stabilizing then the conditions of Lemma 4.1 will be satisfied and asymptotic stability is also achieved.

As mentioned earlier, the ideas presented here have been investigated further in [1] and [2].

Example 2.5 (the ball and beam). This example summarizes the result of [37]. Consider the ball-and-beam system

$$\left\{ \begin{array}{rcl} \dot{x}_1 & = & x_2, \\ \dot{x}_2 & = & -G\sin(x_3) + x_1 x_4^2, \\ \dot{x}_3 & = & x_4, \\ \dot{x}_4 & = & \frac{1}{Mx_1^2+J}[\tau - 2Mx_1x_2x_4 - MGx_1\cos(x_3)], \end{array} \right. \tag{104}$$

with three strictly positive real numbers G, M, and J, four state variables x_1 to x_4, and one control τ. See [14] for an interpretation of the state variables and a derivation of the dynamics. We wish to stabilize the system using measurement of x_1 and x_3 only. It can be shown that there exists a semiglobally stabilizing, and locally exponentially stabilizing, control $\bar{u}(x_1, x_2, x_3, x_4)$. See [39] for the case when M, J are known. For the case where M, J are unknown but have known bounds, the procedure is to use the results of [39] to get a result for the (x_1, x_2, x_3) subsystem and then apply Lemma 2.2

and Lemma 4.1 to get a result for the full system. See [37] for a complete discussion. From the results of [22, Thm. 7], Assumption ULP is satisfied with $\vartheta = 0$ for the closed-loop system with $\bar{u}(x_1, x_2, x_3, x_4)$ as the control. To implement this control, we build the observer

$$\left\{\begin{array}{lcl} \dot{\hat{x}}_1 & = & \hat{x}_2 + L\ell_1(x_1 - \hat{x}_1), \\ \dot{\hat{x}}_2 & = & -G\sin(x_3) + L^2\ell_2(x_1 - \hat{x}_1), \\ \dot{\hat{x}}_3 & = & \hat{x}_4 + L\ell_1(x_3 - \hat{x}_3), \\ \dot{\hat{x}}_4 & = & L^2\ell_2(x_3 - \hat{x}_3), \end{array}\right. \tag{105}$$

and we let

$$\tau = \bar{u}(x_1, \hat{x}_2, x_3, \Delta(\hat{x}_4)) \tag{106}$$

where Δ is defined as in Example 2.4. Note that Δ does not need to act on $\hat{x}_2$ because, coincidentally, $\bar{u}$ can be chosen so that the x_2 dependence is already bounded. Again we choose to saturate the state $\hat{x}_4$ instead of the entire control τ because the state x_4 has physical significance. If we define the observer error

$$e_1 = L(x_1 - \hat{x}_1)\ , \quad e_2 = x_2 - \hat{x}_2\ , \quad e_3 = L(x_3 - \hat{x}_3)\ , \quad e_4 = x_4 - \hat{x}_4\ , \tag{107}$$

we have the error dynamics

$$\left\{\begin{array}{lcl} \dot{e}_1 & = & Le_2 - L\ell_1 e_1, \\ \dot{e}_2 & = & -L\ell_2 e_1 + x_1 x_4^2, \\ \dot{e}_3 & = & Le_4 - L\ell_1 e_3, \\ \dot{e}_4 & = & -L\ell_2 e_3 + \frac{1}{Mx_1^2+J}[\tau - 2Mx_1x_2x_4 - MGx_1\cos(x_3)]. \end{array}\right. \tag{108}$$

The bounds in (69) are satisfied and we can achieve bounded solutions from any compact set of initial conditions $(x, \hat{x})$. Furthermore, since $\bar{u}$ is locally exponentially stabilizing, asymptotic stability is also achieved.

Note that, as for the system in Example 2.4, we could choose the initial conditions of $\hat{x}_1$ and $\hat{x}_3$ so that $e_1(0) = 0$ and $e_3(0) = 0$. This is possible because x_1 and x_3 are measured. This, in turn, would remove the need for introducing the function Δ in the control $\bar{u}$. Building a reduced-order observer is also possible.

3. A generalized version of Theorem 1.2.

3.1. Assumptions and results. The proof of Theorem 1.2 follows from an appropriate application of Lemmas 2.3 and 2.4. An even more general case can be considered. Indeed, let the control system be[5]

$$\left\{\begin{array}{lcl} \dot{z} & = & A(z, u, d(t)), \\ y & = & C(z, d(t)). \end{array}\right. \tag{109}$$

We assume only that the point $z = 0$ is semiglobally practically stabilizable by UCO static state feedback, as in the following assumption.

Assumption S-GPS. There exist two integers N_y and N_u so that, for each pair of compact sets $(\mathcal{K}_{zs}, \mathcal{K}_{zl})$, neighborhoods of 0 and with $\mathcal{K}_{zs} \subset \mathcal{K}_{zl}$, we can find

[5] May be augmented with the dynamics of a controller in the case of a dynamically stabilizable system.

1. a positive C^1 function V, zero at 0, which is defined on $\mho$, an open set containing $\mathcal{K}_{zl}$, and such that there exist three positive real numbers ϑ_l, c_s, and c_l satisfying

$$c_s < c_l, \qquad \{z : V(z) \leq \vartheta_l\} \subset \mathcal{K}_{zs}, \qquad \mathcal{K}_{zl} \subset \{z : V(z) \leq c_s\} \tag{110}$$

and so that the set $\{z : V(z) \leq c_l\}$ is compact and contained in $\mho$.

2. a C^2 function $\bar{u}(z)$ which is zero at 0, is defined on $\mho$, and is UCO (i.e. (115) holds) with $n_y \leq N_y$, $n_u \leq N_u$, such that, for the system

$$\dot{z} = A(z, \bar{u}(z), d(t)), \tag{111}$$

we have

$$\dot{V}_{(111)} \leq -\Phi(z), \tag{112}$$

where $\Phi(z)$ is continuous on $\mho$ and positive definite on $\{z : \vartheta_s \leq V(z) \leq c_l\}$ for some real number ϑ_s satisfying

$$0 < \vartheta_s < \vartheta_l. \tag{113}$$

The meaning of this assumption, as we shall make precise later, is that, once a pair of compact sets $(\mathcal{K}_{zs}, \mathcal{K}_{zl})$ is chosen, we know the existence of a UCO control law $\bar{u}$ so that Assumption ULP holds for the system (14). We shall prove the following proposition.

PROPOSITION 3.1. *If Assumption S-GPS holds then the point $z = 0$ of the system (109) is semiglobally practically stabilizable by dynamic output feedback.*

Proof of Theorem 1.2. If the equilibrium $z = 0$ of the system (2) is globally stabilizable by a C^2 state feedback $\bar{u}(z)$, i.e., $z = 0$ is a globally asymptotically stable equilibrium of

$$\dot{z} = A(z, \bar{u}(z)), \tag{114}$$

then, according to the converse Lyapunov theorem [22, Thm. 7], there exists a C^1 function V defined on R^n which is positive definite on $\mathbb{R}^n \backslash \{0\}$ and proper on $\mathbb{R}^n$ so that $\dot{V}_{(114)}$ is negative definite on $\mathbb{R}^n \backslash \{0\}$. It follows that point 1, (112), and (113) in Assumption S-GPS hold for any pair of compact sets $(\mathcal{K}_{zs}, \mathcal{K}_{zl})$. Therefore, if $\bar{u}(z)$ is also UCO, Assumption S-GPS holds. Thus Theorem 1.2 follows from Proposition 3.1. □

3.2. Proof of Proposition 3.1.

Our idea for proving Proposition 3.1 is, instead of using $\bar{u}(z)$ which cannot be "measured", to use an approximation $\widehat{\bar{u}}$. To get this approximation, we use the fact that $\bar{u}$ is UCO, i.e.,[6]

$$\bar{u}(z) = \Psi\left(y, y^{(1)}, \ldots, y^{(n_y)}, u, u^{(1)}, \ldots, u^{(n_u)}\right) . \tag{115}$$

Following [43], the control u and its n_u derivatives can be obtained if we augment the dynamics of the controller. But for y and its n_y derivatives, we shall need an observer. Our proof is made in three steps. The first two steps—dirty derivatives of y and dynamic extension—concern the dynamic output feedback design. In the third step, we shall establish practical stability.

[6] Note the strong requirement that Φ is independent of d.

For the first two steps of this proof, the compact sets $\mathcal{K}_{zs}$ and $\mathcal{K}_{zl}$ are arbitrary but fixed. So from Assumption S-GPS, $\bar{u}$, V, and $\mho$ are given. Then, the following real number is well defined:

$$\bar{u}_{\max} = \max_{\{z : V(z) \leq c_l\}} \{|\bar{u}(z)|\} \ . \tag{116}$$

And, by picking ϑ_1 as an arbitrary real number in $(0, 1/8)$, let κ be a C^1 class-K_∞ function satisfying

$$\kappa(\vartheta_s) = \vartheta_1 \ , \quad \kappa(\vartheta_l) \geq 8\vartheta_1 \ , \quad \kappa(c_s) \geq 1 \ , \quad \kappa(c_l) > 1 + \kappa(c_s). \tag{117}$$

This function exists since our assumption gives

$$0 < \vartheta_s < \vartheta_l \leq c_s < c_l. \tag{118}$$

Then we let

$$V_1(z) = \kappa(V(z)), \qquad c_1 = \kappa(c_s). \tag{119}$$

So Assumption ULP is satisfied and we have

$$\{z : V_1(z) \leq 8\vartheta_1\} \subset \mathcal{K}_{zs}, \qquad \mathcal{K}_{zl} \subset \{z : V_1(z) \leq c_1\} \, . \tag{120}$$

Let us also pick ρ as

$$\rho = \frac{\vartheta_1}{2}. \tag{121}$$

3.2.1. Dirty derivatives of the measurement. With n_y the number of derivatives of y needed in (115) to reconstruct $\bar{u}$, we see, by induction on the order of derivation (see also the notation section), that there exist $n_y + 1$ smooth functions C_i and an integer $m_u \leq n_y$ such that, for each solution of

$$\left\{ \begin{array}{lcl} \dot{z} & = & A(z, u_0, d(t)), \\ \dot{u}_0 & = & u_1, \\ & \vdots & \\ \dot{u}_{m_u - 1} & = & u_{m_u}, \end{array} \right. \tag{122}$$

we have, for all t where the solution makes sense,

$$y^{(i)} = C_i\left(z(t), u_0(t), \ldots, u_{m_u}(t), d(t)\right), \qquad i = 1, \ldots, n_y + 1. \tag{123}$$

Then, for the system

$$\left\{ \begin{array}{lcl} \dot{y} & = & y^{(1)}, \\ & \vdots & \\ \dot{\overbrace{y^{(n_y)}}} & = & C_{n_y+1}\left(z(t), u_0(t), \ldots, u_{m_u}(t), d(t)\right), \end{array} \right. \tag{124}$$

we can propose the following approximate observer:

$$\left\{ \begin{array}{lclcl} \dot{\widehat{y}}_0 & = & \widehat{y}_1 & + & L\ell_0 \left(y - \widehat{y}_0\right), \\ & \vdots & & & \\ \dot{\widehat{y}}_{n_y - 1} & = & \widehat{y}_{n_y} & + & L^{n_y}\ell_{n_y - 1}\left(y - \widehat{y}_0\right), \\ \dot{\widehat{y}}_{n_y} & = & & & L^{n_y+1}\ell_{n_y}\left(y - \widehat{y}_0\right) \ + \ C_{n_y+1}\left(0, u_0, \ldots, u_{m_u}, 0\right), \end{array} \right. \tag{125}$$

with the ℓ_i's chosen as the coefficients of a Hurwitz polynomial associated with a matrix A_o and the real number L to be specified later. It is important to note that (125) is not a true observer since $(y, y^{(1)}, \ldots, y^{(n_y)})$ is not a solution of (125).

3.2.2. Dynamic extension. To reconstruct $\bar{u}$, we need to know n_u derivatives of u. Similarly, to implement the above observer, we need m_u derivatives of u. So, by letting[7]

$$l_u = \max\{n_u, m_u\} + 1 \ , \tag{126}$$

we see that, by adding l_u integrators to the system (109) to be controlled, we shall have u and its required derivatives as measured state components of the system

$$\left\{ \begin{array}{lcl} \dot{z} & = & A(z, u_0, d(t)), \\ \dot{u}_0 & = & u_1, \\ & \vdots & \\ \dot{u}_{l_v - 1} & = & v. \end{array} \right. \tag{127}$$

To design the control v for this augmented system we can use Lemma 2.3. By letting

$$\xi_1 = u_0 - \bar{u}(z) \ , \quad \xi_i = \frac{u_{i-1}}{K^{i-1}} \qquad i = 1, \ldots, l_u \ , \tag{128}$$

with K a positive real number to be specified later, we get the system

$$\left\{ \begin{array}{lcl} \dot{z} & = & A(z, \bar{u}(z) + \xi_1, d(t)) \ \doteq \ h(z, \xi_1, d(t)), \\ \dot{\xi}_1 & = & K\,\xi_2 - \frac{\partial \bar{u}}{\partial z}(z) A(z, \bar{u}(z) + \xi_1, d(t)), \\ \dot{\xi}_2 & = & K\,\xi_3, \\ & \vdots & \\ \dot{\xi}_{l_u} & = & K^{1-l_u}\, v, \end{array} \right. \tag{129}$$

which is in the form of the system (51) written in the ξ_i's coordinates. As we mentioned earlier, Assumption ULP is satisfied by the system

$$\dot{z} = h(z, 0, d(t)) \ . \tag{130}$$

Then we choose coefficients, a_i's, of a Hurwitz polynomial associated with a matrix A_c. Let P_c be the solution of

$$A_c^T P_c + P_c A_c = -I \ . \tag{131}$$

Let $\mathcal{K}_{\xi l}$ be an arbitrary compact set where we choose to initialize ξ. We let

$$\mu_1 = \max\left\{1, \max_{\xi \in \mathcal{K}_{\xi l}} \{\xi^T P_c \xi\}\right\} \ . \tag{132}$$

Then Lemma 2.3 gives the bound K_*, the intermediate control

$$\begin{aligned} v &= -K^{l_u}(a_1 \xi_1 + \ldots + a_{l_u} \xi_{l_u}) \\ &= -K^{l_u}\left(a_1 \left[u_0 - \bar{u}(z)\right] + a_2 \xi_2 + \ldots + a_{l_u} \xi_{l_u}\right) \end{aligned} \tag{133}$$

and the intermediate Lyapunov function

$$W_1(z, \xi) = \frac{c_1 V_1(z)}{c_1 + 1 - V_1(z)} + \frac{\mu_1 \xi^T P_c \xi}{\mu_1 + 1 - \xi^T P_c \xi} \ . \tag{134}$$

[7] In fact the result holds with $l_u = \max\{n_u, m_u\}$ but such a choice leads to more complicated notation.

We have

$$W_1(z,\xi) \le c_1^2 + \mu_1^2 \qquad \forall\,(z,\xi)\, \in\, \mathcal{K}_{zl} \times \mathcal{K}_{\xi l} \tag{135}$$

and, for $K \ge K_*$,

$$\dot{W}_{1\,(127,133)} \le -\Phi_1(z,\xi), \tag{136}$$

where $\Phi_1(z,\xi)$ is positive definite on $\{(z,\xi) : \vartheta_1 + \rho \le W_1(z,\xi) \le c_1^2 + \mu_1^2 + 1\}$.

For future use we define the set

$$\Gamma = \{(z,\xi) : W_1(z,\xi) \le c_1^2 + \mu_1^2 + 1\}\ . \tag{137}$$

This set is compact. (See (24).) We also define the real number $c_2 = c_1^2 + \mu_1^2$. To summarize, by denoting by Z the state vector $(z^T, \xi^T)^T$, we can write the system (127), (133) as

$$\dot{Z} = H_0(Z, d(t)) \tag{138}$$

and we have that Assumption ULP is satisfied for this system with $V = W_1$, $\vartheta = \vartheta_1 + \rho$, and $c = c_2$.

3.2.3. A dynamic output feedback. To summarize our design, we can propose the following dynamic output feedback for the system (109):

(139)

$$\left\{\begin{array}{ll}
\dot{\widehat{y}}_0 & = \widehat{y}_1 + L\ell_0\,(y - \widehat{y}_0)\,, \\
 & \vdots \\
\dot{\widehat{y}}_{n_y-1} & = \widehat{y}_{n_y} + L^{n_y}\ell_{n_y-1}\,(y - \widehat{y}_0)\,, \\
\dot{\widehat{y}}_{n_y} & = \qquad L^{n_y+1}\ell_{n_y}\,(y - \widehat{y}_0) + C_{n_y+1}\,(0, u, K\xi_2, \ldots, K^{m_u}\xi_{m_u+1}, 0)\,, \\
\hline
\dot{u} & = K\,\xi_2, \\
 & \vdots \\
\dot{\xi}_{l_u-1} & = K\,\xi_{l_u}, \\
\dot{\xi}_{l_u} & = K^{1-l_u}\,v, \\
\hline
v & = -K^{l_u}\left(a_1[u - \Delta(\widehat{\overline{u}})] + a_2\xi_2 + \ldots + a_{l_u}\xi_{l_u}\right),
\end{array}\right.$$

where

$$\widehat{\overline{u}} = \Psi\left(y, \widehat{y}_1, \ldots, \widehat{y}_{n_y}, u, K\xi_2, \ldots, K^{n_u}\xi_{n_u+1}\right), \tag{140}$$

and

$$\Delta(s) = s\ \min\left\{1, \frac{\bar{u}_{\max}}{|s|}\right\} \tag{141}$$

with $\bar{u}_{\max}$ given in (116). This function Δ, already encountered in Examples 2.4 and 2.5, is one of the many possible ways of disregarding estimates when they are not in a given compact set and therefore make no sense. More specifically, this function guarantees that the assumption of Lemma 2.4 holds. And, in particular, we have

$$|\Delta(s_1) - \Delta(s_2)| \le \min\{|s_1 - s_2|\,, 2\bar{u}_{\max}\}\ . \tag{142}$$

It would also be possible to saturate each component $\widehat{y}_i$ independently. See Examples 2.4 and 2.5.

For the controller (139), we have chosen the $\widehat{y}_i$s and ξ_is as coordinates for its realization. Indeed, it is for this state that we shall be able to prove the practical stability.

Finally, note that we cannot implement the controller with ξ_1 as one of its state components since $\dot{\xi}_1$ involves unknown quantities.

3.2.4. Practical stability. To study the closed-loop system (109), (139), we use the coordinates $Z = (z^T, \xi^T)^T$ and e where

$$e_i = L^{n_y - i} \left(y^{(i)} - \widehat{y}_i \right) . \tag{143}$$

The closed-loop dynamics can be written

$$\left\{ \begin{array}{rcl} \dot{Z} & = & H(Z, e, d(t)), \\ \dot{e} & = & L A_o \, e + \Xi_e(z, \xi, d(t)), \end{array} \right. \tag{144}$$

where $\Xi_e(z, \xi, d)$ is a vector whose components are zero except the last one which is

$$\begin{aligned} \Xi_e(z, \xi, d)_{n_y} = {} & C_{n_y+1} \left(z, \xi_1 + \bar{u}(z), K\xi_2, \ldots, K^{m_u} \xi_{m_u+1}, d \right) \\ & - C_{n_y+1} \left(0, \xi_1 + \bar{u}(z), K\xi_2, \ldots, K^{m_u} \xi_{m_u+1}, 0 \right) . \end{aligned} \tag{145}$$

This system is in the form (67) considered in Lemma 2.4 with the Z dynamics playing the role of the z dynamics in that lemma.

From the conclusion of the dynamic extension stage and the facts

$$e = 0 \implies \widehat{\bar{u}} = \bar{u}(z), \qquad Z \in \Gamma \implies H(Z, 0, d(t)) = H_0(Z, d(t)), \tag{146}$$

which follow from

$$Z \in \Gamma \implies V(z) \leq c_l, \tag{147}$$

$$\implies \Delta(\bar{u}(z)) = \bar{u}(z), \tag{148}$$

Assumption ULP is satisfied. We also have

$$H(Z, e, d) - H(Z, 0, d) = \left[\begin{array}{c} 0 \\ \vdots \\ 0 \\ K \, a_1 \left[\Delta(\widehat{\bar{u}}) - \Delta(\bar{u}(z)) \right] \end{array} \right] . \tag{149}$$

But, with (123), (128), and the compactness of Γ and D, the following function $\overline{\Psi}$ on $\mathbb{R}^2_{\geq 0}$ is well defined:

$$\begin{aligned} \overline{\Psi}(K, E) = \sup_{(z,e,d,\theta,i,L) \in \mathcal{S}} \Bigg\{ \Bigg| & \tfrac{\partial \Psi}{\partial y^{(i)}} \Big(y, y^{(1)} - \tfrac{\theta}{L^{n_y - 1}} e_1, \ldots, y^{(n_y)} \\ & - \theta e_{n_y}, \xi_1 + \bar{u}(z), K\xi_2, \ldots, K^{n_u} \xi_{n_u+1} \Big) \Bigg| \Bigg\} \end{aligned} \tag{150}$$

where

$$\mathcal{S} = \Gamma \times \{ e : |e| \leq E \} \times D \times [0, 1] \times \{1, \ldots, n_y\} \times [1, \infty) . \tag{151}$$

Then, from (115), (140), and (142), we have, for $L \geq 1$ and all (Z, e, d) in $\Gamma \times \mathbb{R}^{(n_y+1)} \times D$,

$$|H(Z,e,d) - H(Z,0,d)| \leq K\,|a_1|\, \min\left\{|\widehat{u} - \bar{u}(z)|\,,\, 2\bar{u}_{\max}\right\}\ , \tag{152}$$

$$\leq K\,|a_1|\, \min\left\{\overline{\Psi}(K,|e|)|e|\,,\, 2\bar{u}_{\max}\right\}\ , \tag{153}$$

$$\leq \gamma(|e|), \tag{154}$$

for some bounded and continuous function $\gamma : \mathbb{R}_{\geq 0} \to \mathbb{R}_{\geq 0}$ satisfying $\gamma(0) = 0$. Note that γ depends on K which is fixed at this stage. Also $\Xi_e(z, \xi, d)$, given in (145), is bounded on the compact set $\Gamma \times D$ by a positive real number ν_1, independent of L but dependent on K. Now, let P_o satisfy the matrix equation

$$A_o^T P_o + P_o A_o = -I. \tag{155}$$

Also let $\mathcal{K}_{yl}$ be a compact set where we choose to initialize the estimated derivatives of y. Since, from (123), the $y^{(i)}$s are bounded on $\mathcal{K}_{zl} \times \mathcal{K}_{\xi l} \times D$, the following positive real number is well defined and depends on K but not on L:

$$k = \sup_{(z,\xi,(\widehat{y}_i),d) \in \mathcal{K}_{zl} \times \mathcal{K}_{\xi l} \times \mathcal{K}_{yl} \times D} \left\{|y^{(i)} - \widehat{y}_i|\right\}\ . \tag{156}$$

If we then choose

$$\mu_2(L) = \ln(1 + k\lambda_{\max}\{P_o\}L^{2n_y}) \tag{157}$$

we have, for the initial condition $e(0)$,

$$\ln(1 + e^T(0)P_o e(0)) \leq \mu_2(L) \tag{158}$$

and the limit (70) is satisfied. So Lemma 2.4 gives us a bound L_*, depending on K, and the final Lyapunov function

$$W_2(Z,e) = \frac{c_2 W_1(Z)}{c_2 + 1 - W_1(Z)} + \frac{\mu_2(L)\ln(1 + e^T P_o e)}{\mu_2(L) + 1 - \ln(1 + e^T P_o e)} \tag{159}$$

so that for $L \geq L_*$, we have

$$\mu_2(L) \geq 1 \tag{160}$$

and

$$\dot{W}_{2\,(144)} \leq -\Phi_2(Z,e), \tag{161}$$

where $\Phi_2(Z,e)$ is positive definite on $\{(Z,e) : \vartheta_1 + 2\rho \leq W_2 \leq c_2^2 + \mu_2(L)^2 + 1\}$. Since the set $\mathcal{K}_{zl} \times \mathcal{K}_{\xi l} \times \mathcal{K}_{yl}$ is contained in $\{(Z,e) : W_2 \leq c_2^2 + \mu_2(L)^2\}$ and $\rho = \frac{\vartheta_1}{2}$, we conclude that the solutions initialized in $\mathcal{K}_{zl} \times \mathcal{K}_{\xi l} \times \mathcal{K}_{yl}$ remain forever in the set $\{(Z,e) : W_2(Z,e) \leq c_2^2 + \mu_2(L)^2\}$ and are captured by the set $\{(Z,e) : W_2(Z,e) \leq 2\vartheta_1\}$. Then we remark that, c_1, μ_1, μ_2 being larger than 1, we have

$$(W_2(Z,e) \leq 2\vartheta_1) \Longrightarrow \left(e^T P_o e \leq \exp(4\vartheta_1) - 1,\ \xi^T P_c \xi \leq 8\vartheta_1,\ V_1(z) \leq 8\vartheta_1\right). \tag{162}$$

Since the real number ϑ_1 can be chosen arbitrarily small and (120) holds, we have proved the following.

For any pair of compact sets $(\mathcal{K}_{zs}, \mathcal{K}_{zl})$, neighborhoods of 0, with $\mathcal{K}_{zs} \subset \mathcal{K}_{zl}$, we can find compact sets $(\mathcal{K}_{ys}, \mathcal{K}_{\xi s})$ and $(\mathcal{K}_{yl}, \mathcal{K}_{\xi l})$; gains ℓ_is; a_is; a bound K_*; a bounding function $L_*(K)$; integers l_u, n_y; and functions $\bar{u}$, Δ, Ψ so that, for each $K \geq K_*$, $L \geq L_*(K)$, the dynamic output feedback (139) *in closed loop with the system* (109) *makes all the solutions, with initial condition in $\mathcal{K}_{zl} \times \mathcal{K}_{yl} \times \mathcal{K}_{\xi l}$, be captured by the set $\mathcal{K}_{zs} \times \mathcal{K}_{ys} \times \mathcal{K}_{\xi s}$.*

4. The small gain theorem for asymptotic stability. Up to this point we have focused on boundedness of solutions only . However, we have constructed Lyapunov functions to guarantee that, in appropriate coordinates, the states become ultimately arbitrarily small. Now, if the linear approximation in these coordinates is exponentially stable we are effectively done. If the linear approximation is not exponentially stable, then the problem reduces to studying the local stability on the center manifold. See [8]. Because the center manifold analysis can be quite involved, we choose to develop a sufficient condition, other than exponential stability, that can be checked a priori.

Our approach will be to appeal to the notion of "small gain." We will state here a version of the small nonlinear gain theorem, expressed in terms closely related to the nonlinear L_∞-gain from input to state. This is inspired by Sontag's input-to-state (ISS) stability definition [33]. We start with the following definition and give an illustrative fact.

DEFINITION 5. *The system*

$$\dot{x} = h(x, u, t) \tag{163}$$

with $x \in \mathbb{R}^n$, $u \in \mathbb{R}^m$, *and* $t \in \mathbb{R}_{\geq 0}$ *is said to be* uniformly (ϵ, δ) input-to-state stable (*uniformly* (ϵ, δ) *ISS*) *if there exist a class-KL function* β, *a class-K function* γ, *called the gain, and strictly positive real numbers* δ, ϵ *such that, for each* $t_\circ \geq 0$, *for each initial state* $x(t_\circ) = x_\circ$ *satisfying* $|x_\circ| \leq \delta$ *and for each measurable control* $u(\cdot)$ *satisfying* $||u||_{t_\circ} \leq \epsilon$, *the solution of* (163) *exists for each* $t \geq t_\circ$ *and satisfies*

$$|x(t)| \leq \beta(|x_\circ|, t - t_\circ) + \gamma(||u||_{t_\circ}). \tag{164}$$

FACT 4.1 ([35],[20],[45]). *For the system* (163),

1. *if* h *does not depend explicitly on time and the equilibrium point* $x = 0$ *of the system*

$$\dot{x} = h(x, 0) \tag{165}$$

is locally asymptotically stable, then the system is (uniformly) (ϵ, δ) *ISS.*

2. *if* $\frac{\partial h}{\partial x}(x, 0, t)$ *is bounded for sufficiently small* x *uniformly in* t, $h(x, u, t)$ *is locally Lipschitz in* (x, u) *uniformly in* t, *and the equilibrium point* $x = 0$ *of the system*

$$\dot{x} = h(x, 0, t) \tag{166}$$

is uniformly locally asymptotically stable, then the system (163) *is uniformly* (ϵ, δ) *ISS. Moreover, if* $x = 0$ *of* (166) *is locally exponentially stable then* γ *can be taken to be of the form* $\gamma(s) = ks$, *for some positive real number* k, *and* β *can be taken to be of the form* $\beta(s, t) = bse^{-at}$, *for some strictly positive real numbers* b *and* a.

Proof. See the appendix. □

Remark 4.1. For the local exponential stability case, this result was presented in [45]. For the time invariant case, the result is essentially contained in [35, Thm. 2]. For the case where h is differentiable, the proof of this fact can be constructed from theorems in [20, §4.5.2].

The local asymptotic stability of the equilibrium point of interconnected uniformly (ϵ, δ) ISS subsystems can then be analyzed using the following result.

LEMMA 4.1 (small gain). *Consider the feedback interconnection*

$$\left\{\begin{array}{rclrcl}\dot{x}_1 & = & h_1(x_1,u_1,v,t), & u_1 & = & x_2,\\ \dot{x}_2 & = & h_2(x_2,u_2,v,t), & u_2 & = & x_1,\end{array}\right. \tag{167}$$

with $x_i \in \mathbb{R}^{n_i}$ for $i = 1, 2$ and $v \in \mathbb{R}^m$. Define $x = (x_1^T, x_2^T)^T$. Assume h_i is locally Lipschitz in (x_i, u_i, v) and piecewise continuous in t. Assume the ith subsystem is uniformly (ϵ_i, δ_i) ISS with respect to both u_i and v (characterized by δ_i, ϵ_i^u, ϵ_i^v, β_i, γ_i^u and γ_i^v).[8]

Suppose there exist strictly positive real numbers ω and λ such that[9]

$$\left.\begin{array}{r}(1+\lambda)\gamma_1^u \circ (1+\lambda)\gamma_2^u(s) \le s\\ (1+\lambda)\gamma_2^u \circ (1+\lambda)\gamma_1^u(s) \le s\end{array}\right\} \quad \forall s \in [0,\omega] . \tag{168}$$

Under these conditions, the feedback interconnection is uniformly (ϵ, δ) ISS.

More specifically, define

$$\left\{\begin{array}{rcl}\phi_1(s) & = & (1+\lambda^{-1})\left(\beta_1(s,0)+\gamma_1^u\left((1+\lambda^{-1})(1+\lambda^{-1})(\beta_2(s,0))\right)\right),\\ \phi_2(s) & = & (1+\lambda^{-1})\left(\beta_2(s,0)+\gamma_2^u\left((1+\lambda^{-1})(1+\lambda^{-1})(\beta_1(s,0))\right)\right),\\ \phi(s) & = & \phi_1(s)+\phi_2(s)\end{array}\right. \tag{169}$$

and

$$\left\{\begin{array}{rcl}r_1(s) & = & (1+\lambda^{-1})(\gamma_1^v+\gamma_1^u \circ (1+\lambda^{-1})(1+\lambda)(\gamma_2^v))(s),\\ r_2(s) & = & (1+\lambda^{-1})(\gamma_2^v+\gamma_2^u \circ (1+\lambda^{-1})(1+\lambda)(\gamma_1^v))(s),\\ r(s) & = & r_1(s)+r_2(s).\end{array}\right. \tag{170}$$

Then, for any pair (ϵ, δ) satisfying

$$\epsilon \le \min\{\epsilon_1^v, \epsilon_2^v\} \quad , \qquad \phi(\delta)+r(\epsilon) < \min\{\delta_1, \delta_2, \epsilon_1^u, \epsilon_2^u, \omega\} \tag{171}$$

and for each class-K_∞ function α there exists a class-KL function β such that, for each $t_\circ \ge 0$, for each initial state satisfying $|x(t_\circ)| \le \delta$, and for each measurable input $v(\cdot)$ satisfying $||v||_{t_\circ} \le \epsilon$, the solution of (167) *exists for each $t \ge t_\circ$ and satisfies*

$$|x(t)| \le \beta(|x(t_\circ)|, t-t_\circ) + (r+\alpha)(||v||_{t_\circ}). \tag{172}$$

If each subsystem is uniformly globally *ISS and inequality* (168) *holds for all $s \in [0, \infty)$, then inequality* (172) *holds for all initial conditions and all measurable inputs $v(\cdot)$.*

Proof. See the appendix. □

Remark 4.2. 1. Notice that when $||v||_{t_\circ} = 0$, the lemma provides an asymptotic stability result. For the local case, this lemma can be seen as a generalization of [7, Lem. 4.13] where, there, $\gamma_2 \equiv 0$. In the global case, this lemma is a generalization of the result that the cascade of an ISS system and a globally asymptotically stable (GAS) system is GAS.

2. This lemma is a form of the small nonlinear gain theorem (see [10]) which includes explicitly the effects of initial conditions. Its condition (168) was introduced in [24]. For other purely input–output results see [24] and [32] and the references

[8] For example, $|x_1(t)| \le \beta_1(|x_1(t_\circ)|, t-t_\circ) + \gamma_1^u(||u_1||_{t_\circ}) + \gamma_1^v(||v||_{t_\circ})$.

[9] See Fact A.2

therein. In [16], a generalization of this lemma is presented dealing, in particular, with practical stability and the input–output case.

To make this small gain result more efficient we remark that [20, Thm. 4.10], reproduced here, gives us a way to compute effectively the gain function γ.

LEMMA 4.2. *Let B_r^n be the set $\{x \in \mathbb{R}^n : |x| \leq r\}$, $V : \mathbb{R}_{\geq 0} \times B_r^n \to \mathbb{R}$ be a C^1 function, α_1, α_2, α_3, and α_5^{-1} be class-K functions defined on $[0, r]$, and $h : B_r^n \times B_\epsilon^m \times \mathbb{R}_{\geq 0} \to \mathbb{R}^n$ be piecewise continuous in t and locally Lipschitz in (x, u). Assume ϵ satisfies*

$$\epsilon \leq \alpha_5^{-1}(\alpha_2^{-1}(\alpha_1(r))) \tag{173}$$

and, for all $t \geq 0$, for all (x, u) in $B_r^n \times B_\epsilon^m$, we have

$$\alpha_1(|x|) \leq V(t, x) \leq \alpha_2(|x|) \tag{174}$$

and

$$|x| \geq \alpha_5(|u|) \quad \Longrightarrow \quad \frac{\partial V}{\partial t} + \frac{\partial V}{\partial x} h(x, u, t) \leq -\alpha_3(|x|) \ . \tag{175}$$

Under these conditions, the system

$$\dot{x} = h(x, u, t) \tag{176}$$

is uniformly (ϵ, δ) ISS with

$$\delta = \alpha_2^{-1}(\alpha_1(r)) \ , \quad \gamma(s) = \alpha_1^{-1}(\alpha_2(\alpha_5(s))). \tag{177}$$

Furthermore, if $\alpha_i(s) = k_i s^2$, for $i = 1, \ldots, 3$ and some $k_i > 0$ then

$$\beta(s, t) = \sqrt{\frac{k_2}{k_1}}\, s \exp\left(-\frac{k_3}{2k_2}\, t\right) . \tag{178}$$

Example 4.1. Let us consider the system

$$\left\{ \begin{array}{rcl} \dot{z} & = & -z^3 + y, \\ \dot{y} & = & u - z|z|^j, \end{array} \right. \tag{179}$$

where j is some nonnegative real number. We can apply Lemma 2.2 to deduce that the point $(0, 0)$ is semiglobally practically stabilizable by the output feedback

$$u = -Ky \tag{180}$$

with K large enough. To study whether we have asymptotic stability, we check to see if Lemma 4.1 applies.

First we consider the system

$$\dot{x}_1 = -x_1^3 + u_1. \tag{181}$$

To get an expression for γ_1, we apply Lemma 4.2. We have

$$\overset{\cdot}{\widehat{\tfrac{1}{2} x_1^2}}_{(181)} = -x_1^4 + x_1\, u_1, \tag{182}$$

$$\leq -\tfrac{7}{8} x_1^4 - \tfrac{1}{8} |x_1| \left(|x_1|^3 - 8|u_1|\right) . \tag{183}$$

It follows that

$$\gamma_1(s) = 2s^{\frac{1}{3}}, \qquad \epsilon_1 = \delta_1 = +\infty \ . \tag{184}$$

Similarly, for the system

$$\dot{x}_2 = -Kx_2 + u_2|u_2|^j \ , \tag{185}$$

we get

$$\gamma_2(s) = \frac{1}{K-1}|s|^{j+1}, \qquad \epsilon_2 = \delta_2 = +\infty \ . \tag{186}$$

Therefore by choosing K large enough, we can meet the constraint (168) for some λ strictly positive and $\omega = 1$ if

$$j \geq 2 \ . \tag{187}$$

In this condition, we know that the equilibrium $(0,0)$ of (179) and (180) is locally asymptotically stable.

In fact the condition (187) given by Lemma 4.1 is not necessary. Indeed, we have

$$\overset{\cdot}{\overbrace{\frac{|z|^{j+2}}{j+2} + \frac{y^2}{2}}}\Bigg|_{(179)-(180)} = -Ky^2 - |z|^{j+4} \ . \tag{188}$$

This implies global asymptotic stability for all nonnegative j.

Example 4.2 (A continuation of Example 2.1). Consider again the system (34) of Example 2.1. We have seen that the semiglobal stabilizability of the z subsystem, the definite sign of G, as well as the existence of a lower bound for G, are sufficient conditions for the existence of semiglobally practically stabilizing feedback.

We study now whether we have not only practical stability but also asymptotic stability when

$$A(0,0) = 0 \ , \quad \bar{u}(0) = 0 \ , \quad F(0,0,d) = 0 \qquad \forall d \in D \ . \tag{189}$$

For this study and with the notation of Example 2.1, consider the system

$$\dot{x} = f(z,x,d(t)) + g(z,x,d(t))(-K\text{sgn}(g)x) \tag{190}$$

with input z and a disturbance d. We consider the analysis on the set

$$B(\delta) \doteq \{(z,x) : \ \max\{|z|,|x|\} \leq \delta\} \ , \tag{191}$$

where δ is some strictly positive real number. Because of smoothness, compactness of D, and the definition of f, we can write

$$|f| \leq \gamma_f(|z|) + k_1|x| \qquad \forall((z,x),d) \in B(\delta) \times D, \tag{192}$$

where γ_f is any class-K function satisfying

$$\gamma_f(s) \geq \max_{|z| \leq s,\, d \in D} \{|f(z,0,d)|\} \tag{193}$$

and k_1 is some positive real number independent of K. Recall also that $b \leq |g| = |G|$.

We show now that the system (190) is locally asymptotically stable when $z = 0$ and that it has the uniform (ϵ, δ) ISS property with respect to z. Indeed we have, for $((z, x), d) \in B(\delta) \times D$,

$$\frac{1}{2}\overset{\cdot}{\widehat{x^2}}_{(190)} \le -Kbx^2 + |x|[k_1|x| + \gamma_f(|z|)], \tag{194}$$

$$\le -x^2 - |x|\left[(Kb - k_1 - 1)|x| - \gamma_f(|z|)\right]. \tag{195}$$

So, from Lemma 4.2, we have established that, for $K > \frac{k_1+2}{b}$, the system (190) is uniformly (ϵ_x, δ_x) with

$$\left\{ \begin{array}{rcl} \delta_x & = & \delta\ , \\ \gamma_f(\epsilon_x) & < & \delta\ , \\ \gamma_x(s) & = & \frac{1}{Kb-k_1-1}\gamma_f(s)\ , \\ \beta_x(s,t) & = & s\exp(-t)\ . \end{array} \right. \tag{196}$$

On the other hand, we notice with Fact 4.1 that the asymptotic stability assumption for the z subsystem implies the existence of a class-K function γ_z and two strictly positive real numbers δ_z and ϵ_z such that the system (see (35))

$$\dot{z} = A(z, \bar{u}(z) + x) \tag{197}$$

with x as input is (ϵ_z, δ_z) ISS with gain function γ_z and class-KL function β_z.

So let us assume the existence of strictly positive real numbers λ, M, ϖ such that

$$(1 + \lambda)\gamma_z \circ \frac{1}{M}\gamma_f(s) \le s \quad \forall s \in [0, \varpi]\ . \tag{198}$$

Then by imposing the constraint:

$$K \ge \max\left\{\frac{(1 + \lambda)M + 1 + k_1}{b}\ , \ \frac{k_1 + 2}{b}\right\}\ , \tag{199}$$

the conditions of Lemma 4.1 are satisfied with

$$\omega = \min\{\varpi\ , \ (1 + \lambda)\gamma_z(\varpi)\}\ . \tag{200}$$

This result gives that the system

$$\left\{ \begin{array}{rcl} \dot{z} & = & A(z, \bar{u}(z) + x), \\ \dot{x} & = & f(z, x, d(t)) + g(z, x, d(t))(-K\mathrm{sgn}(g)x) \end{array} \right. \tag{201}$$

has a basin of attraction for local asymptotic stability. Precisely, as shown with full details in §5.2.3, there exists a strictly positive number ϑ_0 independent of K, such that the basin of attraction contains the set

$$\mathcal{A} = \{(z, x) : \ |(z, x)| < \vartheta_0\}\ . \tag{202}$$

To complete our proof of semiglobal stabilizability under the condition in (198), it remains to establish that the solutions of the closed-loop system are captured by $\mathcal{A}$. But this follows easily by choosing, in the design, the compact set $\mathcal{K}_s$ so that $\mathcal{K}_s \subset \mathcal{A}$ and picking K large enough.

5. A generalized version of Theorem 1.1.

5.1. Assumptions and results. As was done for Theorem 1.2, we prove here a proposition from which Theorem 1.1 follows directly. We consider again the system (109) under the following assumptions (see (145)).

$$\Xi_e(0,0,d) = 0 \quad A(0,0,d) = 0 \quad \forall d \in D \tag{203}$$

and

Assumption S. We can find

1. a strictly positive real number c_l and a positive C^1 function V which is zero at 0, defined on $\mho$, an open neighborhood of 0, so that the set $\{z : V(z) \leq c_l\}$ is a neighborhood of 0, compact and contained in $\mho$,

2. a C^2 function $\bar{u}(z)$ which is zero at 0, is defined on $\mho$, and is UCO (i.e., (6) holds), such that

$$\dot{V}_{(111)} \leq -\Phi(z) \tag{204}$$

where $\Phi(z)$ is continuous on $\mho$ and positive definite on $\{z : V(z) \leq c_l\} \setminus \{0\}$.

PROPOSITION 5.1. *Suppose the system* (109) *is so that Assumption S and* (203) *hold and there exist strictly positive real numbers* λ, M, ϖ *such that*

$$(1+\lambda)\gamma_z \circ \frac{1}{M}\gamma_{0,(\xi,e)}(s) \leq s \qquad \forall s \; \in [0,\varpi], \tag{205}$$

where $\gamma_{0,(\xi,e)}$ *is a class-*K *function satisfying* (*see* (145))

$$\gamma_{0,(\xi,e)}(s) \geq \max_{(z,d):\,|z|\leq s,\, d\in D} \left\{ \left| \frac{\partial \bar{u}}{\partial z}(z) A(z,\bar{u}(z),d) \right| \,,\, |\Xi_e(z,0,d)| \right\} \tag{206}$$

and γ_z *is the* (ϵ,δ) *gain, with respect to* u *and uniform in* d, *of the system*

$$\dot{z} = A(z,\bar{u}(z)+u,d(t)) \; . \tag{207}$$

Under these conditions, there exists a dynamic output feedback making the origin of the closed-loop system uniformly asymptotically stable with a basin of attraction such that its projection contains any strict compact subset of $\{z : V(z) \leq c_l\}$.

Proof of Theorem 1.1. As already mentioned in the proof of Theorem 1.2, there exists a C^1 function V defined and proper on $\mathbb{R}^n$ and positive definite on $\mathbb{R}^n \setminus \{0\}$ and a C^2 UCO control law so that Assumption S holds for any strictly positive real number c_l. Also, this control being locally exponentially stabilizing, it follows from Fact 4.1 that $\gamma_z(s)$ in (205) is linearly bounded on a neighborhood of 0. On the other hand, with the functions involved in (206) being at least C^1, the function $\gamma_{0,(\xi,e)}(s)$ can be chosen linear on a neighborhood of 0. Inequality (205) follows readily. The conclusion of Theorem 1.1 then follows from Proposition 5.1. □

5.2. Proof of Proposition 5.1.

5.2.1. Practical stabilization. Let us first notice that Assumption S implies the existence of a class-K function α_1 so that

$$V(z) \leq c_l \quad \Longrightarrow \quad \alpha_1(|z|) \leq V(z). \tag{208}$$

Then, let us pick three strictly positive real numbers ϑ_s, ϑ_l, and c_s so that

$$\vartheta_s < \vartheta_l \leq c_s < c_l \tag{209}$$

and define the following compact sets:

$$\mathcal{K}_{zs} = \{z : V(z) \le \vartheta_l\}\,, \qquad \mathcal{K}_{zl} = \{z : V(z) \le c_s\}\,. \tag{210}$$

We are in the condition where the controller design in the proof of Theorem 1.2 applies. So, for any strictly positive real number ϑ_1, and any compact sets $(\mathcal{K}_{yl}, \mathcal{K}_{\xi l})$, we can find, in particular, a real number K_{*1}, a compact set

$$\Gamma \supset \mathcal{K}_{zl} \times \left\{ \xi : \xi^T P_c \xi \le \sup_{\xi \in \mathcal{K}_{\xi l}} \{\xi^T P_c \xi\} \right\}, \tag{211}$$

and positive functions $L_{*1}(K)$, $\mu_2(K)$, so that, for each $K \ge K_{*1}$, $L \ge L_{*1}(K)$, the dynamic output feedback (139) in closed-loop with the system (109) makes all the solutions, with initial condition in $\mathcal{K}_{zl} \times \mathcal{K}_{yl} \times \mathcal{K}_{\xi l}$, remain in the set $\{(z, \widehat{y}, \xi) : (z, \xi) \in \Gamma,\ e^T P_o\, e \le \exp(\mu_2(L) + 1) - 1\}$ and be captured by

$$\mathcal{R} = \left\{(z, \widehat{y}, \xi) \ : \ |z| \le \alpha_1^{-1}(\vartheta_l)\,,\ e^T P_o\, e \le \exp(4\vartheta_1) - 1\,,\ \xi^T P_c \xi \le 8\vartheta_1 \right\}, \tag{212}$$

where P_o is given by (155) and P_c is given by (131).

To study under which condition we have attractivity of a single point, we remark that the closed-loop system is made of the interconnection of

$$\dot{z} = A(z, \bar{u}(z) + \xi_1, d(t)) \tag{213}$$

with:

$$\left\{ \begin{array}{rcl} \dot{\xi} & = & K\, A_c\, \xi + \Xi_\xi(z, \xi, e, d(t)), \\ \dot{e} & = & L\, A_o\, e + \Xi_e(z, \xi, d(t)), \end{array} \right. \tag{214}$$

where Ξ_e is defined in (145) and

$$\Xi_\xi(z, \xi, e, d) = \begin{pmatrix} -\frac{\partial \bar{u}}{\partial z}(z) A(z, \bar{u}(z) + \xi_1, d) \\ 0 \\ \vdots \\ 0 \\ K a_1 \left[\Delta(\widehat{u}) - \bar{u}(z)\right] \end{pmatrix}. \tag{215}$$

From Assumption S, Fact 4.1 applies. So there exist a class-K function γ_z and two strictly positive real numbers δ_z and ϵ_z such that the system

$$\dot{z} = A(z, \bar{u}(z) + u, d(t)) \tag{216}$$

is uniformly (ϵ_z, δ_z) ISS stable with gain function γ_z. It follows then from Lemma 4.1 that local asymptotic stability can be proved if the (ξ, e)-subsystem is also uniformly $(\epsilon_{(\xi,e)}, \delta_{(\xi,e)})$ ISS for some strictly positive real numbers $\epsilon_{(\xi,e)}, \delta_{(\xi,e)}$, and gain $\gamma_{(\xi,e)}$ satisfying a small gain condition like (168).

5.2.2. Input-to-state stability of the $(\boldsymbol{\xi}, \boldsymbol{e})$-subsystem. With (115), (140) and the function $\overline{\Psi}$ defined in (150), we have, for all e, $K \ge 0$, and $L \ge 1$,

$$|\widehat{u} - \bar{u}(z)| \le \overline{\Psi}(K, |e|)\, |e| \qquad \forall ((z, \xi), d) \in \Gamma \times D\,. \tag{217}$$

Also there exists a positive real number ν_4 satisfying, for all $((z,\xi),d)$ in the compact set $\Gamma \times D$,

$$\left|\frac{\partial \bar{u}}{\partial z}(z)A(z,\bar{u}(z)+\xi_1,d) - \frac{\partial \bar{u}}{\partial z}(z)A(z,\bar{u}(z),d)\right| \le \nu_4\,|\xi|. \tag{218}$$

So, with (142), (148), for all $((z,\xi),d)$ in $\Gamma \times D$ and all e, we have

$$|\Xi_\xi(z,\xi,e,d) - \Xi_\xi(z,0,0,d)| \le \nu_4\,|\xi| + \overline{\Psi}(K,|e|)\,K\,|a_1|\,|e|\ . \tag{219}$$

Similarly, from (145), we see that there exists a positive function $\overline{C}$ satisfying, for $K \ge 1$,

$$|\Xi_e(z,\xi,d) - \Xi_e(z,0,d)| \le \overline{C}(K)\,K^{m_u}\,|\xi| \qquad \forall((z,\xi),d) \in \Gamma \times D\ . \tag{220}$$

With (203), let $\gamma_{0,(\xi,e)}$ be any class-K function satisfying

$$\gamma_{0,(\xi,e)}(s) \ge \max_{(z,d):\,|z|\le s,\,d\in D}\left\{\left|\frac{\partial \bar{u}}{\partial z}(z)A(z,\bar{u}(z),d)\right|\ ,\ |\Xi_e(z,0,d)|\right\}. \tag{221}$$

We have, for all $((z,\xi),d)$ in $\Gamma \times D$ and all e,

$$\left\{\begin{array}{lcl}\dot{\overbrace{\xi^T P_c\,\xi}} & \le & -K\xi^T\xi + 2\lambda_{\max}\{P_c\}|\xi|\left[\nu_4\,|\xi| + \gamma_{0,(\xi,e)}(|z|) + \overline{\Psi}(K,|e|)\,K\,|a_1|\,|e|\right],\\ \dot{\overbrace{e^T P_o\,e}} & \le & -Le^Te + 2\lambda_{\max}\{P_o\}|e|\left[\overline{C}(K)\,K^{m_u}\,|\xi| + \gamma_{0,(\xi,e)}(|z|)\right].\end{array}\right. \tag{222}$$

Then, from the properties of V, there exists a strictly positive real number $\epsilon_{(\xi,e)}$ satisfying

$$|z| \le \epsilon_{(\xi,e)} \qquad \Longrightarrow \qquad z \in \mathcal{K}_{zl}. \tag{223}$$

Let also

$$r = \sqrt{\frac{\sup_{\xi\in\mathcal{K}_{\xi l}}\{\xi^T P_c \xi\}}{\lambda_{\max}\{P_c\}}}\,. \tag{224}$$

Then, with (211), we have

$$\left(|z| \le \epsilon_{(\xi,e)}\ ,\ \left|\begin{pmatrix}\xi\\ e\end{pmatrix}\right| \le r\right) \qquad \Longrightarrow \qquad ((z,\xi) \in \Gamma\ ,\ |e| \le r)\,. \tag{225}$$

Finally, define[10]

$$\begin{aligned} K_{*2} &= 2\max\left\{\lambda_{\max}\{P_c\}\,(3+4\nu_4)\,,\right.\\ &\qquad \left. 1 + \tfrac{\max\{\lambda_{\max}\{P_c\},\lambda_{\max}\{P_o\}\}\,(\lambda_{\max}\{P_c\}+\lambda_{\max}\{P_o\})}{r^2\min\{\lambda_{\min}\{P_c\},\lambda_{\min}\{P_o\}\}}\,\gamma_{0,(\xi,e)}\big(\epsilon_{(\xi,e)}\big)^2\right\},\\ L_{*2}(K) &= 2\lambda_{\max}\{P_o\} + K + 2\tfrac{\left(\lambda_{\max}\{P_c\}K|a_1|\overline{\Psi}(K,r)\right)^2+\left(\lambda_{\max}\{P_o\}K^{m_u}\,\overline{C}(K)\right)^2}{\lambda_{\max}\{P_c\}}\,. \end{aligned} \tag{226}$$

[10] The second argument in the max guarantees the condition (173) of Lemma 4.2 holds.

We have established that the condition

$$d \in D \ , \quad |z| \leq \epsilon_{(\xi,e)} \text{ and } \left| \begin{pmatrix} \xi \\ e \end{pmatrix} \right| \leq r$$

implies, for all $K \geq K_{*2}$ and $L \geq L_{*2}(K)$,

$$\overbrace{\xi^T P_c \xi + e^T P_o e}^{\cdot} \leq -\frac{K}{2} \left(\xi^T \xi + e^T e\right) + \left(\lambda_{\max}\{P_c\} + \lambda_{\max}\{P_o\}\right) \gamma_{0,(\xi,e)}(|z|)^2. \tag{227}$$

Then by applying Lemma 4.2, we see that, for $K \geq K_{*2}$ and $L \geq L_{*2}(K)$, the (ξ, e)-subsystem is uniformly $(\epsilon_{(\xi,e)}, \delta_{(\xi,e)})$ ISS with

(228)

$$\left\{ \begin{array}{rcl} \gamma_{(\xi,e)}(s) & = & \sqrt{\frac{2}{K-2}} \sqrt{\frac{\max\{\lambda_{\max}\{P_c\},\lambda_{\max}\{P_o\}\} \left(\lambda_{\max}\{P_c\}+\lambda_{\max}\{P_o\}\right)}{\min\{\lambda_{\min}\{P_c\},\lambda_{\min}\{P_o\}\}}} \gamma_{0,(\xi,e)}(|z|) \ , \\ \delta_{(\xi,e)} & = & \sqrt{\frac{\min\{\lambda_{\min}\{P_o\} \ , \lambda_{\min}\{P_c\}\}}{\max\{\lambda_{\max}\{P_o\}, \ \lambda_{\max}\{P_c\}\}}} \, r, \\ \beta_{(\xi,e)}(s,t) & = & \sqrt{\frac{\max\{\lambda_{\max}\{P_o\} \ , \lambda_{\max}\{P_c\}\}}{\min\{\lambda_{\min}\{P_o\} \ , \lambda_{\min}\{P_c\}\}}} \, s \exp\left(-\frac{1}{\max\{\lambda_{\max}\{P_o\} \ , \lambda_{\max}\{P_c\}\}} t \right) . \end{array} \right.$$

5.2.3. Uniform asymptotic stability.

With Lemma 4.1, we can now conclude that the origin is a uniformly asymptotically stable equilibrium point of the closed-loop system under consideration with domain of attraction containing

$$\mathcal{P} = \left\{ (z, \xi, e) : \phi(|(z, \xi, e)|, K) < \min\{\delta_{(\xi,e)}, \delta_z, \epsilon_{(\xi,e)}, \epsilon_z, \varpi, (1+\lambda)\gamma_z(\varpi)\} \right\} \tag{229}$$

if

1. there exist strictly positive real numbers λ, M, ϖ such that

$$(1+\lambda)\gamma_z \circ \frac{1}{M} \gamma_{0,(\xi,e)}(s) \leq s \qquad \forall s \ \in [0, \varpi] \tag{230}$$

2. K and L are chosen to satisfy

(231)

$$K \geq \max\left\{ K_{*1} \ , \ K_{*2} \ , \ 2 + 2 \frac{\max\{\lambda_{\max}\{P_c\},\lambda_{\max}\{P_o\}\} \left(\lambda_{\max}\{P_c\}+\lambda_{\max}\{P_o\}\right)}{\min\{\lambda_{\min}\{P_c\},\lambda_{\min}\{P_o\}\}} M^2 \left(1+\lambda\right)^2 \right\} ,$$
$$L \geq \max\left\{ L_{*1}(K), L_{*2}(K) \right\} .$$

In (229), the function $\phi(s, K)$ is obtained from (169) as

$$\left\{ \begin{array}{ll} \phi_1(s) & = \left(1+\lambda^{-1}\right) \left(\beta_z(s,0) + \gamma_z\left((1+\lambda^{-1})(1+\lambda^{-1})(\beta_{(\xi,e)}(s,0))\right)\right) , \\ \phi_2(s,K) & = \left(1+\lambda^{-1}\right) \left(\beta_x(s,0) + \gamma_{(\xi,e)}\left((1+\lambda^{-1})(1+\lambda^{-1})(\beta_z(s,0))\right)\right) , \\ \phi(s,K) & = \phi_1(s) + \phi_2(s,K), \end{array} \right. \tag{232}$$

where $\gamma_{(\xi,e)}$, dependent on K but not on L, and $\beta_{(\xi,e)}$ are given by (228). From (228), we see that there exists a class-K function ϱ independent of K and L satisfying

$$\phi(s, K) \leq \varrho(s) \tag{233}$$

for all $s \geq 0$ and K, L satisfying (231). It follows from (229) and (233) that there exists a strictly positive number ϑ_0, independent of K and L, such that the set

$$\mathcal{A} = \left\{ (z, \xi, e) : \ \max\left\{ |z|, \xi^T P_c \xi, e^T P_o e \right\} < \vartheta_0 \right\} \tag{234}$$

is contained in $\mathcal{P}$ and therefore in the domain of attraction for all K and L satisfying (231). Then since, in the controller (139), the gains ℓ_i and a_i and the bound $u_{\max}$ are chosen independent of ϑ_1 and ϑ_l, ϑ_0 does not depend on ϑ_1 and ϑ_l. Therefore, we can choose ϑ_1 and ϑ_l strictly positive and such that

$$\vartheta_1 < \min\left\{\frac{1}{4}\ln(1+\vartheta_0)\,,\ \frac{\vartheta_0}{8}\right\}, \qquad \vartheta_l < \alpha_1(\vartheta_0). \tag{235}$$

With such a choice, we are guaranteed that $\mathcal{A}$ contains $\mathcal{R}$ defined in (212). This implies that the solutions are captured by the set $\mathcal{A}$.

6. Other examples.

6.1. Minimum phase i/o linearizable systems.

Many results in the spirit of Theorems 1.1 and 1.2 can be formulated for minimum phase i/o linearizable systems using the tools developed in this paper. Consider the C^1 system

$$\left\{\begin{array}{rcl} \dot{z} & = & h(z,x,\zeta), \\ \dot{x}_1 & = & x_2, \\ & \vdots & \\ \dot{x}_r & = & \zeta, \\ \dot{\zeta} & = & f(z,x,\zeta,d(t)) + g(z,x,\zeta,d(t))u, \\ y & = & x_1, \end{array}\right. \tag{236}$$

$y \in \mathbb{R}$, $u \in \mathbb{R}$, $z \in \mathbb{R}^m$ and $x = (x_1,\dots,x_r)^T \in \mathbb{R}^r$ and $\zeta \in \mathbb{R}$. We assume a well-defined relative degree and the global minimum phase property as follows.

Assumption HFG. The sign of g is constant and the magnitude of g is bounded away from zero.

Assumption MP. $z = 0$ is a globally asymptotically stable equilibrium for the system

$$\dot{z} = h(z,0,0). \tag{237}$$

We will also assume semiglobal stabilizability for the origin of the (z,x) subsystem:

Assumption RSE. The equilibrium point (0,0) of the (z,x) subsystem, with ζ as input, is semiglobally stabilizable by C^ℓ ($\ell \geq 2$) partial state feedback depending only x. Furthermore, this feedback locally exponentially stabilizes the origin of the x subsystem.

Note that, with Assumption MP, Assumption RSE holds in the following cases:

1. [34], [31] the state z remains bounded for all "disturbances" x and ζ which converge to zero. A special case is when the z subsystem is globally ISS with respect to x and ζ.
2. [36, Thm. 6.2] h is globally Lipschitz.
3. [7], [36] h depends only on z and x_1.
4. [38] h depends on only one component of the vector $(x_1,\dots,x_r,\zeta)^T$.

Then, using Example 2.1 and Proposition 3.1, respectively, we have the following results.

COROLLARY 6.1. *If Assumptions HFG, MP, and RSE hold, then the origin of the system* (236) *is semiglobally practically stabilizable by C^ℓ ($\ell \geq 2$) and UCO state feedback.*

COROLLARY 6.2. *If Assumptions HFG, MP, and RSE hold, then then the origin of* (236) *is semiglobally practically stabilizable by dynamic output feedback.*

The feedback used to prove Corollary 6.1 is of the form $u = -\mathrm{sgn}(g)K(\zeta - \bar{u}(x))$, where $\bar{u}$ is the feedback given by assumption RSE. See Example 2.1. We then remark that the dynamic extension of §3.2.2 is not needed in Corollary 6.2 because the practically stabilizing feedback of Corollary 6.1 is UCO without using u or its derivatives.

Local exponential stability for the origin of the x subsystem is not used in either of these results. It will be used, together with the next assumption, to guarantee asymptotic stabilizability.

Assumption LSG. With γ_z the local gain function of the z subsystem with (x, ζ) as input, there exist a class-K function γ_f and positive real numbers λ, M, and ϖ such that

$$(1+\lambda)\gamma_z \circ \frac{1}{M}\gamma_f(s) \le s \qquad \forall s \in [0, \varpi], \tag{238}$$

$$\gamma_f(s) \ge \sup_{|z|\le s, d\in D} \{|f(z,0,0,d)|\}. \tag{239}$$

We remark that local exponential stability of the origin of the system (237) and $f(0,0,0,d) = 0$ for all $d \in D$ are sufficient to guarantee that Assumption LSG holds.

COROLLARY 6.3. *If Assumptions HFG, MP, RSE, and LSG hold, then the origin of the system* (236) *is semiglobally stabilizable by* C^ℓ ($\ell \ge 2$) *and UCO state feedback.*

Proof. Define $\xi = \zeta - \bar{u}(x)$. With the feedback law mentioned above, the (x, ξ) subsystem has the form

(240)

$$\begin{aligned} \dot{x} &= E(x, \bar{u}(x) + \xi), \\ \dot{\xi} &= f(z, x, \bar{u}(x) + \xi, d(t)) - |g(z, x, \bar{u}(x) + \xi, d(t))|K\xi - \tfrac{\partial \bar{u}}{\partial x}(x)E(x, \bar{u}(x) + \xi), \end{aligned}$$

where

$$|E(x, \bar{u}(x) + \xi) - E(x, \bar{u}(x))| \le |\xi| \ . \tag{241}$$

We will show that, for K sufficiently large, the (x, ξ) subsystem with z as input is uniformly (ϵ, δ) ISS with

$$\beta(s,t) = k_1 s \exp(-k_2 t), \qquad \gamma(s) = \frac{k_3}{\sqrt{K}}\gamma_f(s) \tag{242}$$

for some positive real numbers ϵ, δ, k_1, k_2, k_3, which can all be taken independent of K. The equilibrium point $x = 0$ of

$$\dot{x} = E(x, \bar{u}(x)) \tag{243}$$

is locally exponentially stable. This guarantees the existence of a function $V(x)$ and strictly positive real numbers c_1, c_2, c_3, and r such that, for all $|x| \le r$,

$$\begin{aligned} c_1|x|^2 &\le V(x) \le c_2|x|^2, \\ \dot{V}_{(243)} &\le -|x|^2, \\ |\tfrac{\partial V}{\partial x}| &\le c_3|x| \ . \end{aligned} \tag{244}$$

We restrict our analysis to the set

$$B(\bar{\delta}) \doteq \{(z, x, \xi) : \max\{|z|, |x|, |\xi|\} \le \bar{\delta}\}, \tag{245}$$

where $\bar{\delta} \leq r$ is some strictly positive real number. Because of smoothness and compactness of D we can write

$$\left| f(z,x,\bar{u}(x)+\xi,d) - \frac{\partial \bar{u}}{\partial x}(x) E(x,\bar{u}(x)+\xi) \right| \leq \gamma_f(|z|) + c_4|x| + c_5|\xi| \qquad \forall ((z,x,\xi),d) \in B(\bar{\delta}) \times D, \tag{246}$$

where γ_f is defined in (239) and c_4 and c_5 are some positive real numbers independent of K. With assumption HFG, let $0 < b \leq |g|$. Then, for all $((z,x,\xi),d) \in B(\bar{\delta}) \times D$,

$$\dot{\overbrace{V(x)+\xi^2}}_{(240)} \leq -|x|^2 + c_3|x||\xi| - 2Kb|\xi|^2 + 2|\xi|[\gamma_f(|z|) + c_4|x| + c_5|\xi|], \tag{247}$$

$$\leq -\tfrac{1}{2}|x|^2 - (Kb - 4c_4^2 - c_3^2 - 2c_5)|\xi|^2 + \tfrac{1}{Kb}\gamma_f^2(|z|). \tag{248}$$

For K sufficiently large, the uniform (ϵ, δ) ISS property, with β and γ of the form (242), follows by applying Lemma 4.2. Compare with (227), (228).

From Assumption MP and Fact 4.1, the z subsystem is (ϵ_z, δ_z) ISS with respect to (x, ζ) with gain function γ_z. To identify the gain function with respect to (x, ξ) observe that, $\bar{u}$ being at least C^1, there exists a positive real number k_4 such that

$$|x| \leq \epsilon_z \quad \Longrightarrow \quad \bar{u}(x) \leq k_4|x|. \tag{249}$$

Then, we can take

$$\gamma_z^{(x,\xi)}(s) = \gamma_z((1+k_4)s). \tag{250}$$

Finally, applying Lemma 4.1 to the interconnection of the z and (x, ξ) subsystems, one finds that condition (238) of Assumption LSG is sufficient to guarantee local asymptotic stability for K sufficiently large. Moreover, as in Example 4.2, Lemma 4.1 demonstrates that a neighborhood $\mathcal{A}$ can be described, independent of K, which is contained in the basin of attraction for all K sufficiently large. Then, semiglobal stabilizability follows from Corollary 6.1. □

COROLLARY 6.4. *If Assumptions HFG, MP, RSE, and LSG hold, then the origin of the system* (236) *is semiglobally stabilizable by dynamic output feedback.*

Sketch of proof. The proof is the same as that of the previous corollary. In this instance, the closed-loop system has the state (z, x, ξ, e) and the (x, ξ, e) subsystem is uniformly (ϵ, δ) ISS with β and γ again of the form (242). The conclusion follows from the small gain theorem and Corollary 6.2 with K chosen large enough.

Weaker versions of this last corollary have been published. In [15, §4.7] a similar local result is established for systems with locally exponentially stable zero dynamics. In [19], a similar global result is established for globally Lipschitz nonlinearities. More recently, for the case where the $\dot{z}$ equation in (236) is linear in z, it has been shown in [11] that the equilibrium point $(z, x) = (0, 0)$ is locally stabilizable by output feedback. This result provided estimates for the region of attraction but did not guarantee arbitrarily large domains of attraction. In all of these cases, Assumption LSG is automatically satisfied. When the system does not have zero dynamics it was shown in [18] that the equilibrium point $x = 0$ is semiglobally stabilizable by output feedback. For these results, high gain observers are used. In the special case where only x_1 appears in h, f is generated by differentiation of nonlinearities that depend only on x_1 and z, and $g(x_1)$ is known (see (61)), it has been shown in [42] that the system (236) is semiglobally stabilizable by output feedback, under an assumption similar to Assumption LSG, but without requiring high gain observers. See Example 2.3. If, in addition, the inverse dynamics satisfy an input-to-state stability property with respect to x_1 then, as was shown in [29], the system (236) is *globally* stabilizable by output feedback. This generalized the results of [17] and [26], [27] where it was required that the system be linear up to output injection.

6.2. A nonminimum phase i/o linearizable system. Consider the nonminimum phase system on $\mathbb{R}^3$ with y as the output

$$\left\{\begin{array}{rcl} \dot{z}_1 &=& -z_1 + z_2 - z_1 y^2, \\ \dot{z}_2 &=& z_2^2 + y + z_2 z_1^2, \\ \dot{y} &=& u + z_2. \end{array}\right. \tag{251}$$

The origin of the zero dynamics

$$\left\{\begin{array}{rcl} \dot{z}_1 &=& -z_1 + z_2, \\ \dot{z}_2 &=& z_2^2 + z_2 z_1^2 \end{array}\right. \tag{252}$$

is unstable. Indeed, any solution with initial condition satisfying $z_2(0) > 0$ exhibits finite escape time. Instead of a decomposition into z and y subsystems, we view the system (251) as in Lemma 2.3 with z_1 playing the role of z and (z_2, y) as the block of integrators. Although the assumptions of Lemma 2.3 cannot be satisfied because of the presence of y in the $\dot{z}_1$ equation, the result is still valid. Namely, for K_1 large enough, the control

$$u = -K_1^2 \left(z_2 + \frac{y}{K_1} \right) \tag{253}$$

is semiglobally stabilizing. This can be checked by looking at the time derivative of

$$W = c \frac{z_1^2}{c + 1 - z_1^2} + \mu \frac{\frac{3}{2} z_2^2 + z_2 \frac{y}{K_1} + (\frac{y}{K_1})^2}{\mu + 1 - (\frac{3}{2} z_2^2 + z_2 \frac{y}{K_1} + (\frac{y}{K_1})^2)} . \tag{254}$$

Local exponential convergence follows from the exponential stability of the undriven z_1 subsystem as discussed above. To conclude semiglobal output feedback stabilizability from Propositions 3.1 and 5.1, it remains to verify that the control (253) is UCO. This property holds trivially since we have

$$z_2 = \dot{y} - u. \tag{255}$$

7. Conclusion. We have developed tools for semiglobal stabilization by partial state and output feedback with, as a main application, semiglobal output feedback stabilization for nonlinear systems that admit a uniformly completely observable stabilizing function. Our approach for this problem uses the observer idea of [11] and the dynamic extension of [43]. This result can be seen as an extension of the result given in [40].

An important feature in our approach is to consider the issue of bounded solutions separate from convergence to the equilibrium. To guarantee convergence we have imposed a sufficient, but not necessary, small gain condition which generalizes local exponential stability assumptions.

We have given several applications illustrating our tools:

- We have shown that semiglobal stabilizability by uniformly completely observable state feedback is a sufficient condition for semiglobal practical stabilization by output feedback. Stabilization itself is obtained if an extra local small gain property is satisfied. We have applied this result to input–output linearizable systems.
- We have given output feedback solutions for certain robotics problems (Example 2.4), for the ball and beam (Example 2.5), and for a nonminimum phase system (§6.2).
- We have applied our semiglobal stabilization design to the almost disturbance decoupling problem to eliminate the vanishing regions of the attraction problem discussed in [21] and [25]. See Example 2.2.

Appendix A. Appendix.

A.1. Proof of Fact 4.1. For a strictly positive real number r and a positive integer n, define the set B_r^n, a closed subset of $\mathbb{R}^n$, by

$$B_r^n = \{x \in \mathbb{R}^n : |x| \leq r\} \ . \tag{256}$$

Claim. *Under the conditions of Fact* 4.1, *there exist a strictly positive real number* r, *a* C^1 *function* $V : \mathbb{R}_{\geq 0} \times B_r^n \to \mathbb{R}_{\geq 0}$, *class-$K$ functions* α_1, α_2, α_3, *and* α_5^{-1} *defined on* $[0, r]$ *such that, for all* $t \geq 0$ *and all* (x, u) *in* $B_r^n \times B^m_{\alpha_5^{-1}(r)}$, *we have*

$$\alpha_1(|x|) \leq V(t,x) \leq \alpha_2(|x|) \tag{257}$$

and

$$|x| \geq \alpha_5(|u|) \qquad \Longrightarrow \qquad \frac{\partial V}{\partial t} + \frac{\partial V}{\partial x} h(x,u,t) \leq -\alpha_3(|x|) \ . \tag{258}$$

In the time-invariant case, V can be taken independent of t.

Proof. We start this proof by defining a function $d_\circ : [0, r] \to \mathbb{R}_{\geq 0}$ as follows:

1. For the time-invariant case, as in [35, Eqs. (13), (14)] but assuming only LAS of $x = 0$ for $\dot{x} = h(x, 0)$, we know that there exist a strictly positive real number r, a C^1 function $V : B_r^n \to \mathbb{R}_{\geq 0}$, and functions α_1, α_2, α_3 of class-K defined on $[0, r]$ such that

$$\alpha_1(|x|) \leq V(x) \leq \ \alpha_2(|x|) \forall x \in B_r^n, \tag{259}$$

$$\frac{\partial V}{\partial x}(x) h(x,0) + \alpha_3(|x|) < \ 0 \qquad \forall x \in B_r^n \backslash \{0\}. \tag{260}$$

Then following [35, Proofs of Lem. 3.1 and 3.2], there exists a piecewise constant function $d_\circ : [0, r] \to \mathbb{R}_{\geq 0}$ such that

$$\left\{ \begin{array}{lcl} d_\circ(0) & = & 0 \ , \\ d_\circ(s) & > & 0 \qquad \forall s \in (0, r] \ , \\ |u| \leq d_\circ(|x|) & \Rightarrow & \frac{\partial V}{\partial x}(x) h(x,0) < -\alpha_3(|x|). \end{array} \right. \tag{261}$$

2. For the time-variant case, from the assumptions on h, there exist strictly positive real numbers r and L, a C^1 function $V : \mathbb{R}_{\geq 0} \times B_r^n \to \mathbb{R}_{\geq 0}$, and class-$K$ functions α_1, α_2, α_3, and α_4 defined on $[0, r]$ such that, for all (x, t) in $B_r^n \times \mathbb{R}_{\geq 0}$,

$$\left\{ \begin{array}{rcl} \alpha_1(|x|) & \leq & V(t,x) \leq \alpha_2(|x|), \\ \frac{\partial V}{\partial t}(t,x) + \frac{\partial V}{\partial x}(t,x) h(x,0,t) & \leq & -2\alpha_3(|x|), \\ |\frac{\partial V}{\partial x}(t,x)| & \leq & \alpha_4(|x|), \end{array} \right. \tag{262}$$

(see [20, Thm. 4.7] for example) and

$$(263) \qquad |h(x,u,t) - h(x,0,t)| \leq L|u| \qquad \forall (x,u,t) \in B_r^n \times B_r^m \times \mathbb{R}_{\geq 0} \ .$$

From (262) and (263) it follows that

$$(264) \qquad \begin{aligned} &\frac{\partial V}{\partial t}(t,x) + \frac{\partial V}{\partial x}(t,x)h(x,u,t) \\ &\leq -2\alpha_3(|x|) + \alpha_4(|x|)L|u| \qquad \forall (x,u,t) \in B_r^n \times B_r^m \times \mathbb{R}_{\geq 0}. \end{aligned}$$

In this case, we let the function $d_\circ$ be

$$(265) \qquad \left\{ \begin{array}{rcl} d_\circ(0) & = & 0, \\ d_\circ(s) & = & \frac{\alpha_3(s)}{L\alpha_4(s)} \quad \forall s \in (0,r]. \end{array} \right.$$

From our definitions of $d_\circ$, we have

$$(266) \qquad \inf_{\sigma \in [\tau, r]} \{\sigma, d_\circ(\sigma)\} > 0 \qquad \forall \tau \in (0,r].$$

So let θ be the function defined on $\mathbb{R}_{\geq 0}$ by

$$(267) \qquad \left\{ \begin{array}{rcl} \theta(0) & = & 0 \ , \\ \theta(s) & = & \displaystyle\frac{1}{s}\int_0^s \left(\frac{\tau}{1+\tau} \inf_{\sigma \in [\tau,r]} \{\sigma, d_\circ(\sigma)\} \right) d\tau \qquad \forall s \in (0,r] \ , \\ \theta(s) & = & \theta(r)\displaystyle\frac{s}{r} \qquad \forall s \in (r, +\infty) \ . \end{array} \right.$$

This function is of class-K and the definitions of $d_\circ$ imply, for all (x,u) in $B_r^n \times B_r^m$,

$$(268) \qquad |u| \leq \theta(|x|) \quad \Longrightarrow \quad \frac{\partial V}{\partial t}(t,x) + \frac{\partial V}{\partial x}(t,x)h(x,u,t) \leq -\alpha_3(|x|) \ .$$

So, the claim follows by defining $\alpha_5 = \theta^{-1}$.

If the equilibrium point is locally exponentially stable, then from the assumptions on h, it is well known (see [20, Thm. 4.5]) that the functions α_i, for $i = 1,2,3$, can be taken to be of the form $\alpha_i(s) = k_i r^2$ for some $k_i > 0$. Furthermore, α_4 can be taken of the form $\alpha_4(s) = k_4 s$ for some $k_4 > 0$. It follows that α_5 has the form $\alpha_5(s) = \frac{Lk_4}{k_3}s$.

Finally, Fact 4.1 is established using Lemma 4.2 and the fact that, when $\alpha_i(s) = k_i s^2$, for $i = 1, \ldots, 3$, $\alpha_5 = \frac{Lk_4}{k_3}s$ implies

$$(269) \qquad \alpha_1^{-1}(\alpha_2(\alpha_5(s))) = \sqrt{\frac{k_2}{k_1}\frac{Lk_4}{k_3}}s. \qquad \square$$

A.2. Proof of Lemma 4.1. We will make use of the following facts (see [24]).

FACT A.1. *Let γ be a function of class-K and let g be a function of class-K_∞. We have*

$$(270) \qquad \gamma(a+b) \leq \gamma \circ (Id + g)(a) + \gamma \circ (Id + g^{-1})(b) \ .$$

Proof. The proof follows from considering the two cases $b \leq g(a)$ and $b \geq g(a)$. $\square$

FACT A.2. *Let γ_1 and γ_2 be functions of class-K and let λ and ω be positive real numbers. Then*

$$\begin{aligned} &(1+\lambda)\gamma_1 \circ (1+\lambda)\gamma_2(s) \le s \qquad \forall s \in [0,\omega] \\ &\qquad \Longrightarrow (1+\lambda)\gamma_2 \circ (1+\lambda)\gamma_1 \le s \qquad \forall s \in [0,(1+\lambda)\gamma_2(\omega)]. \end{aligned} \tag{271}$$

Proof. We prove Fact A.2 by contradiction. Assume there exists an $s' \in [0,(1+\lambda)\gamma_2(\omega)]$ such that

$$(1+\lambda)\gamma_2 \circ (1+\lambda)\gamma_1(s') > s'. \tag{272}$$

Since $(1+\lambda)\gamma_2$ is of class-K, this inequality implies that $((1+\lambda)\gamma_2)^{-1}(s')$ is well defined and that

$$(1+\lambda)\gamma_1(s') > ((1+\lambda)\gamma_2)^{-1}(s') \doteq t. \tag{273}$$

This further implies

$$(1+\lambda)\gamma_1 \circ (1+\lambda)\gamma_2(t) > t, \tag{274}$$

which is a contradiction since $t \in [0,\omega]$. □

We now prove Lemma 4.1.

CLAIM. *For the class-K functions ϕ and r defined in* (169) *and* (170), *for positive real numbers δ and ϵ satisfying* (171), *and for each $t_\circ \ge 0$, we have*

$$\{|x(t_\circ)| \le \delta\ ,\ ||v||_{t_\circ} \le \epsilon\} \quad \Longrightarrow \quad \{|x(t)| \le \phi(|x(t_\circ)|) + r(||v||_{t_\circ})\ \forall t \ge t_\circ\}\ . \tag{275}$$

Proof. Define

$$\bar{\delta} = \min\{\delta_1, \delta_2, \epsilon_1^u, \epsilon_2^u, \omega\}. \tag{276}$$

The positive real numbers on the right-hand side come from the uniform (ϵ_i, δ_i) ISS assumption on each subsystem and condition (168). Notice, from (164) and (169), that for any pair (δ, ϵ) satisfying (171) we have

$$\delta < \bar{\delta}. \tag{277}$$

Then, from the assumptions on the system (167), for any initial condition $x(t_\circ)$ satisfying $|x(t_\circ)| \le \delta$ and any measurable function $v(\cdot)$ satisfying $||v||_{t_\circ} \le \min\{\epsilon_1^v, \epsilon_2^v\}$, one can find a strictly positive real number T, possibly infinite, corresponding to a maximal interval $[t_\circ, t_\circ + T)$ such that there exists a unique solution to the feedback interconnection satisfying $|x_i(t)| < \bar{\delta}$ for all $t \in [t_\circ, t_\circ + T)$. Define

$$||x_i||_{t_\circ}^T = \sup_{t_\circ \le \tau < t_\circ + T} |x_i(\tau)|. \tag{278}$$

For ease of notation, we take

$$\gamma_i = \gamma_i^u, \qquad d_i = \gamma_i^v(||v||_{t_\circ}^T). \tag{279}$$

From the uniform (ϵ_i, δ_i) ISS assumption and causality of the feedback interconnection, for all $t \in [t_\circ, t_\circ + T)$, we have

$$\left\{ \begin{array}{lcl} |x_1(t)| & \leq & \beta_1(|x_1(t_\circ)|, t - t_\circ) + \gamma_1(||x_2||_{t_\circ}^T) + d_1, \\ |x_2(t)| & \leq & \beta_2(|x_2(t_\circ)|, t - t_\circ) + \gamma_2(||x_1||_{t_\circ}^T) + d_2. \end{array} \right. \tag{280}$$

Now, using Fact A.1 and (168), we get

$$||x_1||_{t_\circ}^T \leq \beta_1(|x(t_\circ)|, 0) + \gamma_1\left(\beta_2(|x(t_\circ)|, 0) + \gamma_2(||x_1||_{t_\circ}^T) + d_2\right) + d_1 \tag{281}$$

$$\begin{aligned} &\leq \beta_1(|x(t_\circ)|, 0) + \gamma_1\left((1 + \lambda^{-1})(\beta_2(|x(t_\circ)|, 0) + d_2)\right) \\ &\qquad + \gamma_1\left((1 + \lambda)\gamma_2(||x_1||_{t_\circ}^T)\right) + d_1 \end{aligned} \tag{282}$$

$$\begin{aligned} &\leq \beta_1(|x(t_\circ)|, 0) + \gamma_1\left((1 + \lambda^{-1})(1 + \lambda^{-1})(\beta_2(|x(t_\circ)|, 0))\right) \\ &\qquad + \gamma_1\left((1 + \lambda^{-1})(1 + \lambda)(d_2)\right) + (1 + \lambda)^{-1}||x_1||_{t_\circ}^T + d_1 \ . \end{aligned} \tag{283}$$

From this we conclude:

$$\begin{aligned} ||x_1||_{t_\circ}^T \leq (1 + \lambda^{-1})&\left[\beta_1(|x(t_\circ)|, 0) + \gamma_1((1 + \lambda^{-1})(1 + \lambda^{-1})(\beta_2(|x(t_\circ)|, 0))\right] \\ &+ (1 + \lambda^{-1})\left(d_1 + \gamma_1\left((1 + \lambda^{-1})(1 + \lambda)(d_2)\right)\right). \end{aligned} \tag{284}$$

Using the definition of ϕ_1 in (169), r_1 in (170), and d_i in (279), we get

$$||x_1||_{t_\circ}^T \leq \phi_1(|x(t_\circ)|) + r_1(||v||_{t_\circ}^T). \tag{285}$$

We can repeat the analysis for x_2 obtaining the class-K functions ϕ_2 and r_2 defined in (169) and (170), respectively. Then choose

$$\phi(s) = \phi_1(s) + \phi_2(s), \qquad r(s) = r_1(s) + r_2(s). \tag{286}$$

Then, since δ and ϵ are strictly positive real numbers satisfying (171), by contradiction it is easy to see that if $|x(t_\circ)| \leq \delta$ and $||v||_{t_\circ} \leq \epsilon$, T must be infinite which establishes the claim. Note for the global case, there are no restrictions on $|x(t_\circ)|$ or $||v||_{t_\circ}$.

CLAIM. *Let (ϵ, δ) be an arbitrary pair satisfying* (171). *For each strictly positive real number σ and each $(x(t_\circ), v)$ satisfying*

$$|x(t_\circ)| \leq \delta \ , \quad ||v||_{t_\circ} \leq \epsilon \ , \tag{287}$$

there exists a time T so that the corresponding solution satisfies

$$|x(t)| \leq \sigma + r(||v||_{t_\circ}) \qquad \forall t \geq t_\circ + T. \tag{288}$$

Moreover

1. T depends only on σ and m defined as

$$m = \phi(|x(t_\circ)|) + r(||v||_{t_\circ}). \tag{289}$$

2. T is zero for any σ if m is zero.

3. For any fixed $\sigma > 0$, T increases with m.

4. For any fixed $m > 0$, T decreases with σ.

Proof. Given a pair (ϵ, δ) satisfying (171), let $x(t_\circ)$ and v satisfy

$$|x(t_\circ)| \leq \delta, \qquad ||v||_{t_\circ} \leq \epsilon. \tag{290}$$

Then, given the strictly positive real number σ, we pick $t_1(\sigma, m)$ to satisfy

$$\beta_1(m, t_1) + \gamma_1\left((1+\lambda^{-1})^2 \beta_2(m, t_1)\right) \leq \frac{\frac{1}{2}\sigma}{(1+\lambda^{-1})} \ . \tag{291}$$

It is possible to pick such a t_1 because β_i is of class-KL and γ_1 is of class-K. We will show that T given as

$$T = 2t_1 \mathcal{I}\left(\frac{\ln\left(\frac{m}{\frac{1}{2}\sigma}\right)}{\ln(1+\lambda)}\right), \tag{292}$$

where $\mathcal{I}(s)$ is the smallest nonnegative integer greater than or equal to s, is sufficient to establish (288). From (291), (292) this choice satisfies points 1–4 of the claim.

From the uniform (ϵ_i, δ_i) ISS property and the previous claim, we have

$$|x_2(t)| \leq \beta_2(m, t_1) + \gamma_2(||x_1||_{t_s}) + d_2 \qquad \forall t \geq t_1 + t_s \tag{293}$$

for each $t_s \geq t_\circ$. Using this information, we can establish that

$$|x_1(t)| \leq \beta_1(m, t_1) + \gamma_1\left(\beta_2(m, t_1) + \gamma_2(||x_1||_{t_s}) + d_2\right) + d_1 \qquad \forall t \geq 2t_1 + t_s \tag{294}$$

for each $t_s \geq t_\circ$. Using the choice of t_1 in (291), the definition of d_i in (279), and Fact A.1 we get

$$|x_1(t)| \leq \frac{\frac{1}{2}\sigma}{(1+\lambda^{-1})} + (1+\lambda)^{-1}||x_1||_{t_s} + (1+\lambda^{-1})^{-1} r(||v||_{t_\circ}) \quad \forall t \geq 2t_1 + t_s \tag{295}$$

for each $t_s \geq t_\circ$. From this we get

$$||x_1||_{2t_1+t_s} \leq \frac{\frac{1}{2}\sigma}{(1+\lambda^{-1})} + (1+\lambda)^{-1}||x_1||_{t_s} + (1+\lambda^{-1})^{-1} r(||v||_{t_\circ}) \quad \forall t_s \geq t_\circ. \tag{296}$$

Since, from the previous claim, we have

$$||x_1||_{t_\circ} \leq m \ , \tag{297}$$

it follows by induction that, for any positive integer j,

$$||x_1||_t \leq (1+\lambda)^{-j} m + \frac{1}{2}\sigma + r(||v||_{t_\circ}) \qquad \forall t \geq j2t_1 + t_\circ. \tag{298}$$

Thus, with T defined in (292), we have obtained

$$||x_1||_t \leq \sigma + r(||v||_{t_\circ}) \qquad \forall t \geq T + t_\circ. \tag{299}$$

The analysis can be repeated for $x_2(t)$ to establish the claim. □

With the previous claim established, by following the same lines as in [23, Lem. 2.1.4], we can construct a family of mappings $\{T_m\}_{m>0}$ with

1. for each fixed $m > 0$, $T_m : \mathbb{R}_{>0} \to \mathbb{R}_{>0}$ is continuous, strictly decreasing, and onto;
2. for each fixed $\sigma > 0$, $T_m(\sigma)$ is increasing as m increases, and $\lim_{m\to\infty} T_m(\sigma) = \infty$;

such that, for each pair (ϵ, δ) satisfying (171), we have

$$\{|x(t_\circ)| \leq \delta,\ ||v||_{t_\circ} \leq \epsilon,\ m > 0\} \Longrightarrow \{|x(t)| \leq \sigma + r(||v||_{t_\circ})\ \forall t \geq T_m(\sigma) + t_\circ\}\,, \tag{300}$$

$$\{|x(t_\circ)| = ||v||_{t_\circ} = 0\} \Longrightarrow \{|x(t)| = 0 \quad \forall t \geq t_\circ\}\,, \tag{301}$$

where m is defined in (289).

The discussion now follows the proof of [23, Prop. 2.1.5] closely. For each $m > 0$, denote $\psi_m \doteq T_m^{-1}$. Then for each $m > 0$, $\psi_m : \mathbb{R}_{>0} \to \mathbb{R}_{>0}$ is continuous, strictly decreasing, and onto. And, for each fixed $t > 0$, $\psi_m(t)$ is nondecreasing as m increases. We also write $\psi_m(0) = \infty$ which is consistent with ψ_m being strictly decreasing and onto. Finally, we extend this family with

$$\psi_0(t) = 0 \quad \forall t \in \mathbb{R}_{\geq 0}. \tag{302}$$

To summarize the situation, we have established the following implications, for each pair (ϵ, δ) satisfying (171):

$$\{|x(t_\circ)| \leq \delta\ ,\ ||v||_{t_\circ} \leq \epsilon\} \Longrightarrow \{|x(t)| \leq \psi_m(t - t_\circ) + r(||v||_{t_\circ})\ \forall t \geq t_\circ\}\,, \tag{303}$$

$$\{|x(t_\circ)| \leq \delta\ ,\ ||v||_{t_\circ} \leq \epsilon\} \Longrightarrow \{|x(t)| \leq m\ \forall t \geq t_\circ\}\,, \tag{304}$$

with m given in (289).

Now, as in the proof of [23, Prop. 2.1.5], for any $s \geq 0$, $d \geq 0$, and $t \geq 0$, let

$$\bar{\psi}(s, d, t) \doteq \min\{\psi_{\phi(s)+r(d)}(t), \phi(s)\}, \tag{305}$$

where ϕ may need to be extended to be defined for all $s \geq 0$. Since ϕ and r are increasing, for any fixed d, t, $\bar{\psi}(\cdot, d, t)$ is a nondecreasing function and, for any fixed s, t, $\bar{\psi}(s, \cdot, t)$ is a nondecreasing function. Similarly, since, for any fixed m, $\psi_m(\cdot)$ decreases to 0 as $t \to \infty$, the same holds, for fixed s, d, for $\bar{\psi}(s, d, \cdot)$. Finally, if the pair (ϵ, δ) satisfies (171), we have

$$\begin{aligned} &\{|x(t_\circ)| \leq \delta\ ,\ ||v||_{t_\circ} \leq \epsilon\} \\ &\quad\Longrightarrow \quad \{|x(t)| \leq \bar{\psi}(|x(t_\circ)|, ||v||_{t_\circ}, t - t_\circ) + r(||v||_{t_\circ})\ \forall t \geq t_\circ\}\,. \end{aligned} \tag{306}$$

Now, for any class-K_∞ function α, we can find a class-K_∞ function α_1 such that

$$\phi \circ \alpha_1 \leq \alpha. \tag{307}$$

Then, for each $t \geq 0$, we have

$$\bar{\psi}(s, d, t) \leq \bar{\psi}(\alpha_1(d), d, t) + \bar{\psi}(s, \alpha_1^{-1}(s), t). \tag{308}$$

This follows from considering the two cases, $s \leq \alpha_1(d)$ and $d \leq \alpha_1^{-1}(s)$, and by using the monotonicity properties of $\bar{\psi}$. But from the definition of $\bar{\psi}$, it is clear that

$$\bar{\psi}(\alpha_1(d), d, t) \leq \phi(\alpha_1(d)) \leq \alpha(d). \tag{309}$$

Finally, the term $\bar{\psi}(s, \alpha_1^{-1}(s), t)$ can be bounded by a class-KL function $\beta(s, t)$ as in the proof of [23, Prop. 2.1.5]. Combining these manipulations, we have (172). □

REFERENCES

[1] A. Abichou, *Stabilisation de systemes mecaniques avec bifurcation fourche. Commande nonlineaire d' un robot hydraulique*, Ph. D. thesis, 1993.

[2] K. Aouchiche and B. d'Andrea Novel, *Nonlinear dynamic output feedback for an equilibrist robot*, in IEEE SMC Conf. at Le Touquet, 1993, pp. 39–44.

[3] A. Bacciotti, *Potentially global stabilizability*, IEEE Trans. Automat. Control, AC-31 (1986), pp. 974-976.

[4] ———, *Linear feedback: the local and potentially global stabilization of cascade systems*, in Proc. IFAC Nonlinear Control Systems Design Symposium, Bordeaux, 1992, pp. 21–25.

[5] H. Berghuis and H. Nijmeijer, *Global regulation of robots using only position measurements*, Systems Control Lett., 21 (1993), pp. 289–293.

[6] C. Byrnes and A. Isidori, *New results and examples in nonlinear feedback stabilization*, Systems Control Lett., 12 (1989), pp. 437–442.

[7] ———, *Asymptotic stabilization of minimum phase nonlinear systems*, IEEE Trans. Automat. Control, 36 (1991), pp. 1122–1137.

[8] J. Carr, *Applications of Centre Manifold Theory*, Springer-Verlag, Berlin, New York, 1981.

[9] J.-M. Coron and L. Praly, *Adding an integrator for the stabilization problem*, Systems Control Lett., 17 (1991), pp. 89–104.

[10] C. Desoer and M. Vidyasagar, *Feedback Systems: Input-Output Properties*, Academic Press, New York, 1975.

[11] F. Esfandiari and H. K. Khalil, *Output feedback stabilization of fully linearizable systems*, Internat. J. Control, 56 (1992), pp. 1007–1037.

[12] R. Freeman and P. V. Kokotovic, *Backstepping design of robust controllers for a class of nonlinear systems*, in Proc. IFAC Nonlinear Control Systems Design Symposium, 1992, pp. 307–312.

[13] J. W. Grizzle and P. E. Moraal, *Newton, observers and nonlinear discrete-time control*, in Proc. 29th Conference on Decision and Control, IEEE, Honolulu, Hawaii, 1990, pp. 760–767.

[14] J. Hauser, S. Sastry, and P. Kokotovic, *Nonlinear control via approximate input-output linearization: the ball and beam example*, IEEE Trans. Automat. Control, 37 (1992), pp. 392–398.

[15] A. Isidori, *Nonlinear Control Systems*, Springer-Verlag, Berlin, New York, 1989.

[16] Z. P. Jiang, A. R. Teel, and L. Praly, *Small gain theorem for ISS systems and applications*, Math. Control Signals Systems, 7 (1995), pp. 95–120.

[17] I. Kanellakopoulos, P. V. Kokotovic, and A. S. Morse, *A toolkit for nonlinear feedback design*, Systems Control Lett., 18 (1992), pp. 83–92.

[18] H. K. Khalil and F. Esfandiari, *Semiglobal stabilization of a class of nonlinear systems using output feedback*, IEEE Trans. Automat. Control, 38 (1993), pp. 1412-1415.

[19] H. K. Khalil and A. Saberi, *Adaptive stabilization of a class of nonlinear systems using high-gain feedback*, IEEE Trans. Automat. Control, AC-32 (1987), pp. 1031-1035.

[20] H.K. Khalil, *Nonlinear Systems*, Macmillan Publishing Company, New York, 1992.

[21] P. V. Kokotovic and R. Marino, *On vanishing stability regions in nonlinear systems with high-gain feedback*, IEEE Trans. Automat. Control, AC-31 (1986), pp. 967–970.

[22] J. Kurzweil, *On the inversion of Lyapunov's second theorem on stability of motion*, Amer. Math. Soc. Transl., Ser. 2, 24 (1956), pp. 19–77.

[23] Y. Lin, *Lyapunov function techniques for stabilization*, Ph. D. dissertation, Rutgers University, New Brunswick, NJ, 1992.

[24] I. M. Y. Mareels and D. J. Hill, *Monotone stability of nonlinear feedback systems*, J. Math. Systems Estim. Control, 2 (1992), pp. 275–291.

[25] R. Marino, W. Respondek, and A.J. van der Schaft, *Almost disturbance decoupling for single-input single-output nonlinear systems*, IEEE Trans. Automat. Control, 34 (1989), pp. 1013-1017.

[26] R. Marino and P. Tomei, *Dynamic output feedback linearization and global stabilization*, Systems Control Lett., 17 (1991), pp. 115–121.

[27] ———, *Robust output feedback stabilization of single input output nonlinear systems*, in Proc. 30th Conference on Decision and Control, IEEE, Brighton, 1991, pp. 2503–2508.

[28] F. Mazenc, L. Praly, and W. P. Dayawansa, *Global stabilization by output feedback: examples and counterexamples*, Systems Control Lett., 22 (1994), pp. 119–125.

[29] L. Praly and Z. P. Jiang, *Stabilization by output feedback for systems with ISS inverse dynamics*, Systems Control Lett., 21 (1993), pp. 19–33.

[30] A. Saberi and H. K. Khalil, *Quadratic-type Lyapunov functions for singularly perturbed*

systems, IEEE Trans. Automat. Control, AC-29 (1984), pp. 542–550.

[31] P. SEIBERT AND R. SAUREZ, *Global stabilization of nonlinear cascade systems*, Systems Control Lett., 14 (1990), pp. 347–352.

[32] J. S. SHAMMA, *The necessity of the small-gain theorem for time-varying and nonlinear systems*, IEEE Trans. Automat. Control, 36 (1991), pp. 1138–1147.

[33] E. D. SONTAG, *Smooth stabilization implies coprime factorization*, IEEE Trans. Automat. Control, AC-34 (1989), pp. 435–443.

[34] ———, *Remarks on stabilization and input-to-state stability*, in Proc. 28th Conference on Decision and Control, IEEE, Tampa, FL, 1989, pp. 1376–1378.

[35] ———, *Further facts about input to state stabilization*, IEEE Trans. Automat. Control, 35 (1990), pp. 473–476.

[36] H. J. SUSSMANN AND P. V. KOKOTOVIC, *The peaking phenomenon and the global stabilization of nonlinear systems*, IEEE Trans. Automat. Control, 36 (1991), pp. 424–439.

[37] A. R. TEEL, *Semi-global stabilization of the 'ball and beam' using 'output' feedback*, in Proc. 1993 American Control Conference, American Automatic Control Council, San Francisco, CA, 1993, pp. 2577–2581.

[38] ———, *Semi-global stabilization of minimum phase nonlinear systems in special normal forms*, Systems Control Lett., 19 (1992), pp. 187–192.

[39] ———, *Using saturation to stabilize single-input partially linear composite nonlinear systems*, in Proc. IFAC Nonlinear Control Systems Design Symposium, 1992, pp. 224–229.

[40] A. TEEL AND L. PRALY, *Global stabilizability and observability imply semi-global stabilizability by output feedback*, Systems Control Lett., 22 (1994), pp. 313–325.

[41] ———, *Semi-global stabilization by output feedback: a worked example*, CAS tech. report E137, École des Mines des Paris, Fontainbleau Cédex, France, 1992.

[42] ———, *Semi-global stabilization by linear, dynamic output feedback for siso minimum phase nonlinear systems*, in Proc. 12th IFAC World Congress, Sydney, Australia, vol. 8, 1993, pp. 39–42.

[43] A. TORNAMBÈ, *Output feedback stabilization of a class of non-minimum phase nonlinear systems*, Systems Control Lett., 19 (1992), pp. 193–204.

[44] J. TSINIAS, *Sufficient Lyapunov-like conditions for stabilization*, Math. Control Signals Systems, 2 (1989), pp. 343–357.

[45] M. VIDYASAGAR AND A. VANNELLI, *New relationships between input-output and Lyapunov stability*, IEEE Trans. Automat. Control, 27 (1982), pp. 481–483.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 33, No. 5, pp. 1489–1509, September 1995

© 1995 Society for Industrial and Applied Mathematics

009

CONSISTENCY OF PRIMAL-DUAL APPROXIMATIONS FOR CONVEX OPTIMAL CONTROL PROBLEMS*

STEPHEN E. WRIGHT[†]

Abstract. Problems in the optimal control of linear systems with convex costs are recast in a primal-dual (minimax) framework. An approximation scheme which leads to primal and dual optimal control problems in discrete time having similar structure to the original primal-dual pair is introduced. The discretization is shown to be variationally consistent in the sense of epi/hypo-convergence, so that any limit point of solutions for the approximate minimax problems will solve the original primal and dual problems.

Key words. convex optimal control, duality, Lagrangians, discrete approximation, Euler approximation, epi-convergence, epi/hypo-convergence, variational convergence

AMS subject classifications. 49M15, 49M25

1. Introduction.

In this paper we present a new approach for discretizing convex problems in optimal control based on the *direct* use of minimax formulations. We analyze the method through the concept of epi/hypo-convergence [2]–[4]. There are several advantages to be gained from this viewpoint. First, many nonsmooth problems and most exact penalty representations can be written equivalently as minimax problems with smooth data and simple constraints. Second, epi/hypo-convergence involves convergence of both solutions *and* multipliers, affording better sensitivity analysis at approximate solutions. Finally, the perturbational properties of epi/hypo-convergence make it easier to extend convergence results to a broader class of approximations.

A number of authors have studied methods of discretization for constrained optimal control problems. Some have used direct approximations of the primal problem, including Cullum [9], [10] and Daniel [11]. Others have introduced penalties into the primal formulation; see, for instance, the papers of Chen and Mills [7], Cullum [8], and Russell [27]. Also various dual formulations have been considered by Hager [13], Hager and Ianculescu [15], and Pirronneau and Polak [18]. The use of primal-dual representations for approximation in optimal control as in the current paper is new.

We begin in §2 with the formulation of the optimal control problem and a discussion of Rockafellar's primal-dual representation [22]. In §3 we introduce the basic approximation scheme, given by restricting attention to (primal and dual) controls which are piecewise constant. It is demonstrated that this leads to a dual pair of optimal control problems in discrete time. Section 4 is devoted to an exposition of the relevant facts from the theory of closed saddle functions and epi/hypo-convergence that will be needed later. In §5, we state and prove our main result about the consistency of the approximation scheme introduced earlier. It is assumed at this point that the trajectory and cost corresponding to an approximate control can be calculated precisely at the grid points (as could be done in the autonomous case). The final

*Received by the editors November 18, 1992; accepted for publication (in revised form) April 11, 1994.

[†]Department of Mathematics and Statistics, Miami University, Oxford, Ohio 45056 (wrightse@muohio.edu). This work was initiated at the Department of Mathematics, University of Washington, Seattle, and completed at the IBM Thomas J. Watson Research Center, Yorktown Heights, New York.

section extends the consistency theorem to approximation by Euler's finite-difference scheme.

We will deal with consistency of the approximations, obtaining convergence under the general assumption of continuity on the data. Rates of convergence are not discussed in this paper. Such results typically require the data to have Lipschitz-continuous derivatives (perhaps of several orders). Recent attempts to treat fairly general problems in optimal control, removing extraneous differentiability assumptions, have been undertaken by Dontchev and Hager [12], Mordukhovich [16], and Veliov [28], to name a few. Earlier work was restricted to problems of very simple structure—for example, unconstrained problems—or quadratic problems with simple bounds on endpoints; cf. Bosarge and Johnson [6], Chen and Mills [7], Hager [13], [14], and Mathis and Reddien [17]. All of the above papers treat primal or dual formulations of control problems. At this time, nothing is known about rates for primal-dual approximations. The current paper should provide a natural framework for adapting the concepts of "epi-distance" (cf. [5]) to study this issue.

2. Primal-dual representation of optimal control problems.

The problem we shall consider is the following:

($\mathcal{P}^1$) minimize the functional

$$F(u) = \int_{t_0}^{t_1} [p_t \cdot u_t + \varphi_t(u_t) - c_t \cdot x_t + \psi_t^*(q_t - D_t u_t - C_t x_t)]dt$$
$$+ p_e \cdot u_e + \varphi_e(u_e) - c_e \cdot x_{t_1} + \psi_e^*(q_e - D_e u_e - C_e x_{t_1})$$

over the control space $\mathcal{U}^1 = \mathcal{L}_k^1 \times \mathbf{R}^{k_e}$, with the dynamics given by

$$\dot{x}_t = A_t x_t + B_t u_t + b_t \text{ a.e.}, \quad x_{t_0} = B_e u_e + b_e.$$

We shall also work with the problems ($\mathcal{P}^r$) (for $r \in [1, \infty]$) given by replacing the space $\mathcal{U}^1$ by $\mathcal{U}^r = \mathcal{L}_k^r \times \mathbf{R}^{k_e}$.

The functions φ_t, ψ_t^* (for each $t \in [t_0, t_1]$) and φ_e and ψ_e^* are assumed to be proper, lower semicontinuous convex functions. (A function with values in $\overline{\mathbf{R}} = \mathbf{R} \cup \{\pm\infty\}$ is said to be *proper* if it does not take actually the value $-\infty$ and is not identically $+\infty$.) In addition, we require that φ_t and ψ_t vary epi-continuously with t and that the data elements p_t, q_t, A_t, B_t, b_t, C_t, c_t, D_t are all continuous with respect to $t \in [t_0, t_1]$.

This formulation of an optimal control problem was introduced by Rockafellar [22]. Its main advantage is that it leads to a clean duality theory, exposing those aspects of the problem which are important in the study of optimality conditions and sensitivity. It can also serve as a model in its own right; cf. [21], [22]. Since φ_t, φ_e, ψ_t^*, and ψ_e^* may take values in $\overline{\mathbf{R}}$, constraints on the controls and trajectories (as well as endpoint constraints) can be modelled implicitly. In addition, the expression $\psi_t^*(q_t - D_t u_t - C_t x_t)$ can be viewed as a penalty representation for a constraint of the form

$$q_t \leq D_t u_t - C_t x_t.$$

For $r \in [1, \infty]$, we will also consider the dual problems

($\mathcal{D}^r$) maximize the functional

$$G(v) = \int_{t_0}^{t_1} [q_t \cdot v_t - \psi_t(v_t) - b_t \cdot y_t - \varphi_t^*(D_t^* v_t + B_t^* y_t - p_t)]dt$$
$$+ q_e \cdot v_e - \psi_e(v_e) - b_e \cdot y_{t_0} - \varphi_e^*(D_e^* v_e + B_e y_{t_0} - p_e)$$

over the control space $\mathcal{V}^r = \mathcal{L}_l^r \times \mathbf{R}^{l_e}$ with the dynamics given by

$$-\dot{y}_t = A_t^* y_t + C_t^* v_t + c_t \text{ a.e.}, \qquad y_{t_1} = C_e^* v_e + c_e.$$

Here the superscript $*$ denotes the transpose for a matrix and the *convex conjugate* for a function: $h^*(r) = \sup_s\{r \cdot s - h(s)\}$. (If h is proper, lower semicontinuous, and convex, then $h = (h^*)^*$. See [20].)

Instead of approximating the problems $(\mathcal{P}^r)$ and $(\mathcal{D}^r)$ directly, we will work with a corresponding minimax representation for optimality. To this end, we introduce the functional

$$\mathcal{J}(u,v) = \int_{t_0}^{t_1} J_t(u_t, v_t)dt + J_e(u_e, v_e) - \gamma(u,v),$$

where

$$J_t(u_t, v_t) = \begin{cases} \infty & \text{if } \varphi_t(u_t) = \infty, \\ p_t \cdot u_t + \varphi_t(u_t) - v_t \cdot D_t u_t + q_t \cdot v_t - \psi_t(v_t) & \text{otherwise,} \end{cases}$$

$$J_e(u_e, v_e) = \begin{cases} \infty & \text{if } \varphi_e(u_e) = \infty, \\ p_e \cdot u_e + \varphi_e(u_e) - v_e \cdot D_e u_e + q_e \cdot v_e - \psi_e(v_e) & \text{otherwise,} \end{cases}$$

and

$$\begin{aligned}\gamma(u,v) &= \int_{t_0}^{t_1} y_t \cdot (B_t u_t + b_t)dt + y_{t_0} \cdot (B_e u_e + b_e)dt \\ &= \int_{t_0}^{t_1} x_t \cdot (C_t^* v_t + c_t)dt + x_{t_1} \cdot (C_e^* v_e + c_e).\end{aligned}$$

To avoid ambiguity, $\mathcal{J}$ is defined using the convention that $\infty - \infty = \infty$. This is used twice: first in evaluating the integrals and then in adding the integrals to $J_e(u_e, v_e)$. The expression $\int_{t_0}^{t_1} J_t(u_t, v_t)dt$ is taken to be ∞ if and only if $J_t(u_t, v_t)$ is not majorized by any integrable function; it is taken to be $-\infty$ if and only if $J_t(u_t, v_t)$ is majorized by an integrable function but not minorized by an integrable function. Of course, the term $\gamma(u,v)$ is always finite.

A control pair $(\bar{u}, \bar{v})$ is said to be a *saddle point* for $\mathcal{J}$ on $\mathcal{U}^r \times \mathcal{V}^{r'}$ if, for all $(u,v) \in \mathcal{U}^r \times \mathcal{V}^{r'}$, it is true that

$$\mathcal{J}(\bar{u}, v) \le \mathcal{J}(\bar{u}, \bar{v}) \le \mathcal{J}(u, \bar{v}).$$

A theory of duality for problems $(\mathcal{P}^r)$ and $(\mathcal{D}^{r'})$, linking the primal and dual problems to the quest for saddle points of $\mathcal{J}$, was developed recently by Rockafellar [22]. The next two theorems summarize the facts that are most relevant to our purpose here.

THEOREM 1 [22]. *Consider* $r, r' \in [1, \infty]$.

(i) *The functional* F *is lower semicontinuous and convex on* $\mathcal{U}^r$ *and takes values which are finite or* $+\infty$. *If* ψ_t^* *and* ψ_e^* *are finite-valued everywhere, then* F *is* (inf)-*proper.*

(ii) *The functional* G *is upper semicontinuous and concave on* $\mathcal{V}^{r'}$ *and takes values which are finite or* $-\infty$. *If* φ_t^* *and* φ_e^* *are finite-valued everywhere, then* G *is* (sup)-*proper.*

(iii) *The problems* $(\mathcal{P}^r)$ *and* $(\mathcal{D}^{r'})$ *are the primal and dual problems associated with the problem of finding a saddle point of* $\mathcal{J}$ *on* $\mathcal{U}^r \times \mathcal{V}^{r'}$, *i.e.,*

$$F(u) = \sup_{v \in \mathcal{V}^{r'}} \mathcal{J}(u,v) \text{ and } G(v) = \inf_{u \in \mathcal{U}^r} \mathcal{J}(u,v).$$

In particular, the optimal values satisfy $\inf(\mathcal{P}^r) \geq \sup(\mathcal{D}^{r'})$*, and a pair* $(\bar{u}, \bar{v})$ *is a saddle point for* $\mathcal{J}$ *over* $\mathcal{U}^r \times \mathcal{V}^{r'}$ *if and only if* $\bar{u}$ *solves* $(\mathcal{P}^r)$*,* $\bar{v}$ *solves* $(\mathcal{D}^{r'})$*, and* $\inf(\mathcal{P}^r) = \sup(\mathcal{D}^{r'})$.

THEOREM 2 [22]. *Consider* $r, r' \in [1, \infty]$. *Assume that* $\varphi_t^*, \psi_t^*, \varphi_e^*, \psi_e^*$ *are finite everywhere.*

(i) *The problems* $\mathcal{P}^r$ *and* $\mathcal{D}^{r'}$ *both admit optimal solutions, and*

$$\min(\mathcal{P}^r) = \max(\mathcal{D}^{r'}) \quad (\textit{finite}).$$

(ii) *A pair* $(\bar{u}, \bar{v})$ *is a saddle point of* $\mathcal{J}$ *over* $\mathcal{U}^r \times \mathcal{V}^{r'}$ *if and only if* $\bar{u}$ *solves* $(\mathcal{P}^r)$ *and* $\bar{v}$ *solves* $(\mathcal{D}^{r'})$.

(iii) *Any optimal solution of* $(\mathcal{P}^r)$ *is actually in* $\mathcal{U}^\infty$*, and any optimal solution of* $(\mathcal{D}^{r'})$ *is actually in* $\mathcal{V}^\infty$.

The arguments given by Rockafellar in deriving this duality result (as well as the properness of the functionals F and G) depend heavily on the finiteness assumption made on the conjugate integrands. This assumption is equivalent to requiring the functions $\varphi_t, \psi_t, \varphi_e, \psi_e$ to be coercive. (A function $h : \mathbf{R}^d \mapsto \overline{\mathbf{R}}$ is said to be *coercive* if $\lim_{|w| \to \infty} h(w)/|w| = \infty$. For example, h is coercive if the effective domain of h is bounded. Also, strong convexity implies coercivity.)

Statement (iii) of Theorem 2 tells us that in the search for solutions to problems $(\mathcal{P}^r)$ and $(\mathcal{D}^{r'})$, we need only consider the problem of finding a saddle point of $\mathcal{J}$ over $\mathcal{U}^\infty \times \mathcal{V}^\infty$. Indeed, in applications it is often more natural to restrict attention to essentially bounded controls anyway.

Even so, we still need to work with the larger spaces $\mathcal{U}^r$ and $\mathcal{V}^{r'}$ in our discussion of approximations. The results we give are of the epi/hypo-convergence variety. Thus, any cluster point of a sequence of solutions to the approximate problems will solve the original problem. The difficulty in applying such a result is establishing whether the approximate solutions actually cluster at all. Of course, clustering is more likely in a weaker topology than in a stronger one. Thus the sharpest result is that which guarantees epi/hypo-convergence relative to the weakest topology available. On the other hand, many applications require that the trajectories corresponding to approximate controls converge uniformly (or cluster in the uniform norm) to an optimal trajectory. This requires working with the weak topologies for $\mathcal{L}^r$ on the controls.

3. Approximation by step functions. We now introduce an approximation scheme for finding a saddle point of $\mathcal{J}$ on $\mathcal{U}^r \times \mathcal{V}^{r'}$. The idea is to approximate the controls by feasible *step functions* on $[t_0, t_1]$. Of course, there might not be any feasible step functions, so our hypotheses will eventually require the effective domains $\operatorname{dom} \varphi_t$ and $\operatorname{dom} \psi_t$ to be constant with respect to t. Note, however, that this does not necessarily restrict φ_t and ψ_t from varying with t. More generally, one can obtain similar results when $\operatorname{dom} \varphi_t = W_t \bar{U} + w_t$ and $\operatorname{dom} \psi_t = Z_t \bar{V} + z_t$, where W_t, w_t, Z_t, z_t are continuous with respect to t, and the sets $\bar{U}$ and $\bar{V}$ are fixed convex polyhedrons. It can also be shown that a related procedure can be applied if the graphs of $\operatorname{dom} \varphi_t$ and $\operatorname{dom} \psi_t$ are convex in $\mathbf{R}^k \times \mathbf{R}$ and $\mathbf{R}^l \times \mathbf{R}$.

Let $\pi = (t_0 = a_0 < a_1 < \cdots < a_T = t_1)$ be a partition of the interval $[t_0, t_1]$. Consider the following subsets of $\mathcal{U}^\infty$ and $\mathcal{V}^\infty$:

$$\mathcal{U}_\pi = \{u \in \mathcal{U}^\infty | u_t \text{ is constant a.e. on } [a_{\tau-1}, a_\tau] \text{ for } \tau = 0, \ldots, T\},$$
$$\mathcal{V}_\pi = \{v \in \mathcal{V}^\infty | v_t \text{ is constant a.e. on } [a_{\tau-1}, a_\tau] \text{ for } \tau = 0, \ldots, T\}.$$

The approximate problem is

$(\mathcal{S}_\pi)$ find a saddle point of $\mathcal{J}$ relative to $\mathcal{U}_\pi \times \mathcal{V}_\pi$.

This is a finite-dimensional problem, and we shall show in §5 that it represents a variational approximation to the problem of finding a saddle point of $\mathcal{J}$ relative to $\mathcal{U}^r \times \mathcal{V}^{r'}$. The remainder of this section is devoted to describing how the approximate problem $(\mathcal{S}_\pi)$ leads to a pair of optimal control problems in *discrete time* which are dual to each other. To simplify the discussion, we shall assume that $t_0 = 0$, $t_1 = 1$, and $D_e = 0$. Furthermore, we will work only with the special case where the given partition is $\pi = (0, 1/T, 2/T, \ldots, (T-1)/T, 1)$. It will be clear how to extend the process to the general case.

Let $\mathcal{A}$ be the fundamental matrix for the homogeneous differential equation

$$\dot{x}_t = A_t x_t,$$

i.e., the $\mathbf{R}^{n\times n}$-valued function on $[0,1]$ which satisfies

$$\dot{\mathcal{A}}_t = A_t \mathcal{A}_t, \quad \mathcal{A}_0 = I.$$

Then the unique solution to the initial-value problem

$$\dot{x}_t = A_t x_t + B_t u_t + b_t \text{ a.e.}, \quad x_0 = B_e u_e + b_e$$

is given by

$$x_t = \mathcal{A}_t \left[B_e u_e + b_e + \int_0^t \mathcal{A}_s^{-1} (B_s u_s + b_s) ds \right]. \tag{1}$$

Now suppose that u_t is constant on $[(\tau-1)/T, \tau/T)$ for each $\tau = 0, \ldots, T$. Using (1) we write

$$x_{(\tau-1)/T} = \mathcal{A}_{(\tau-1)/T} \left[x_0 + \int_0^{(\tau-1)/T} \mathcal{A}_s^{-1}(B_s u_s + b_s) ds \right],$$

$$x_{\tau/T} = \mathcal{A}_{\tau/T} \left[x_0 + \int_0^{\tau/T} \mathcal{A}_s^{-1}(B_s u_s + b_s) ds \right].$$

We combine these two equations to get the following representation for $x_{\tau/T}$:

$$x_{\tau/T} = \left[\mathcal{A}_{\tau/T} \mathcal{A}_{(\tau-1)/T}^{-1} \right] x_{(\tau-1)/T} + \left[\int_{(\tau-1)/T}^{\tau/T} \mathcal{A}_{\tau/T} \mathcal{A}_s^{-1} B_s ds \right] u_{\tau/T} + \left[\int_{(\tau-1)/T}^{\tau/T} \mathcal{A}_{\tau/T} \mathcal{A}_s^{-1} b_s \right].$$

This formula leads us to introduce a discrete-time control system with time periods $\tau = 0, \ldots, T$, controls $\tilde{u}_\tau = u_{\tau/T}$, and states $\tilde{x}_\tau = x_{\tau/T}$. The evolution of this system is described by

$$\begin{aligned} \tilde{x}_\tau &= \tilde{A}_\tau \tilde{x}_{\tau-1} + \tilde{B}_\tau \tilde{u}_\tau + \tilde{b}_\tau, \ \tau = 1, \ldots, T \\ \tilde{x}_0 &= \tilde{B}_0 \tilde{u}_0 + \tilde{b}_0, \end{aligned} \tag{2}$$

where we define

$$\begin{aligned} \tilde{A}_\tau &= \mathcal{A}_{\tau/T} \mathcal{A}_{(\tau-1)/T}^{-1}, \\ \tilde{B}_\tau &= \int_{(\tau-1)/T}^{\tau/T} \mathcal{A}_{\tau/T} \mathcal{A}_s^{-1} B_s ds, \text{ for } \tau = 1, \ldots, T, \\ \tilde{B}_0 &= B_e, \\ \tilde{b}_\tau &= \int_{(\tau-1)/T}^{\tau/T} \mathcal{A}_{\tau/T} \mathcal{A}_s^{-1} b_s ds, \text{ for } \tau = 1, \ldots, T, \\ \tilde{b}_0 &= b_e. \end{aligned}$$

In a similar fashion, corresponding to having v_t constant on each $[(\tau-1)/T, \tau/T)$, we can introduce a dual control system:

$$\begin{aligned} \tilde{y}_\tau &= \tilde{A}_\tau^* \tilde{y}_{\tau+1} + \tilde{C}_\tau^* \tilde{v}_\tau + \tilde{c}_\tau, \quad \tau = T, \dots, 1 \\ \tilde{y}_{T+1} &= \tilde{C}_{T+1}^* \tilde{v}_{T+1} + \tilde{c}_{T+1}, \end{aligned} \tag{3}$$

where we define

$$\begin{aligned} \tilde{C}_\tau &= \int_{(\tau-1)/T}^{\tau/T} C_s \mathcal{A}_s \mathcal{A}_{(\tau-1)/T}^{-1} ds \text{ for } \tau = 1, \dots, T, \\ \tilde{C}_{T+1} &= C_e, \\ \tilde{c}_\tau &= \int_{(\tau-1)/T}^{\tau/T} \left(\mathcal{A}_{(\tau-1)/T}^{-1}\right)^* \mathcal{A}_s^* c_s ds \text{ for } \tau = 1, \dots, T, \\ \tilde{c}_{T+1} &= c_e. \end{aligned}$$

Now, for $(u, v) \in \mathcal{U}_\pi \times \mathcal{V}_\pi$, we can reexpress the value of $\mathcal{J}(u, v)$ in terms of the corresponding $(\tilde{u}, \tilde{v})$ as

$$\begin{aligned} \mathcal{J}(u, v) &= \tilde{\mathcal{J}}(\tilde{u}, \tilde{v}) \\ &:= \sum_{\tau=1}^{T} [\tilde{p}_\tau \cdot \tilde{u}_\tau + \tilde{\varphi}_\tau(\tilde{u}_\tau) - \tilde{v}_\tau \cdot \tilde{D}_\tau \tilde{u}_\tau - \tilde{\psi}_\tau(\tilde{v}_\tau) + \tilde{q}_\tau \cdot \tilde{v}_\tau - \tilde{d}_\tau] \\ &\qquad + \tilde{p}_0 \cdot \tilde{u}_0 + \tilde{\varphi}_0(\tilde{u}_0) + \tilde{q}_{T+1} \cdot \tilde{v}_{T+1} - \tilde{\psi}_{T+1}(\tilde{v}_{T+1}) - \tilde{\gamma}(\tilde{u}, \tilde{v}), \end{aligned}$$

where

$$\tilde{\gamma}(\tilde{u}, \tilde{v}) = \sum_{\tau=1}^{T+1} x_{\tau-1}(\tilde{C}_\tau^* \tilde{v}_\tau + \tilde{c}_\tau) = \sum_{\tau=0}^{T} y_{\tau+1}(\tilde{B}_\tau \tilde{u}_\tau + \tilde{b}_\tau),$$

where the "trajectories" $\tilde{x}$ and $\tilde{y}$ are the solutions of the difference equations (2) and (3). The functions $\tilde{\varphi}_\tau$ and $\tilde{\psi}_\tau$ are given by

$$\begin{aligned} \tilde{\varphi}_0(\tilde{u}_0) &= \varphi_e(\tilde{u}_0), \\ \tilde{\varphi}_\tau(\tilde{u}_\tau) &= \int_{(\tau-1)/T}^{\tau/T} \varphi_t(\tilde{u}_\tau) dt \text{ for } \tau = 1, \dots, T, \\ \tilde{\psi}_\tau(\tilde{v}_\tau) &= \int_{(\tau-1)/T}^{\tau/T} \psi_t(\tilde{v}_\tau) dt \text{ for } \tau = 1, \dots, T, \\ \tilde{\psi}_{T+1}(\tilde{v}_{T+1}) &= \psi_e(\tilde{v}_{T+1}). \end{aligned}$$

The coefficients $\tilde{p}_\tau$, $\tilde{q}_\tau$, $\tilde{D}_\tau$, and $\tilde{d}_\tau$ are given by

$$\begin{aligned} \tilde{p}_\tau &= \int_{(\tau-1)/T}^{\tau/T} \left[p_t - \left(\int_{(\tau-1)/T}^{t} \mathcal{A}_s^{-1} B_s ds \right)^* \mathcal{A}_t^* c_t \right] dt \text{ for } \tau = 1, \dots, T, \\ \tilde{p}_0 &= p_e, \\ \tilde{q}_{T+1} &= q_e, \\ \tilde{q}_\tau &= \int_{(\tau-1)/T}^{\tau/T} \left[q_t - C_t \mathcal{A}_t \left(\int_{(\tau-1)/T}^{t} \mathcal{A}_s^{-1} b_s ds \right) \right] dt \text{ for } \tau = 1, \dots, T, \end{aligned}$$

$$\begin{aligned}\tilde{D}_\tau &= \int_{(\tau-1)/T}^{\tau/T}\left[D_t + C_t\mathcal{A}_t\left(\int_{(\tau-1)/T}^{t}\mathcal{A}_s^{-1}B_s ds\right)\right]dt\\ &= \int_{(\tau-1)/T}^{\tau/T}\left[D_t + \left(\int_{t}^{\tau/T}C_s\mathcal{A}_s ds\right)\mathcal{A}_t^{-1}B_t\right]dt,\\ \tilde{d}_\tau &= \int_{(\tau-1)/T}^{\tau/T}(\mathcal{A}_t^*c_t)\cdot\left(\int_{(\tau-1)/T}^{t}\mathcal{A}_s^{-1}b_s ds\right)dt\\ &= \int_{(\tau-1)/T}^{\tau/T}\left(\int_{t}^{\tau/T}\mathcal{A}_s^*c_s ds\right)\cdot\mathcal{A}_t^{-1}b_t dt.\end{aligned}$$

Clearly $\tilde{\mathcal{J}}$ is a saddle function on $(\mathbf{R}^{k_e}\times\mathbf{R}^{k\cdot T})\times(\mathbf{R}^{l\cdot T}\times\mathbf{R}^{l_e})$. Thus the approximate problem of finding a saddle point of $\mathcal{J}$ over $\mathcal{U}_\pi\times\mathcal{V}_\pi$ leads to the following pair of optimal control problems in *discrete time* which are dual to each other:

$$(\tilde{\mathcal{P}}) \qquad \text{minimize } \tilde{F}(\tilde{u}) = \sup_{\tilde{v}}\tilde{\mathcal{J}}(\tilde{u},\tilde{v}) \text{ over all } \tilde{u}\in\mathbf{R}^{k_e}\times\left(\mathop{\times}_{\tau=1}^{T}\mathbf{R}^k\right)$$

and

$$(\tilde{\mathcal{D}}) \qquad \text{maximize } \tilde{G}(\tilde{v}) = \inf_{\tilde{u}}\tilde{\mathcal{J}}(\tilde{u},\tilde{v}) \text{ over all } \tilde{v}\in\left(\mathop{\times}_{\tau=1}^{T}\mathbf{R}^l\right)\times\mathbf{R}^{l_e}.$$

The functions $\tilde{F}$ and $\tilde{G}$ are defined by

$$\begin{aligned}\tilde{F}(\tilde{u}) = \sum_{\tau=1}^{T}&[\tilde{p}_\tau\cdot\tilde{u}_\tau + \tilde{\varphi}_\tau(\tilde{u}_\tau) - \tilde{c}_\tau\cdot\tilde{x}_{\tau-1} + (\tilde{\psi}_\tau)^*(\tilde{q}_\tau - \tilde{D}_\tau\tilde{u}_\tau - \tilde{C}_\tau\tilde{x}_{\tau-1}) - \tilde{d}_\tau]\\ &+\tilde{p}_0\cdot\tilde{u}_0 + \tilde{\varphi}_0(\tilde{u}_0) - \tilde{c}_{T+1}\cdot\tilde{x}_T + (\tilde{\psi}_{T+1})^*(\tilde{q}_{T+1} - \tilde{C}_{T+1}\tilde{x}_T),\\ \tilde{G}(\tilde{v}) = \sum_{\tau=1}^{T}&[\tilde{q}_\tau\cdot\tilde{v}_\tau - \tilde{\psi}_\tau(\tilde{v}_\tau) - \tilde{b}_\tau\cdot\tilde{y}_{\tau+1} - (\tilde{\varphi}_\tau)^*(\tilde{D}_\tau^*\tilde{v}_\tau + \tilde{B}_\tau^*\tilde{y}_{\tau+1} - \tilde{p}_\tau) - \tilde{d}_\tau]\\ &+\tilde{q}_{T+1}\cdot\tilde{v}_{T+1} - \tilde{\psi}_{T+1}(\tilde{v}_{T+1}) - \tilde{b}_0\cdot\tilde{y}_1 + (\tilde{\varphi}_0)^*(\tilde{B}_0\tilde{y}_1 - \tilde{p}_0).\end{aligned}$$

We see that $\tilde{F}$ is convex, while $\tilde{G}$ is concave. The problems $(\tilde{\mathcal{P}})$ and $(\tilde{\mathcal{D}})$ belong to the realm of convex programming. These problems have a very special structure which allows the application of various "decomposition" techniques, such as the *finite envelope method*, which has been studied by Rockafellar and Wets [24]–[26]. A computer implementation for solving problems $(\tilde{\mathcal{P}})$ and $(\tilde{\mathcal{D}})$ (in the extended linear-quadratic case) via this method has been developed by the author [29]; the code was used by Zhu and Rockafellar [31] as the basis for their numerical experiments in solving large-scale optimization problems. In [30], the finite envelope method is shown to converge for a class of saddle point problems more general than the class of extended linear-quadratic problems.

A drawback to using the discretization described in this section is the computational (and numerical) burden of calculating the integrals defining the coefficients for the discrete-time problem, in addition to the calculation of the fundamental matrix $\mathcal{A}$. A computationally simpler scheme using finite differences will be examined in §6, where we explore (from a variational point of view) its relationship to the discretization used here. Of course, in the autonomous case (where the coefficients are constant) the integrals can be calculated explicitly.

4. Epi/hypo-convergence of closed saddle functions.

In this section we set up a basic framework in which to study variational approximations for saddle point problems, which will then be used in §§5 and 6 to analyze the consistency of our approximation scheme for optimal control problems.

First we review the concept of *closed saddle functions*, which play a role in minimax theory similar to that played by closed convex functions in minimization theory. For a more detailed presentation of this topic, we refer the reader to the paper by Rockafellar [19]; see also [20]. Let X and Y be Banach spaces. A bivariate function $K : X \times Y \to \overline{\mathbf{R}}$ is called a *saddle function* if $x \mapsto K(x, y)$ is a convex function for each $y \in Y$ and $y \mapsto K(x, y)$ is a concave function for each $x \in X$. We shall think of such functions as representing a "minimax" problem, where we minimize with respect to $x \in X$ and maximize with respect to $y \in Y$. Note that the inequality

$$\inf_{x \in X} \sup_{y \in Y} K(x, y) \geq \sup_{y \in Y} \inf_{x \in X} K(x, y)$$

is always valid. In the case that

$$\inf_{x \in X} \sup_{y \in Y} K(x, y) = \sup_{y \in Y} \inf_{x \in X} K(x, y)$$

the common value is called the *saddle value* for K. A pair $(\bar{x}, \bar{y})$ is said to be a *saddle point* for K if, for all $(x, y) \in X \times Y$, it is true that

$$K(\bar{x}, y) \leq K(\bar{x}, \bar{y}) \leq K(x, \bar{y}).$$

The *minimax* problem associated with K is that of finding the saddle value and saddle points for K whenever either of these exist.

We define the *effective domain* of K as

$$\operatorname{dom} K := \operatorname{dom}_1 K \times \operatorname{dom}_2 K := \{x | K(x, \cdot) < \infty\} \times \{y | K(\cdot, y) > -\infty\}.$$

K is said to be *proper* if $\operatorname{dom} K \neq \emptyset$. As with minimization problems, where we interpret the effective domain of an $\overline{\mathbf{R}}$-valued objective function as specifying the "constraint" set, the effective domain of a bivariate function is the set to which we are necessarily restricted in the search for saddle points.

To make full use of the usual theory of minimization (or maximization), we need to impose some sort of regularity hypotheses on K. Ideally, we would like to assume that $K(x, y)$ is lower semicontinuous in x and upper semicontinuous in y. It turns out that this requirement is too restrictive: it strictly prohibits the use of saddle functions which take both of the values ∞ and $-\infty$, thus excluding the possibility of modeling constraints on both x and y through the use of infinite penalties. To deal with this difficulty, Rockafellar [19] introduced an equivalence relation for saddle functions. Within this framework, the natural regularity condition to impose is that a saddle function be equivalent to certain upper and lower regularizations of itself. Functions satisfying this condition, which we describe below, will be called *closed.*

By $\operatorname{cl}_1 K$ we denote the lower semicontinuous regularization of K in x, that is,

$$\operatorname{cl}_1 K(x, y) = \liminf_{x' \to x} K(x', y).$$

The *lower closure* $\underline{\operatorname{cl}}_1 K$ is then defined as the function which satisfies, for each $y \in Y$,

$$\underline{\operatorname{cl}}_1 K(\cdot, y) = \begin{cases} \operatorname{cl}_1 K(\cdot, y) & \text{if } \operatorname{cl}_1 K(\cdot, y) > -\infty, \\ -\infty & \text{otherwise.} \end{cases}$$

In other words, $\underline{\mathrm{cl}}_1\, K(\cdot, y)$ is the pointwise supremum of all continuous linear functionals on X which are majorized by $K(\cdot, y)$. Similarly, we define the *upper closure* of K by

$$\overline{\mathrm{cl}}_2\, K(x, \cdot) = \begin{cases} \mathrm{cl}_2\, K(x, \cdot) & \text{if } \mathrm{cl}_2\, K(x, \cdot) < +\infty, \\ +\infty & \text{otherwise,} \end{cases}$$

where $\mathrm{cl}_2\, K$ is the upper semicontinuous regularization of K in y. Note that if L is another saddle function on $X \times Y$ for which $K \le L$, then $\underline{\mathrm{cl}}_1\, K \le \underline{\mathrm{cl}}_1\, L$ and $\overline{\mathrm{cl}}_2\, K \le \overline{\mathrm{cl}}_2\, L$. We say that two saddle functions K and L are *equivalent* if they have the same upper and lower closures, i.e., if

$$\underline{\mathrm{cl}}_1\, K = \underline{\mathrm{cl}}_1\, L \text{ and } \overline{\mathrm{cl}}_2\, K = \overline{\mathrm{cl}}_2\, L.$$

The function K is said to be *closed* if it is equivalent to both $\underline{\mathrm{cl}}_1\, K$ and $\overline{\mathrm{cl}}_2\, K$, in which case we write $\overline{K} = \overline{\mathrm{cl}}_2\, K$ and $\underline{K} = \underline{\mathrm{cl}}_1\, K$. If K is closed and $\underline{K} \le L \le \overline{K}$, then we write $L \in [\underline{K}, \overline{K}]$. The next two propositions are immediate consequences of the definitions.

PROPOSITION 1. *Suppose K is a closed saddle function.*

(i) $L \in [\underline{K}, \overline{K}]$ *if and only if K and L are equivalent.*

(ii) *If L is also closed, then L is equivalent to K if and only if $\underline{K} \le \overline{L}$ and $\underline{L} \le \overline{K}$.*

PROPOSITION 2. *The following are true for any saddle function K:*

(i) *If* $\mathrm{dom}_1\, K = \emptyset$, *then* $\overline{\mathrm{cl}}_2\, K \equiv \infty$.

(ii) *If* $\mathrm{dom}_2\, K = \emptyset$, *then* $\underline{\mathrm{cl}}_1\, K \equiv -\infty$.

In particular, the only closed improper saddle functions are $K \equiv \infty$ and $K \equiv -\infty$.

The following shows why the term *equivalent* is justified as applied to saddle functions.

PROPOSITION 3. *Suppose that K is closed and equivalent to L.*

(i) $\mathrm{cl}(\mathrm{dom}\, L) = \mathrm{cl}(\mathrm{dom}\, K)$.

(ii) $L(x, y) = \begin{cases} -\infty & \textit{if } x \in \mathrm{dom}_1\, K,\ y \in Y \backslash \mathrm{cl}(\mathrm{dom}_2\, K), \\ \infty & \textit{if } x \in X \backslash \mathrm{cl}(\mathrm{dom}_1\, K),\ y \in \mathrm{dom}_2\, K. \end{cases}$

(iii) *If a saddle value exists for K, then it is also the saddle value for L.*

(iv) *If $(\bar{x}, \bar{y})$ is a saddle point for K, then it is also a saddle point for L.*

(v) *L is finite on* $\mathrm{dom}_1\, \overline{K} \times \mathrm{dom}_2\, \underline{K}$.

We now discuss the notion of epi/hypo-convergence, which is a generalization of epi-convergence. There are actually several distinct concepts of epi/hypo-convergence which are useful in different settings; see [3], [4], [23]. In this paper, we shall use a variant of the definitions used in [4] and [23]. In what follows, we denote by s and w the strong (norm) and weak topologies (respectively) on a Banach space.

Consider a sequence $\{K_n\}$ of closed saddle functions on $X \times Y$. We define the *epi/hypo-limits superior* and *inferior* of $\{K_n\}$ to be

$$(\text{seq-e}_\text{s}/\text{h}_\text{w}\text{-ls}\, K_n)(x, y) = \sup_{y_n \xrightarrow[w]{} y} \inf_{x_n \xrightarrow[s]{} x} \limsup_{n \to \infty} K_n(x_n, y_n),$$

$$(\text{seq-h}_\text{s}/\text{e}_\text{w}\text{-li}\, K_n)(x, y) = \inf_{x_n \xrightarrow[w]{} x} \sup_{y_n \xrightarrow[s]{} y} \liminf_{n \to \infty} K_n(x_n, y_n).$$

We say that K_n *epi/hypo-converges* to K if

$$\text{seq-e}_\text{s}/\text{h}_\text{w}\text{-ls}\, \overline{K}_n \le \overline{K} \text{ and } \text{seq-h}_\text{s}/\text{e}_\text{w}\text{-li}\, \underline{K}_n \ge \underline{K}.$$

In this case, we write $K = \text{e/h-lim}\, K_n$. Note that this is actually a convergence of equivalence classes. It is clear from the definitions that if $\{K_{n_k}\}$ is any subsequence of $\{K_n\}$, then $K = \text{e/h-lim}\, K_{n_k}$.

The next theorem demonstrates the variational aspect of epi/hypo-convergence. It is analogous to a similar result given by Attouch, Azé, and Wets [4] and can be proven by the same argument used there. The pair $(\bar{x}, \bar{y})$ is said to be an *ε-saddle point* (for $\varepsilon \geq 0$) for K if

$$\sup_y K(\bar{x}, y) - \varepsilon \leq K(\bar{x}, \bar{y}) \leq \inf_x K(x, \bar{y}) + \varepsilon.$$

THEOREM 3. *Consider a sequence $\{K_n\}$ of closed saddle functions which epi/hypo-converges to the closed saddle function K. Suppose that $(\bar{x}_k, \bar{y}_k)$ is an ε-saddle point for K_{n_k} for some subsequence of $\{K_n\}$. If $\bar{x}_k \xrightarrow[w]{} \bar{x}$, $\bar{y}_k \xrightarrow[w]{} \bar{y}$, and $\varepsilon_k \longrightarrow \varepsilon$, then $(\bar{x}, \bar{y})$ is a 2ε-saddle point for $[\underline{K}, \overline{K}]$ and*

$$K(\bar{x}, \bar{y}) \in [\limsup_{n \to \infty} K_n(\bar{x}_n, \bar{y}_n) - \varepsilon, \liminf_{n \to \infty} K_n(\bar{x}_n, \bar{y}_n) + \varepsilon].$$

In the next two sections we will deal with saddle functions of the form

$$(x, y) \mapsto \Phi(x) - \Psi(y) - \Gamma(x, y).$$

The following result will play a key role in the analysis. Recall that a sequence $\{f_n\}$ of convex functions *Mosco epi-converges* to a function f if

$$\inf_{x_n \xrightarrow[s]{} x} \limsup_{n \to \infty} f(x_n) \leq f(x) \leq \inf_{x_n \xrightarrow[w]{} x} \liminf_{n \to \infty} f(x_n) \text{ for all } x,$$

in which case we write $f = \text{M-e-lim}\, f_n$. We refer the reader to the book by Attouch [1] for an exposition of epi-convergence.

THEOREM 4. *Let $\Gamma : X \times Y \longrightarrow \mathbf{R}$ be a continuous biaffine map. Suppose $\Phi, \Phi_n : X \longrightarrow \overline{\mathbf{R}}$ and $\Psi, \Psi_n : Y \longrightarrow \overline{\mathbf{R}}$ are proper, lower semicontinuous convex functions such that $\Phi = \text{M-e-lim}\, \Phi_n$ and $\Psi = \text{M-e-lim}\, \Psi_n$. Let K_n and K be any saddle functions on $X \times Y$ for which*

$$K_n(x, y) = \Phi_n(x) - \Psi_n(y) - \Gamma(x, y) \textit{ whenever } \infty - \infty \textit{ does not occur, and}$$
$$K(x, y) = \Phi(x) - \Psi(y) - \Gamma(x, y) \textit{ whenever } \infty - \infty \textit{ does not occur.}$$

Then K_n and K are equivalent to closed, proper saddle functions, and K_n epi/hypo-converges to K.

Proof. It is easy to see that $\overline{\mathrm{cl}}_2\, K$ and $\underline{\mathrm{cl}}_1\, K$ are given by

$$\overline{\mathrm{cl}}_2\, K(x, y) = \begin{cases} \infty & \text{if } \Phi(x) = \infty, \\ \Phi(x) - \Psi(y) - \Gamma(x, y) & \text{otherwise,} \end{cases}$$
$$\underline{\mathrm{cl}}_1\, K(x, y) = \begin{cases} -\infty & \text{if } \Psi(y) = \infty, \\ \Phi(x) - \Psi(y) - \Gamma(x, y) & \text{otherwise.} \end{cases}$$

It is clear from these descriptions that $\overline{\mathrm{cl}}_2\, K$ and $\underline{\mathrm{cl}}_1\, K$ are both saddle functions and that $\underline{\mathrm{cl}}_1(\overline{\mathrm{cl}}_2\, K) = \underline{\mathrm{cl}}_1\, K$ and $\overline{\mathrm{cl}}_2(\underline{\mathrm{cl}}_1\, K) = \overline{\mathrm{cl}}_2\, K$. The properness of K follows from that of Φ and Ψ. By the same argument, each K_n is equivalent to a proper, closed saddle function.

Now fix $(\bar{x}, \bar{y}) \in X \times Y$. We will show that $(\text{seq-e}_{\text{s}}/\text{h}_{\text{w}}\text{-ls}\, \overline{K}_n)(\bar{x}, \bar{y}) \leq \overline{K}(\bar{x}, \bar{y})$; the proof that $(\text{seq-h}_{\text{s}}/\text{e}_{\text{w}}\text{-li}\, \underline{K}_n)(\bar{x}, \bar{y}) \geq \underline{K}(\bar{x}, \bar{y})$ is similar. If $\overline{K}(\bar{x}, \bar{y}) = \infty$, we're finished. Assume then that $\overline{K}(\bar{x}, \bar{y}) < \infty$ so that $\Phi(\bar{x}) < \infty$. Suppose $\{y_n\}$ converges weakly to $\bar{y}$, and fix $\alpha > \overline{K}(\bar{x}, \bar{y})$. We need to find a sequence $\{x_n\}$, converging in norm to $\bar{x}$, such that $\limsup \overline{K}_n(x_n, y_n) \leq \alpha$. Choose $\alpha_1 \in (\overline{K}(\bar{x}, \bar{y}), \alpha)$. Since $\Phi = \text{M-e-lim}\, \Phi_n$, there is a sequence $\{x_n\}$ converging to $\bar{x}$ such that

$$\limsup \Phi_n(x_n) \leq \Phi(\bar{x}) + \frac{\alpha - \alpha_1}{2}.$$

Also, $\Psi = \text{M-e-lim}\, \Psi_n$ implies that

$$\liminf \Psi_n(y_n) \geq \Psi(\bar{y}) > -[\alpha_1 - \Phi(\bar{x}) + \Gamma(\bar{x}, \bar{y})].$$

Combining these with the sequential $s \times w$-continuity of Γ allows us to find a positive integer N so that the following three inequalities hold whenever $n \geq N$:

$$\begin{aligned} -\Psi_n(y_n) &< \alpha_1 - \Phi(\bar{x}) + \Gamma(\bar{x}, \bar{y}), \\ -\Gamma(x_n, y_n) &< -\Gamma(\bar{x}, \bar{y}) + \frac{\alpha - \alpha_1}{2}, \\ \Phi_n(x_n) &\leq \Phi(\bar{x}) + \frac{\alpha - \alpha_1}{2}. \end{aligned}$$

Thus, for all $n \geq N$, we have

$$\begin{aligned} K(x_n, y_n) &= \Phi_n(x_n) - \Psi_n(y_n) - \Gamma(x_n, y_n) \\ &\leq \Phi(x_n) - \Psi(y_n) - \Gamma(x_n, y_n) \\ &< \left[\Phi(\bar{x}) + \frac{\alpha - \alpha_1}{2}\right] + [\alpha_1 - \Phi(\bar{x}) + \Gamma(\bar{x}, \bar{y})] + \left[\frac{\alpha - \alpha_1}{2} - \Gamma(\bar{x}, \bar{y})\right] \\ &= \alpha \end{aligned}$$

as desired. □

5. Consistency of internal approximations. We now give our main consistency result for the approximation scheme described in §3. Let $\{\pi^\nu : \nu \in \mathbf{N}\}$ be an increasing sequence of partitions of $[t_0, t_1]$ such that $|a_i^\nu - a_{i+1}^\nu| \longrightarrow 0$ uniformly in i as $\nu \longrightarrow 0$. Define

$$\mathcal{J}_\nu(u, v) = \begin{cases} \mathcal{J}(u, v) & \text{if } u \in \mathcal{U}_{\pi^\nu}, v \in \mathcal{V}_{\pi^\nu}, \\ -\infty & \text{if } u \in \mathcal{U}_{\pi^\nu}, v \notin \mathcal{V}_{\pi^\nu}, \\ \infty & \text{if } u \notin \mathcal{U}_{\pi^\nu}. \end{cases}$$

The following theorem will be concerned with this sequence of problems. In addition, we will make use of the following condition on $r, r' \in [1, \infty)$:

(4) the map $(u, v) \mapsto \int_{t_0}^{t_1} v_t \cdot D_t u_t dt$ defines a continuous functional on $\mathcal{U}^r \times \mathcal{V}^{r'}$.

This condition is satisfied if $r \geq r'/(r'-1)$ (or equivalently, if $r' \geq r/(r-1)$), in which case $\mathcal{L}^r \subset (\mathcal{L}^{r'})^*$ and $\mathcal{L}^{r'} \subset (\mathcal{L}^r)^*$. If $D \equiv 0$, then the condition is satisfied for any choice of $r, r' \in [1, \infty)$.

THEOREM 5. *Consider $r, r' \in [1, \infty)$ satisfying condition* (4). *Assume there exist intervals $I_1, \ldots, I_M$ with $[0, 1] = \cup_{i=1}^M I_i$ such that, for each i, one has either*

(a) *$\varphi_t = \bar{\varphi}_t^i + \delta_{U^i}$ for $t \in I_i$, where U^i is a closed, convex set and $\bar{\varphi}_t^i(\cdot)$ is a proper, lower semicontinous convex function with $u \mapsto \int_{I_i} \bar{\varphi}_t^i(u_t)\, dt$ continuous on $\mathcal{U}^r$, or*

(b) *$\varphi_t = \bar{\varphi}^i$ for $t \in I_i$, where $\bar{\varphi}^i$ is a proper,* inf*-compact convex function.*

Similarly, assume there exist intervals $I'_1, \dots, I'_{M'}$ *with* $[0,1] = \cup_{i=1}^{M'} I'_i$ *such that, for each* i*, one has either*

(a′) $\psi_t = \bar{\psi}^i_t + \delta_{V^i}$ *for* $t \in I'_i$*, where* V^i *is a closed, convex set and* $\bar{\psi}^i_t(\cdot)$ *is a proper, lower semicontinous convex function with* $v \mapsto \int_{I'_i} \bar{\psi}^i_t(v_t)\,dt$ *continuous on* $\mathcal{V}^r$*; or*

(b′) $\psi_t = \bar{\psi}^i$ *for* $t \in I'_i$*, where* $\bar{\psi}^i$ *is a proper, inf-compact convex function.*

Then $\mathcal{J}$ *and* $\mathcal{J}_\nu$ *are equivalent to closed saddle functions on* $\mathcal{U}^r \times \mathcal{V}^{r'}$*, and* $\mathcal{J}_\nu$ *epi/hypo-converges to* $\mathcal{J}$*.*

An immediate consequence of Theorem 5 is the following.

COROLLARY. *Assume the hypotheses for Theorem* 5 *are satisfied, and suppose* $(\bar{u}^\nu, \bar{v}^\nu)$ *is a saddle point for* $\mathcal{J}^\nu$*. If, for some* $s \in [\max\{r, r'\}, \infty)$*, one can guarantee that* $(\bar{u}^\nu, \bar{v}^\nu)$ *converges in the weak topology on* $\mathcal{U}^s \times \mathcal{V}^s$*, then the limit point is a saddle point for* $\mathcal{J}$ *relative to* $\mathcal{U}^s \times \mathcal{V}^s$*.*

Note that for $(\bar{u}, \bar{v})$ to be a saddle point for $\mathcal{J}$, it is sufficient to show that the sequence of approximates merely clusters (weakly) at $(\bar{u}, \bar{v})$. This clustering is guaranteed, for example, when the effective domains $\operatorname{dom} \varphi_t$ and $\operatorname{dom} \psi_t$ are bounded for all t (and therefore uniformly bounded, by continuity) or more generally if $\operatorname{dom} \varphi_t \in \alpha_t \mathbf{B}_k$ and $\operatorname{dom} \psi_t \in \beta_t \mathbf{B}_l$ for some $\mathcal{L}^s$ functions α and β (with $s > 1$).

To show the validity of Theorem 5, we use some basic facts about constrained approximation of measurable functions by simple and step functions. We state these explicitly as lemmas but omit the proofs since they are merely variants of well-known arguments in elementary measure theory.

LEMMA 1. *Let* $(\Omega, \mathcal{F}, \mu)$ *be a finite-measure space, and suppose* $1 \leq p \leq \infty$*. Consider* $f \in \mathcal{L}^p_d(\Omega, \mathcal{F}, \mu)$ *and an inf-compact convex function* h *on* $\mathbf{R}^d$*. If* $f(\omega) \in \operatorname{dom} h$ *almost everywhere (a.e.)* $[\mu]$*, then there is a sequence* $\{f_\nu\}$ *of simple functions satisfying*

(i) $h(f_\nu(\omega)) \leq h(f_{\nu+1}(\omega)) \leq h(f(\omega))$ *a.e.* $[\mu]$*.*

(ii) $f_\nu \to f$ *a.e.* $[\mu]$*.*

(iii) $f_\nu \to f$ *in* $\mathcal{L}^p_d(\Omega, \mathcal{F}, \mu)$*.*

LEMMA 2. *Let* $(\Omega, \mathcal{F}, \mu)$ *be a finite-measure space with* $\mathcal{F} = \sigma(\mathcal{F}_0)$*, where* $\mathcal{F}_0$ *is a field on* Ω*. Let* $p \in [1, \infty)$*. Suppose that* $h : \mathbf{R}^d \to \overline{\mathbf{R}}$*. If* $f : \Omega \to \mathbf{R}^d$ *is an* $\mathcal{F}$*-simple function, then, for any* $\varepsilon > 0$*, there exists an* $\mathcal{F}_0$*-simple function* $\tilde{f}$ *with the same range as* f *for which* $\|\tilde{f} - f\|_p < \varepsilon$ *and*

$$\mu\{\omega \in \Omega | h(\tilde{f}(\omega)) > h(f(\omega))\} < \varepsilon.$$

Proof of Theorem 5. We shall prove the theorem for the special case $k_e = l_e = 0$, $b_e = c_e = 0$. In addition, we will assume that $I_i \cap I_j = \emptyset$ whenever $i \neq j$ and that each endpoint of each I_i coincides with the endpoint of some cell for some π^ν. The proof of the general case is similar.

For $u \in \mathcal{L}^r_k$ and $v \in \mathcal{L}^{r'}_l$, define

$$\Phi(u) = \int_{t_0}^{t_1} \varphi_t(u_t)dt, \quad \Psi(v) = \int_{t_0}^{t_1} \psi_t(v_t)dt,$$

and

$$\Gamma(u, v) = \int_{t_0}^{t_1} [v_t \cdot D_t u_t - p_t \cdot u_t - q_t \cdot v_t]dt + \gamma(u, v).$$

Then, for all $(u, v) \in \mathcal{L}^r_k \times \mathcal{L}^{r'}_l$, we have

$$\mathcal{J}(u, v) = \begin{cases} \infty & \text{if } \Phi(u) = \infty, \\ \Phi(u) - \Psi(v) - \Gamma(u, v) & \text{otherwise.} \end{cases}$$

Also define $\Phi_\nu = \Phi + \delta(\cdot|\mathcal{U}_{\pi^\nu})$ and $\Psi_\nu = \Psi + \delta(\cdot|\mathcal{V}_{\pi^\nu})$. We need to show that Φ, Ψ, Γ, Φ_ν, and Ψ_ν satisfy the hypotheses for Theorem 4, thereby obtaining the stated epi/hypo-convergence. It is easy to see that Φ, Ψ, Φ_ν, and Ψ_ν are proper, lower semicontinuous, convex functions and that Γ is a (norm) continuous biaffine functional. Furthermore, $\mathcal{J}_\nu(u,v)$ agrees with $\Phi_\nu(u) - \Psi_\nu(v) - \Gamma(u,v)$ whenever $\infty - \infty$ does not occur. We must therefore show that $\Phi = \text{M-e-lim}\,\Phi_\nu$ and $\Psi = \text{M-e-lim}\,\Psi_\nu$. Since $\Phi_\nu(u)$ is decreasing in ν for all u, it suffices to show

$$(5) \qquad \forall \bar{u} \in \mathcal{U}^r,\ \exists \bar{u}^\nu \underset{s}{\rightarrow} \bar{u} \text{ such that } \bar{u}^\nu \in \mathcal{U}_{\pi^\nu} \text{ and } \limsup \Phi_\nu(\bar{u}^\nu) \le \Phi(\bar{u}).$$

Fix $\bar{u} \in \mathcal{L}^r_k$. Suppose $\Phi(\bar{u}) = \infty$. Then $\limsup \Phi(u^\nu) \le \Phi(\bar{u})$ for any sequence converging (in norm) to $\bar{u}$; there exists such a sequence with $u^\nu \in \mathcal{U}_{\pi^\nu}$ since $\cup \mathcal{U}_{\pi^\nu}$ is dense in $\mathcal{L}^r_k$.

Assume then that $\Phi(\bar{u}) < \infty$. Then $\bar{u}_t \in \operatorname{dom} \varphi_t$ for almost every t. We shall construct an approximating sequence $\{\bar{u}^\nu\}$ on I_i for each i. First consider I_i satisfying hypothesis (a). Then $\bar{u}_t \in U^i$ for almost every $t \in I_i$. By Lemma 1 (taking $h(\xi) = |\xi| + \delta(\xi|U^i)$), there exists a sequence $\tilde{u}^m$ of simple functions, with $\tilde{u}^m_t \in U^i$ for all $t \in I_i$, such that $\|\tilde{u}^m - \bar{u}\|_r \to 0$ as $m \to 0$. For each $m \in \mathbf{N}$, by Lemma 2, there exists u^m which is measurable with respect to one of the partitions π^ν, has the same range as $\tilde{u}^m$, and has $\|\tilde{u}^m - u^m\|_r < 1/m$. Thus, u^m converges to $\bar{u}$ in $\mathcal{L}^r(I_i)$. Since $u^m_t \in U^i$, we have $\int_{I_i} \varphi_t(u^m_t)dt \to \int_{I_i} \varphi_t(\bar{u}_t)dt$, because the map $u \mapsto \int_{I_i} \bar{\varphi}^i_t(u_t)dt$ is continuous on $\mathcal{L}^r_k$.

Now consider I_i satisfying hypothesis (b). Then $\bar{u}_t \in \operatorname{dom} \bar{\varphi}^i$ for almost every $t \in I_i$. By Lemma 1 (with $h = \bar{\varphi}^i$), there exists a sequence $\{\tilde{u}^m\}$ of simple functions such that $\|\tilde{u}^m - \bar{u}\|_r \to 0$ as $m \to \infty$ and $\bar{\varphi}^i(\tilde{u}^m_t) \le \bar{\varphi}^i(\bar{u}_t)$ almost everwhere. In particular we have $\int_{I_i} \varphi_t(\tilde{u}^m_t)dt \le \int_{I_i} \varphi_t(\bar{u}_t)dt$ for all m. For each $m \in \mathbf{N}$ there exists, by Lemma 2, u^m, which is measurable with respect to one of the partitions π^ν and has $\|\tilde{u}^m - u^m\|_r < 1/m$. Moreover, u^m may be chosen to have the same range as $\tilde{u}^m$ and to satisfy

$$\mu\{t \in I_i | \bar{\varphi}^i(u^m_t) > \bar{\varphi}^i(\tilde{u}^m_t)\} < \frac{1}{m \cdot \max\{1, \bar{\alpha}\}},$$

where $\bar{\alpha} = \max\{\bar{\varphi}^i(\tilde{u}^m_t) - \bar{\varphi}^i(\tilde{u}^m_{t'}) | t, t' \in I_i\}$. Hence u^m may be chosen so that

$$\int_{I_i} \varphi_t(u^m_t)dt < \int_{I_i} \varphi_t(\tilde{u}^m_t)dt + (1/m).$$

Thus, the sequence $\{u^m\}$ converges to $\bar{u}$ in $\mathcal{L}^r(I_i)$, and

$$\limsup_{m\to\infty} \int_{I_i} \varphi_t(u^m_t)dt \le \int_{I_i} \varphi_t(\bar{u}_t)dt.$$

So we now have a sequence $\{u^m\}$ of $\cup_\nu \pi^\nu$-step functions which converges (in norm) to $\bar{u}$ with $\limsup_{m\to\infty} \Phi(u^m) \le \Phi(\bar{u})$. Choose $\{\nu_m\}$ with $\nu_m < \nu_{m+1}$ so that $u^m \in \mathcal{U}_{\pi^\nu}$ for $\nu = \nu_m$. We now define a sequence $\{\bar{u}^\nu\}$ satisfying (5). If $\nu_1 = 1$, set $\bar{u}^1 = u^1$; otherwise, choose $\bar{u}^1$ to be an arbitrary element of $\mathcal{U}_{\pi^1}$, and set $\bar{u}^\nu = \bar{u}^1$ for $\nu = 2, \ldots, \nu_1 - 1$. For $\nu = \nu_m, \ldots, \nu_{m-1}$ set $\bar{u}^\nu = u^m$. Thus the requirement of Mosco epi-convergence for Φ_ν is satisfied. A similar argument yields $\Psi = \text{M-e-lim}\,\Psi^\nu$. The conclusion then follows from Theorem 4. □

6. Approximation by finite differences. In the preceding sections, we introduced a model in optimal control and demonstrated that, under certain assumptions, discretization by partitioning of the time interval can be considered as a form of variational approximation (Theorem 5). Such a discretization leads to a problem in *discrete-time* optimal control. The coefficients for the discretized problem are given by integral formulas involving the fundamental matrix associated with the linear dynamics. These integrals could be evaluated numerically by various methods. In this section, however, we will take an alternative route and discretize the original problem directly using finite differences. It will be shown that this also leads to a consistent variational approximation. To simplify the presentation only the Euler forward-difference scheme is considered.

As before, we shall actually approximate the following saddle point problem which is associated with the original optimal control problem:

$$(\mathcal{S}) \qquad \text{find a saddle point } (u,v) \text{ of } \mathcal{J} \text{ relative to } \mathcal{U}^2 \times \mathcal{V}^2.$$

Here we have

$$\mathcal{J}(u,v) = \int_0^1 [p_t \cdot u_t + \varphi_t(u_t) + q_t \cdot v_t - \psi_t(v_t) - v_t \cdot D_t u_t]dt \\ + p_e \cdot u_e + \varphi_e(u_e) + q_e \cdot v_e - \psi_e(v_e) - \gamma(u,v),$$

where

$$\begin{aligned}\gamma(u,v) &= \int_0^1 x_t \cdot (C^* v_t + c)dt + x_1 \cdot (C_e^* v_e + c_e) \\ &= \int_0^1 y_t \cdot (Bu_t + b)dt + y_0 \cdot (B_e u_e + b_e)\end{aligned}$$

and the dynamics are given by

$$\begin{aligned}\dot{x}_t &= A_t x_t + B_t u_t + b_t \text{ a.e.}, & x_0 &= B_e u_e + b_e, \\ -\dot{y}_t &= A_t^* y_t + C_t^* v_t + c_t \text{ a.e.}, & y_1 &= C_e^* v_e + c_e.\end{aligned}$$

We assume that the functions φ_t and ψ_t satisfy suitable conditions for $\mathcal{J}$ to be a closed saddle function (see §§4 and 5).

We partition the unit interval with a uniform step size $h = 1/T$, where T is a positive integer. Associated with this step size will be two approximate saddle point problems. The first is the discrete-time problem given in §3, which consists of using the original saddle function and dynamics but restricting the controls to be constant on each interval $[\tau h, (\tau+1)h)$:

$$(\tilde{\mathcal{S}}_h) \qquad \begin{array}{l}\text{find a saddle point } (u_0^h, \ldots, u_T^h; v_1^h, \ldots, v_{T+1}^h) \text{ of } \tilde{\mathcal{J}}_h \\ \text{relative to } (\mathbf{R}^{k_e} \times \mathbf{R}^{kT}) \times (\mathbf{R}^{lT} \times \mathbf{R}^{l_e}).\end{array}$$

Here we define

$$\begin{aligned}&\tilde{\mathcal{J}}_h(u_0^h, \ldots, u_T^h; v_1^h, \ldots, v_{T+1}^h) \\ &\quad = \sum_{\tau=1}^{T} [\tilde{p}_\tau^h \cdot u_\tau^h + \tilde{\varphi}_\tau^h(u_\tau^h) + \tilde{q}_\tau^h \cdot v_\tau^h - \tilde{\psi}_\tau^h(v_\tau^h) - v_\tau^h \tilde{D}_\tau^h u_\tau^h - d_\tau^h] \\ &\qquad\qquad + p_e \cdot u_0^h + \varphi_e(u_0^h) + q_e \cdot v_{T+1}^h - \psi_e(v_{T+1}^h) - \tilde{\gamma}_h(u_0^h, \ldots, u_T^h; v_1^h, \ldots, v_{T+1}^h),\end{aligned}$$

where

$$\begin{aligned}\tilde\gamma_h(u_0^h,\dots,u_T^h;v_1^h,\dots,v_{T+1}^h) &= \sum_{\tau=1}^{T} \tilde x^h_{\tau-1}\cdot((\tilde C^h_\tau)^* v^h_\tau + \tilde c^h_\tau) + \tilde x^h_T\cdot(C_e^* v^h_{T+1}+c_e)\\ &= \sum_{\tau=1}^{T} \tilde y^h_{\tau+1}\cdot(\tilde B^h_\tau u^h_\tau + \tilde b^h_\tau) + \tilde y^h_1\cdot(B_e u^h_0 + b_e)\end{aligned}$$

and

$$\begin{aligned}\tilde x^h_\tau &= \tilde A^h_\tau \tilde x^h_{\tau-1} + \tilde B^h_\tau u^h_\tau + \tilde b^h_\tau, \text{ for } \tau = 1,\dots,T\\ \tilde x^h_0 &= B_e u^h_0 + b_e,\\ \tilde y^h_\tau &= (\tilde A^h_\tau)^* \tilde y^h_{\tau+1} + (\tilde C^h_\tau)^* v^h_\tau + \tilde c^h_\tau, \text{ for } \tau = T,\dots,1\\ \tilde y^h_{T+1} &= C_e^* v^h_{T+1} + c_e.\end{aligned}$$

The functions $\tilde\varphi^h$, $\tilde\psi^h$ and the coefficients $\tilde A^h_\tau$, $\tilde B^h_\tau$, $\tilde b^h_\tau$, $\tilde C^h_\tau$, $\tilde c^h_\tau$, $\tilde D^h_\tau$, $\tilde d^h_\tau$, $\tilde p^h_\tau$, $\tilde q^h_\tau$ are defined as in §3.

The following approximate problem is given using forward finite differences:

$(\hat{\mathcal{S}}_h)$ find a saddle point $(u_0^h,\dots,u_T^h;v_1^h,\dots,v_{T+1}^h)$ of $\hat{\mathcal{J}}_h$ relative to $(\mathbf{R}^{k_e}\times\mathbf{R}^{kT})\times(\mathbf{R}^{lT}\times\mathbf{R}^{l_e})$.

This problem has the same form as $(\tilde{\mathcal{S}}_h)$, but we replace $\tilde A^h_\tau$, $\tilde B^h_\tau$, $\tilde b^h_\tau$, $\tilde C^h_\tau$, $\tilde c^h_\tau$, $\tilde D^h_\tau$, $\tilde d^h_\tau$, $\tilde p^h_\tau$, and $\tilde q^h_\tau$ with coefficients $\hat A^h_\tau$, $\hat B^h_\tau$, $\hat b^h_\tau$, $\hat C^h_\tau$, $\hat c^h_\tau$, $\hat D^h_\tau$, $\hat d^h_\tau$ $\hat p^h_\tau$, and $\hat q^h_\tau$ defined by

$$\begin{aligned}&\hat A^h_\tau = I + hA_{(\tau-1)h},\\ &\hat B^h_\tau = hB_{(\tau-1)h}, \qquad \hat b^h_\tau = hb_{(\tau-1)h},\\ &\hat C^h_\tau = hC_{(\tau-1)h}, \qquad \hat c^h_\tau = hc_{(\tau-1)h},\\ &\hat D^h_\tau = hD_{(\tau-1)h}, \qquad \hat d^h_\tau = 0,\\ &\hat p^h_\tau = hp_{(\tau-1)h}, \qquad \hat q^h_\tau = hq_{(\tau-1)h}.\end{aligned}$$

We also define $\hat\varphi^h_\tau(u^h_\tau) = \varphi_{(\tau-1)h}(u^h_\tau)$ and $\hat\psi^h_\tau(v^h_\tau) = \psi_{(\tau-1)h}(v^h_\tau)$. The dynamics in this problem are given by

$$\begin{aligned}\hat x^h_\tau &= \hat A^h_\tau \hat x^h_{\tau-1} + \hat B^h_\tau u^h_\tau + \hat b^h_\tau \text{ for } \tau = 1,\dots,T,\\ \hat x^h_0 &= B_e u^h_0 + b_e,\\ \hat y^h_\tau &= (\hat A^h_\tau)^* \hat y^h_{\tau+1} + (\hat C^h_\tau)^* v^h_\tau + \hat c^h_\tau \text{ for } \tau = T,\dots,1,\\ \hat y^h_{T+1} &= C_e^* v^h_{T+1} + c_e.\end{aligned}$$

Note that the trajectories associated with problem $(\hat{\mathcal{S}}_h)$ are denoted by $(\hat x^h_0,\dots,\hat x^h_T)$ and $(\hat y^h_1,\dots,\hat y^h_{T+1})$, whereas the trajectories for $(\tilde{\mathcal{S}}_h)$ are indicated by tildes.

We shall also think of $(\tilde{\mathcal{S}}_h)$ and $(\hat{\mathcal{S}}_h)$ as problems on $\mathcal{U}^2\times\mathcal{V}^2$ by identifying $\mathbf{R}^{k_e}\times\mathbf{R}^{kT}$ and $(\mathbf{R}^{lT}\times\mathbf{R}^{l_e})$ with the subspaces $\mathcal{U}_h$ and $\mathcal{V}_h$ given by

$$\begin{aligned}\mathcal{U}_h &= \{u\in\mathcal{U}^2 \mid u_t \text{ is constant a.e. on } [(\tau-1)h,\tau h) \text{ for } \tau = 1,\dots,T\},\\ \mathcal{V}_h &= \{v\in\mathcal{V}^2 \mid v_t \text{ is constant a.e. on } [(\tau-1)h,\tau h) \text{ for } \tau = 1,\dots,T\}.\end{aligned}$$

Thus the point $(u_0^h, u_1^h, \dots, u_T^h)$ is identified with the point u, where $u_e = u_0^h$ and $u_t = u_\tau^h$ for almost every $t \in [(\tau-1)h, \tau h)$ for $\tau = 1, \dots, T$. The norm on $\mathbf{R}^{k_e} \times \mathbf{R}^{kT}$ is also given through the identification with $\mathcal{U}_h$:

$$\|(u_0^h, \dots, u_T^h)\|_h = \left(h \sum_{\tau=1}^{T} |u_\tau^h|^2 + |u_0^h|^2 \right)^{1/2}.$$

Here $|\cdot|$ denotes the Euclidean norm on $\mathbf{R}^k$ (or $\mathbf{R}^{k_e}$). A similar formula gives the norm on $\mathbf{R}^{lT} \times \mathbf{R}^{l_e}$. The functionals $\tilde{\mathcal{J}}_h$ are extended to $\mathcal{U}^2 \times \mathcal{V}^2$ by taking $\tilde{\mathcal{J}}_h(u,v)$ to be $-\infty$ if $u \in \mathcal{U}_h$ but $v \notin \mathcal{V}_h$, and to be ∞ if $u \notin \mathcal{U}_h$. We extend $\hat{\mathcal{J}}_h$ in the same manner.

As one may expect, the saddle functions $\tilde{\mathcal{J}}_h$ and $\hat{\mathcal{J}}_h$ are closely related. An important aspect of this relationship is given by the following proposition, which follows essentially from the fact that the Euler forward-difference scheme is a first-order method, combined with the continuity of the affine mapping of controls to trajectories given by the dynamics.

PROPOSITION 4. *Suppose that the coefficients $A_t, B_t, b_t, C_t, c_t, D_t, p_t, q_t$ are Lipschitzian in t. Assume there is a nonnegative constant α so that, for all $u \in \mathbf{R}^k$ and $v \in \mathbf{R}^l$, one has*

$$\varphi_s(u) \geq \varphi_t(u) - \alpha|u||s-t| \text{ and } \psi_s(v) \geq \psi_t(v) - \alpha|v||s-t| \text{ for all } s,t \in [0,1].$$

Then there exists an $r \geq 0$ such that

$$(6) \quad \tilde{\mathcal{J}}_h(u,v) - rh(\|u\|+1)(\|v\|+1) \leq \hat{\mathcal{J}}_h(u,v) \leq \tilde{\mathcal{J}}_h(u,v) + rh(\|u\|+1)(\|v\|+1)$$

for all $(u,v) \in \mathcal{U}^2 \times \mathcal{V}^2$ and all $h = 1/T$ with $T \in \mathbf{N}$.

Remarks. (i) The hypotheses on φ_t are equivalent to saying that its effective domain is a fixed set U which does not vary with t and that for a fixed $u \in U$, the function $t \mapsto \varphi_t(u)$ is Lipschitz continous on $[0,1]$ with modulus $\alpha|u|$. The analogous statement for ψ_t is also valid.

(ii) We are tacitly assuming here that the values of $\tilde{\mathcal{J}}_h$ are defined using the same conventions regarding $\infty - \infty$ that are used in defining $\hat{\mathcal{J}}_h$. If this were not the case, then we could simply replace (6) by similar inequalities involving $\overline{\mathrm{cl}}_2\, \tilde{\mathcal{J}}_h$ and $\overline{\mathrm{cl}}_2\, \hat{\mathcal{J}}_h$ (or, equivalently, $\underline{\mathrm{cl}}_1\, \tilde{\mathcal{J}}_h$ and $\underline{\mathrm{cl}}_1\, \hat{\mathcal{J}}_h$).

(iii) Note that (6) is equivalent to the inequalities given by reversing the rôles of $\tilde{\mathcal{J}}_h$ and $\hat{\mathcal{J}}_h$.

Proof of Proposition 4. It is clear that $\tilde{\mathcal{J}}_h(u,v)$ and $\hat{\mathcal{J}}_h(u,v)$ agree whenever $u \notin \mathcal{U}_h$ or $v \notin \mathcal{V}_h$. Similarly, if $(u,v) \in \mathcal{U}_h \times \mathcal{V}_h$ is identified with $(u_0^h, \dots, u_T^h; v_1^h, \dots, v_{T+1}^h)$, then $\tilde{\mathcal{J}}_h(u,v) = \hat{\mathcal{J}}_h(u,v)$ whenever one of

$$\hat{\varphi}_\tau^h(u_\tau^h),\ \hat{\psi}_\tau^h(v_\tau^h),\ \tilde{\varphi}_\tau^h(u_\tau^h), \text{ and } \tilde{\psi}_\tau^h(v_\tau^h)$$

equals $+\infty$ for some τ or when either $\varphi_e(u_0^h)$ or $\psi_e(v_{T+1}^h)$ is $+\infty$. In these cases, (6) is therefore satisfied trivially, regardless of h or r.

Now suppose that $\tilde{\varphi}_\tau(u_\tau^h)$, $\hat{\varphi}_\tau(u_\tau^h)$, $\tilde{\psi}_\tau(v_\tau^h)$, and $\hat{\psi}_\tau(v_\tau^h)$ are finite for each $\tau = 1, \dots, T$ and that $\varphi_e(u_0^h)$ and $\psi_e(v_{T+1}^h)$ are also finite. In this case we may rewrite (6) as

$$|\hat{\mathcal{J}}_h(u_0^h, \dots, u_T^h; v_1^h, \dots, v_{T+1}^h) - \tilde{\mathcal{J}}_h(u_0^h, \dots, u_T^h; v_1^h, \dots, v_{T+1}^h)| \leq rh(\|u\|+1)(\|v\|+1)$$

since both $\hat{\mathcal{J}}_h(u_0^h,\dots,u_T^h;v_1^h,\dots,v_{T+1}^h)$ and $\tilde{\mathcal{J}}_h(u_0^h,\dots,u_T^h;v_1^h,\dots,v_{T+1}^h)$ are finite. We then have

$$\begin{aligned}
&|\hat{\mathcal{J}}_h(u_0^h,\dots,u_T^h;v_1^h,\dots,v_{T+1}^h) - \tilde{\mathcal{J}}_h(u_0^h,\dots,u_T^h;v_1^h,\dots,v_{T+1}^h)| \\
&\quad\le \left|\sum_{\tau=1}^{T}\left[\hat{\varphi}_\tau^h(u_\tau^h) - \tilde{\varphi}_\tau^h(u_\tau^h) + \tilde{\psi}_\tau^h(v_\tau^h) - \hat{\psi}_\tau^h(v_\tau^h)\right]\right| \\
&\qquad\qquad + \left|\sum_{\tau=1}^{T}\left[(\hat{p}_\tau^h - \tilde{p}_\tau^h)\cdot u_\tau^h + (\hat{q}_\tau^h - \tilde{q}_\tau^h)\cdot v_\tau^h + v_\tau^h\cdot(\tilde{D}_\tau^h u_\tau^h - \hat{D}_\tau^h)u_\tau^h + \tilde{d}_\tau^h\right]\right| \\
&\qquad\qquad\quad + |\tilde{\gamma}_h(u_0^h,\dots,u_T^h;v_1^h,\dots,v_{T+1}^h) - \hat{\gamma}_h(u_0^h,\dots,u_T^h;v_1^h,\dots,v_{T+1}^h)|.
\end{aligned}$$

Thus, to prove (6) it suffices to find r_1, r_2, and r_3 so that the following three inequalities are satisfied:

$$\left|\sum_{\tau=1}^{T}\left[\hat{\varphi}_\tau^h(u_\tau^h) - \tilde{\varphi}_\tau^h(u_\tau^h)\right]\right| + \left|\sum_{\tau=1}^{T}\left[\tilde{\psi}_\tau^h(v_\tau^h) - \hat{\psi}_\tau^h(v_\tau^h)\right]\right| \le r_1 h(\|u\| + \|v\|),$$

$$\begin{aligned}
&\left|\sum_{\tau=1}^{T}\left[(\hat{p}_\tau^h - \tilde{p}_\tau^h)\cdot u_\tau^h + (\hat{q}_\tau^h - \tilde{q}_\tau^h)\cdot v_\tau^h + v_\tau^h\cdot(\tilde{D}_\tau^h u_\tau^h - \hat{D}_\tau^h)u_\tau^h + \tilde{d}_\tau^h\right]\right| \\
&\qquad\qquad\qquad\qquad\le r_2 h(\|u\| + 1)(\|v\| + 1),
\end{aligned}$$

$$|\tilde{\gamma}_h(u_0^h,\dots,u_T^h;v_1^h,\dots,v_{T+1}^h) - \hat{\gamma}_h(u_0^h,\dots,u_T^h;v_1^h,\dots,v_{T+1}^h)| \le r_3 h(\|u\| + 1)(\|v\| + 1).$$

In what follows we sketch how to obtain values for these r_i. For a Lipschitz-continuous matrix-valued function ξ on $[0,1]$, let $\mathrm{Lip}(\xi)$ denote the Lipschitz constant for ξ and define $\|\xi\|_\infty = \sup_{t\in[0,1]}|\xi_t|$. It is easily verified that there exists a constant $\beta > 0$ (which depends only on the function A_t) such that the following estimates hold for all $\tau = 1,\dots,T$:

$$\left|\int_{(\tau-1)h}^{\tau h}\xi_t\,dt - h\xi_{(\tau-1)h}\right| \le \frac{h^2}{2}\mathrm{Lip}(\xi), \tag{7}$$

$$\left|\int_{(\tau-1)h}^{\tau h}\left(\xi_t\mathcal{A}_t\int_{(\tau-1)h}^{t}\mathcal{A}_s^{-1}\eta_s\,ds\right)\,dt\right| \le h^2\beta\|\xi\|_\infty\|\eta\|_\infty, \tag{8}$$

$$\left|\int_{(\tau-1)h}^{\tau h}\mathcal{A}_{\tau h}\mathcal{A}_t^{-1}\xi_t\,dt - h\xi_{(\tau-1)h}\right| \le h^2\beta\cdot[\|\xi\|_\infty + \mathrm{Lip}(\xi)], \tag{9}$$

$$\left|\int_{(\tau-1)h}^{\tau h}\xi_t\mathcal{A}_t\mathcal{A}_{(\tau-1)h}^{-1}\,dt - h\xi_{(\tau-1)h}\right| \le h^2\beta\cdot[\|\xi\|_\infty + \mathrm{Lip}(\xi)]. \tag{10}$$

Now, to find r_1 we use the hypotheses on φ_t and ψ_t (via Remark (i) above) and apply inequality (7) with $\xi_t = \varphi_t(u_\tau^h)$ and $\xi_t = \psi_t(v_\tau^h)$. This yields $r_1 = \alpha/2$.

Next we deal with r_2. Note that we have

$$\left|\sum_{\tau=1}^{T}(\hat{p}_\tau^h - \tilde{p}_\tau^h)\cdot u_\tau^h\right| \le \left(\max_\tau|\hat{p}_\tau^h - \tilde{p}_\tau^h|\right)\left(\sum_{\tau=1}^{T}|u_\tau^h|\right) \le \frac{1}{h}\left(\max_\tau|\hat{p}_\tau^h - \tilde{p}_\tau^h|\right)\|u^h\|_h.$$

By applying inequality (7) with $\xi_t = p_t$ and then inequality (8) with $\xi_t = c_t$ and $\eta_s = B_s$ we see that, for each τ,

$$|\hat{p}_\tau^h - \tilde{p}_\tau^h| \le \frac{h^2}{2}\mathrm{Lip}(p) + h^2\beta\|c\|_\infty\|B\|_\infty.$$

Thus we obtain the bound

$$\left|\sum_{\tau=1}^{T}(\hat{p}_\tau^h - \tilde{p}_\tau^h)\cdot u_\tau^h\right| \le h\left[\frac{1}{2}\mathrm{Lip}(p) + \beta\|c\|_\infty\|B\|_\infty\right]\|u^h\|_h.$$

By similar arguments, we find that

$$\left|\sum_{\tau=1}^{T}(\hat{q}_\tau^h - \tilde{q}_\tau^h)\cdot v_\tau^h\right| \le h\left[\frac{1}{2}\mathrm{Lip}(q) + \beta\|b\|_\infty\|C\|_\infty\right]\|v^h\|_h,$$

$$\left|\sum_{\tau=1}^{T}\tilde{d}_\tau^h\right| \le h\beta\|b\|_\infty\|c\|_\infty,$$

$$\left|\sum_{\tau=1}^{T}v_\tau^h\cdot(\hat{D}_\tau^h - \tilde{D}_\tau^h)u_\tau^h\right| \le h\left[\frac{1}{2}\mathrm{Lip}(D) + \beta\|B\|_\infty\|C\|_\infty\right]\|u^h\|_h\|v^h\|_h.$$

Therefore we can take r_2 to be given by

$$r_2 = \max\{\mathrm{Lip}(p), \mathrm{Lip}(q), \mathrm{Lip}(D)\} + \beta\cdot\max\{\|b\|_\infty, \|B\|_\infty, \|c\|_\infty, \|C\|_\infty\}^2.$$

Finally we turn to finding the coefficient r_3. First we see that

$$\begin{aligned}
&|\tilde{\gamma}_h(u_0^h,\dots,u_T^h;v_1^h,\dots,v_{T+1}^h) - \hat{\gamma}_h(u_0^h,\dots,u_T^h;v_1^h,\dots,v_{T+1}^h)| \\
&\qquad\le \left|\sum_{\tau=1}^{T}v_\tau^h\cdot(\tilde{C}_\tau^h\tilde{x}_{\tau-1}^h - \hat{C}_\tau^h\hat{x}_{\tau-1}^h)\right| + \left|v_{T+1}^h\cdot C_e(\tilde{x}_T^h - \hat{x}_T^h)\right| \\
&\qquad\quad + \left|\sum_{\tau=1}^{T}\tilde{c}_\tau^h\cdot\tilde{x}_{\tau-1}^h - \hat{c}_\tau^h\cdot\hat{x}_{\tau-1}^h\right| + \left|c_e\cdot(\tilde{x}_T^h - \hat{x}_T^h)\right| \\
&\qquad\le \frac{1}{h}\|v^h\|_h\cdot\max_\tau\left|\tilde{C}_\tau^h\tilde{x}_{\tau-1}^h - \hat{C}_\tau^h\hat{x}_{\tau-1}^h\right| + \|v^h\|_h\cdot\left|C_e(\tilde{x}_T^h - \hat{x}_T^h)\right| \\
&\qquad\quad + \frac{1}{h}\|v^h\|_h\cdot\max_\tau\left|\tilde{c}_\tau^h\cdot\tilde{x}_{\tau-1}^h - \hat{c}_\tau^h\cdot\hat{x}_{\tau-1}^h\right| + \|v^h\|_h\cdot\left|c_e\cdot(\tilde{x}_T^h - \hat{x}_T^h)\right|.
\end{aligned}\tag{11}$$

Observing that

$$\left|\tilde{C}_\tau^h\tilde{x}_{\tau-1}^h - \hat{C}_\tau^h\hat{x}_{\tau-1}^h\right| \le \left|\tilde{C}_\tau^h - \hat{C}_\tau^h\right|\left|\tilde{x}_{\tau-1}^h\right| + \left|\tilde{x}_{\tau-1}^h - \hat{x}_{\tau-1}^h\right|\left|\hat{C}_\tau^h\right| \tag{12}$$

and

$$\left|\tilde{c}_\tau^h\cdot\tilde{x}_{\tau-1}^h - \hat{c}_\tau^h\cdot\hat{x}_{\tau-1}^h\right| \le \left|\tilde{c}_\tau^h - \hat{c}_\tau^h\right|\left|\tilde{x}_{\tau-1}^h\right| + \left|\tilde{x}_{\tau-1}^h - \hat{x}_{\tau-1}^h\right|\left|\hat{c}_\tau^h\right|, \tag{13}$$

we see that we need upper bounds, in terms of h and $\|u^h\|_h$, for the quantities

$$\left|\tilde{C}_\tau^h - \hat{C}_\tau^h\right|,\ \left|\tilde{x}_\tau^h\right|,\ \left|\tilde{x}_\tau^h - \hat{x}_\tau^h\right|,\ \left|\hat{C}_\tau^h\right|,\ \left|\tilde{c}_\tau^h - \hat{c}_\tau^h\right|,\ \text{and}\ \left|\hat{c}_\tau^h\right|$$

for each τ. By using the expansions

$$\hat{x}_\tau^h = \left[(\hat{A}_\tau^h\cdots\hat{A}_1^h)(B_e u_0^h + b_e)\right] + \sum_{\tau'=1}^{\tau}\left[(\hat{A}_\tau^h\cdots\hat{A}_{\tau'+1}^h)(\hat{B}_{\tau'}^h u_{\tau'}^h + \hat{b}_{\tau'}^h)\right],$$

$$\tilde{x}_\tau^h = \left[(\tilde{A}_\tau^h\cdots\tilde{A}_1^h)(B_e u_0^h + b_e)\right] + \sum_{\tau'=1}^{\tau}\left[(\tilde{A}_\tau^h\cdots\tilde{A}_{\tau'+1}^h)(\tilde{B}_{\tau'}^h u_{\tau'}^h + \tilde{b}_{\tau'}^h)\right]$$

and continuing with repeated use of inequalities (8)–(10) we can obtain

$$\left|\hat{C}_T^h\right| \le h\|C\|_\infty, \qquad \left|\hat{c}_T^h\right| \le h\|c\|_\infty,$$
$$\left|\tilde{C}_T^h - \hat{C}_T^h\right| \le h^2\beta[\|C\|_\infty + \mathrm{Lip}(C)], \qquad \left|\tilde{c}_T^h - \hat{c}_T^h\right| \le h^2\beta[\|c\|_\infty + \mathrm{Lip}(c)],$$
$$\left|\tilde{x}_T^h\right| \le e^{2\|A\|_\infty} \cdot \left[(\|B\|_\infty + |B_e|)\|u^h\|_h + (\|b\|_\infty + |b_e|)\right],$$
$$\left|\tilde{x}_T^h - \hat{x}_T^h\right| \le h\beta e^{2\|A\|_\infty} \cdot \left[\rho(A, B, B_e)\|u^h\|_h + \rho(A, b, b_e)\right],$$

where

$$\rho(A, B, B_e) = \Big(\|B\|_\infty + |B_e|\Big)\Big(\|A\|_\infty + \mathrm{Lip}(A)\Big) + \Big(\|B\|_\infty + \mathrm{Lip}(B)\Big)$$
$$\rho(A, b, b_e) = \Big(\|b\|_\infty + |b_e|\Big)\Big(\|A\|_\infty + \mathrm{Lip}(A)\Big) + \Big(\|b\|_\infty + \mathrm{Lip}(b)\Big).$$

These inequalities supply us with an upper bound for (11) via (12) and (13), yielding a value for r_3. □

The proposition just proved tells us, in particular, that the nets $\{\tilde{\mathcal{J}}_h\}$ and $\{\hat{\mathcal{J}}_h\}$ are "uniformly cofinal" on bounded sets: for a fixed $\rho > 0$ and $r_\rho = r(\rho + 1)^2$, we have

$$\tilde{\mathcal{J}}_h(u, v) - r_\rho h \le \hat{\mathcal{J}}_h(u, v) \le \tilde{\mathcal{J}}_h(u, v) + r_\rho h$$

whenever $\|u\| \le \rho$ and $\|v\| \le \rho$. The uniformity in (6) actually guarantees that the problems $(\tilde{\mathcal{S}}_h)$ and $(\hat{\mathcal{S}}_h)$ are close from a variational standpoint. This is illustrated by the next theorem.

THEOREM 6. *Suppose $\{T_m\}$ is an increasing sequence of positive integers and let $h_m = 1/T_m$. Consider a proper closed saddle function $\mathcal{K}$ defined on $\mathcal{U}^2 \times \mathcal{V}^2$. Under the hypotheses of Proposition 4, the sequence $\{\hat{\mathcal{J}}_{h_m}\}$ epi/hypo-converges to $\mathcal{K}$ if and only if $\{\tilde{\mathcal{J}}_{h_m}\}$ epi/hypo-converges to $\mathcal{K}$.*

Proof. This follows from the definition of epi/hypo-convergence by combining inequality (6) with the fact that weakly convergent sequences are norm-bounded. □

Combined with Theorem 5, the above result gives us sufficient conditions for the problems $\{(\hat{\mathcal{S}}_h)\}$ to be variational approximations to $(\mathcal{S})$.

We close with a few remarks concerning the advantages and disadvantages of the two discretization schemes considered in this paper. The main consistency result (Theorem 5) applies to the basic discretization scheme given by restricting the controls to belong to the class of step functions over a predetermined partition of the time interval. Since the hypotheses for Theorem 5 are very mild, it is applicable to a wide class of problems. The dicretization entails the calculation of various integrals involving the fundamental matrix solution of the linear dynamics. In the autonomous case, where the data are constant with respect to time, these integrals can be expressed in terms of the power series representation of the matrix exponential. In practice, an adequate approximation for the series requires a study of the eigensystems of the matrix A. For general nonautonomous systems, the fundamental matrix solution is unlikely to be available in closed form, so further approximation is needed. The integrals defining the data of problem $(\tilde{\mathcal{P}})$ of §3 must then be calculated with the approximate matrix solution. It considerably simplifies the computations if instead we use the finite-difference scheme given in the current section. The primary drawback is that this is only a first-order scheme. If the data is sufficiently smooth, then computationally efficient higher-order methods should be employed. (Analogously, in the basic discretization scheme we could replace step functions by higher-order piecewise polynomial functions.) We are led then to questions concerning the *rate of convergence* for various discretization schemes and their efficient implementation. These issues will need to be addressed in future work.

REFERENCES

[1] H. ATTOUCH, *Variational Convergence for Functions and Operators*, Pitman, London, 1984.

[2] H. ATTOUCH AND R. J.-B. WETS, *A convergence for bivariate functions aimed at the convergence of saddle values*, in Mathematical Theory of Optimization, J. P. Cecconi and T. Zolezzi, eds., Lecture Notes in Mathematics 979, Springer-Verlag, Berlin, 1981, pp. 1–42.

[3] ———, *A convergence theory for saddle functions*, Trans. Amer. Math. Soc., 280 (1983), pp. 1–41.

[4] H. ATTOUCH, D. AZÉ, AND R. J.-B. WETS, *Convergence of convex-concave saddle functions: applications to convex programming and mechanics*, Ann. Inst. H. Poincaré, Anal. Non Linéaire, 5 (1988), pp. 537–572.

[5] H. ATTOUCH AND R. J.-B. WETS, *Quantitative stability of variational systems: I. The epigraphical distance*, Trans. Amer. Math. Soc., 328 (1991), pp. 695–729.

[6] W. E. BOSARGE, JR. AND O. G. JOHNSON, *Error bounds of high order accuracy for the state regulator problem via piecewise polynomial approximations*, SIAM J. Control, 9 (1971), pp. 15–28.

[7] G. CHEN AND W. H. MILLS, *Finite elements and terminal penalization for quadratic cost optimal control problems governed by ordinary differential equations*, SIAM J. Control Optim., 19 (1981), pp. 744–764.

[8] J. CULLUM, *Penalty functions and nonconvex continuous optimal control problems*, in Computing Methods in Optimization Problems 2, L. A. Zadeh, L. W. Neustadt, and A. V. Balakrishnan, eds., Academic Press, New York, 1969, pp. 55–67.

[9] ———, *Discrete approximations to continuous optimal control problems*, SIAM J. Control, 7 (1969), pp. 32–49.

[10] ———, *An explicit procedure for discretizing continuous optimal control problems*, J. Optim. Theory Appl., 8 (1971), pp. 15–34.

[11] J. W. DANIEL, *The Ritz-Galerkin Method for abstract optimal control problems*, SIAM J. Control, 11 (1973), pp. 53–63.

[12] A. DONTCHEV AND W. W. HAGER, *Lipschitzian stability in nonlinear control and optimization*, SIAM J. Control Optim., 31 (1993), pp. 569–603.

[13] W. W. HAGER, *The Ritz-Trefftz method for state and control constrained optimal control problems*, SIAM J. Numer. Anal., 12 (1975), pp. 854–867.

[14] ———, *Rates of convergence for discrete approximations to unconstrained control problems in a finite dimensional space*, SIAM J. Numer. Anal., 13 (1976), pp. 449–472.

[15] W. W. HAGER AND G. D. IANCULESCU, *Dual approximations in optimal control*, SIAM J. Control Optim., 22 (1984), pp. 423–465.

[16] B. S. MORDUKHOVICH, *Approximation Methods in Problems of Optimization and Control*, Nauka, Moscow, 1988. (In Russian.)

[17] F. H. MATHIS AND G. W. REDDIEN, *Ritz-Trefftz approximations in optimal control*, SIAM J. Control Optim., 17 (1979), pp. 307–310.

[18] O. PIRONNEAU AND E. POLAK, *A dual method for optimal control problems with initial and final boundary constraints*, SIAM J. Control, 11 (1973), pp. 534–549.

[19] R. T. ROCKAFELLAR, *Minimax theorems and conjugate saddle functions*, Math. Scand., 14 (1964), pp. 151–173.

[20] ———, *Convex Analysis*, Princeton University Press, Princeton, NJ, 1970.

[21] ———, *Linear-quadratic programming and optimal control*, SIAM J. Control Optim., 25 (1987), pp. 781–814.

[22] ———, *Hamiltonian trajectories and duality in the optimal control of linear systems with convex costs*, SIAM J. Control Optim., 27 (1989), pp. 1007–1025.

[23] ———, *Generalized second derivatives of convex functions and saddle functions*, Trans. Amer. Math. Soc., 320 (1990), pp. 810–822.

[24] R. T. Rockafellar and R. J.-B. Wets, *A dual solution procedure for quadratic stochastic programs with simple recourse*, in Numerical Methods, V. Pereyra and A. Reinoza, eds., Lecture Notes in Mathematics 1005, Springer-Verlag, Berlin, 1983, pp. 252–265.

[25] ———, *A Lagrangian finite generation technique for solving linear-quadratic problems in stochastic programming*, Math. Programming Stud., 28 (1986), pp. 63–93.

[26] ———, *Linear-quadratic programming problems with stochastic penalties: The finite generation algorithm*, in Numerical Techniques for Stochastic Optimization Problems, Y. Ermoliev and R.J.-B. Wets, eds., Springer-Verlag, Berlin, New York, 1986.

[27] D. L. Russell, *Penalty functions and bounded phase coordinate control*, SIAM J. Control, 2 (1965), pp. 409–422.

[28] V. Veliov, *Second order discrete approximation to linear differential inclusions*, SIAM J. Numer. Anal., 29 (1992), pp. 439–451.

[29] S. E. Wright, *DYNFGM: A computer program for solving extended linear-quadratic problems in optimal control by saddle point generation*, Technical report, Deparment of Mathematics, University of Washington, Seattle, WA, March 1989.

[30] ———, *Convergence and Approximation for Primal-Dual Methods in Large-Scale Optimization*, Ph.D. thesis, University of Washington, Seattle, WA, December 1990.

[31] C. Zhu and R. T. Rockafellar, *Primal-dual projected gradient algorithms for extended linear-quadratic programming*, SIAM J. Optim., (1993), pp. 751–783.

SIAM J. CONTROL OPTIM.
Vol. 33, No. 5, pp. 1510–1529, September 1995

© 1995 Society for Industrial and Applied Mathematics

010

ERROR BOUNDS FOR PIECEWISE CONVEX QUADRATIC PROGRAMS AND APPLICATIONS*

WU LI[†]

Abstract. In this paper, we establish a local error estimate for feasible solutions of a piecewise convex quadratic program and a global error estimate for feasible solutions of a convex piecewise quadratic program. These error estimates provide a unified approach for deriving many old and new error estimates for linear programs, linear complementarity problems, convex quadratic programs, and affine variational inequality problems. The approach reveals the fact that each error estimate is a consequence of some reformulation of the original problem as a piecewise convex quadratic program or a convex piecewise quadratic program. In a sense, even Robinson's result on the upper Lipschitz continuity of a polyhedral mapping can be considered as a special case of error estimates for approximate solutions of a piecewise convex quadratic program. As an application, we derive new (global) error estimates for iterates of the proximal point algorithm for solving a convex piecewise quadratic program.

Key words. local error bound, global error bound, piecewise convex quadratic program, convex quadratic program, monotone linear complementarity problem, affine variational inequality problem, proximal point algorithm, rate of convergence

AMS subject classifications. Primary, 90C31; Secondary, 90C20, 90C33

1. Introduction.

In a series of papers by Luo and Tseng [18]–[22], local error estimates for approximate solutions of a constrained convex minimization problem were proven to be crucial in convergence analysis of descent (or dual ascent) methods for solving constrained convex minimization problems (cf. also [26], [12], [14]). Meanwhile, Ferris [4] showed that the weak sharp minimum property of the solution set of a constrained convex minimization problem is sufficient for the finite convergence of the proximal point algorithm for solving the constrained convex minimization problem. As a consequence, Ferris established the finite convergence of the proximal point algorithm for solving a linear program, since the solution set of a linear program has the weak sharp minimum property [27]. The common feature of an error estimate for approximate solutions and the weak sharp minimum property is the estimation of the distance from an approximate solution to the solution set of the constrained convex minimization problem. In this paper, we give error estimates for feasible solutions of a piecewise convex quadratic program and show some applications of such estimates. In particular, this paper provides a unique perspective to the relationship between Mangasarian and Meyer's weak sharp minimum property of the solution(s) of a linear program [27] and Luo and Tseng's local error estimate for approximate solutions of an affine variational inequality problem [21]. As applications of error estimates for feasible solutions of a convex piecewise quadratic program, we obtain a new global error estimate for approximate solutions of a monotone linear complementarity problem and a new global error estimate for iterates generated by the proximal point algorithm for solving a convex piecewise quadratic program.

Our research on piecewise convex quadratic programs is motivated by a reformulation of certain convex quadratic programs as the problem of minimizing a convex

*Received by the editors January 19, 1993; accepted for publication (in revised form) May 3, 1994.

[†]Department of Mathematics and Statistics, Old Dominion University, Norfolk, Virginia 23529 (wuli@math.odu.edu).

quadratic spline without any constraint [15], [16]. A Newton method with exact line minimization seems very effective for finding the minimizer of a strictly convex quadratic spline [15]. Therefore, it is natural to consider the proximal point algorithm for finding a minimizer of a convex quadratic spline, since one only needs to find the minimizer of a strictly convex quadratic spline in each iteration. Also it is possible to design new iterative algorithms for finding a minimizer of a convex quadratic spline [16]. The error estimates are useful for studying the convergence behavior of new iterative algorithms for solving the unconstrained minimization of a convex quadratic spline (cf. [16]). Moreover, convex piecewise quadratic programs are related to a linear-quadratic minimax problem studied by Rockafellar and Wets in their research on stochastic programming and optimal control problems [34]–[36] (also cf. [37] and some references there).

In §2, we establish local error estimates for feasible solutions of piecewise convex quadratic programs and global error estimates of convex piecewise quadratic programs. In §3, we shows that most existing error estimates for linear programs, linear complementarity problems, and affine variational inequality problems can be recovered easily as consequences of error estimates given in §2. A new global error estimate for monotone linear complementarity problems is derived by using error estimates for convex piecewise quadratic programs. Section 4 is devoted to new error estimates of iterates of the proximal point algorithm, which reveals a new perspective on the relationship between the convergence rate of the proximal point algorithm and the error estimate for feasible solutions. In §5 we mention a few possible extensions of the results to general mathematical programming problems and show how one can view Robinson's result on the upper Lipschitz continuity of a polyhedral mapping as a special case of an error estimate for approximate solutions of a piecewise convex quadratic program.

Now let us give some common assumptions and notations used in the paper. Consider the following minimization problem:

$$f_{\min} := \min_{x \in X} f(x), \tag{1.1}$$

where X is a closed convex polyhedral subset of the n-dimensional Euclidean space $\mathbb{R}^n$. We assume that $f_{\min} > -\infty$ and the solution set of (1.1) is not empty; i.e.,

$$X^* := \{x \in X : f(x) = f_{\min}\} \neq \emptyset.$$

The 2-norm on $\mathbb{R}^n$ is denoted by $\|x\| := \left(\sum_{i=1}^n x_i^2\right)^{1/2}$. The distance from a vector x to the solution set X^* of (1.1) is defined as

$$\operatorname{dist}(x, X^*) := \min\{\|x - x^*\| : x^* \in X^*\}.$$

The transpose of a vector x (or a matrix M) is denoted by x^T (or M^T). For two vectors $x, y \in \mathbb{R}^n$, we write $x \leq y$ if $x_i \leq y_i$ for $1 \leq i \leq n$. A continuous function f on $\mathbb{R}^n$ is called a piecewise (convex) quadratic function if there are finitely many convex polyhedral subsets $\{C_i\}_{i=1}^m$ of $\mathbb{R}^n$ such that $\mathbb{R}^n = \bigcup_{i=1}^m C_i$ and f is a (convex) quadratic function on each C_i. A piecewise convex quadratic function f is called a convex piecewise quadratic function if f is also a convex function. A quadratic spline is a differentiable piecewise quadratic function and a convex quadratic spline is a differentiable convex piecewise quadratic function. We use $f'(x)$ to denote the gradient of $f(x)$ as a column vector.

2. Error bounds for piecewise convex quadratic programs. In this section we give local and global error estimates of $\operatorname{dist}(x, X^*)$ in terms of $(f(x) - f_{\min})$. We first establish error estimates for convex quadratic programs and then extend the results for piecewise convex quadratic programs by using Frank–Wolfe's theorem on solvability of a quadratic program. The main results are Theorems 2.5, 2.6, and 2.7. However, Lemma 2.3, based on the following geometric feature of a polyhedral set and Mangasarian's characterization of the solutions of a convex quadratic program, is crucial for establishing all error estimates in this section.

LEMMA 2.1. *Suppose that $x^k, y^k \in X$ with $x^k \neq y^k$ and*

$$\lim_{k\to\infty} \frac{x^k - y^k}{\|x^k - y^k\|} = z.$$

Then there exist positive constants α_k such that $z^k := y^k + \alpha_k z \in X$ and

$$\lim_{k\to\infty} \frac{\|x^k - z^k\|}{\|x^k - y^k\|} = 0.$$

Proof. Suppose that $X := \{x \in \mathbb{R}^n : Ax \geq b\}$. Let $0 < \lambda < 1$ and $w^k := y^k + \lambda\|x^k - y^k\|z$. Obviously, $(Aw^k)_i \geq (Ay^k)_i \geq b_i$ for $(Az)_i \geq 0$. If $(Az)_i < 0$, then, for k large enough,

$$(Aw^k)_i = (Ax^k)_i + \|x^k - y^k\| \left(A\left(\lambda z - \frac{x^k - y^k}{\|x^k - y^k\|}\right)\right)_i \geq b_i, \tag{2.1}$$

since

$$\lim_{k\to\infty} \left(A\left(\lambda z - \frac{x^k - y^k}{\|x^k - y^k\|}\right)\right)_i = (\lambda - 1)(Az)_i > 0.$$

Since there are only finitely many indices, there exists $k_{\lambda,1} \geq 1$ such that (2.1) holds whenever $(Az)_i < 0$ and $k \geq k_{\lambda,1}$. That is, $w^k \in X$ if $k \geq k_{\lambda,1}$.

On the other hand, since

$$\lim_{k\to\infty} \left\| \frac{x^k - y^k}{\|x^k - y^k\|} - \lambda z \right\| = (1 - \lambda)\|z\| = 1 - \lambda,$$

we obtain that there exists $k_{\lambda,2} \geq 1$ such that, for $k \geq k_{\lambda,2}$,

$$\|x^k - w^k\| = \|x^k - y^k\| \cdot \left\| \frac{x^k - y^k}{\|x^k - y^k\|} - \lambda z \right\| \leq 2(1 - \lambda)\|x^k - y^k\|. \tag{2.2}$$

Set $k_\lambda := \max\{k_{\lambda,1}, k_{\lambda,2}\}$. Then (2.1) and (2.2) hold when $k \geq k_\lambda$.

Let $\lambda_s = 1 - \frac{1}{s}$ for $s = 1, 2\ldots$. Then there exist k_{λ_s} such that (2.1) and (2.2) hold with $\lambda = \lambda_s$ when $k \geq k_{\lambda_s}$. We may assume $k_{\lambda_{s+1}} > k_{\lambda_s}$ for $s = 1, 2, \ldots$. Define $\alpha_k = \lambda_s\|x^k - y^k\|$ and $z^k = y^k + \alpha_k z$ for $k_{\lambda_s} \leq k < k_{\lambda_{s+1}}$ and $s = 1, 2, \ldots$. Then $z^k \in X$ and

$$\frac{\|x^k - z^k\|}{\|x^k - y^k\|} \leq 2(1 - \lambda_s) = \frac{2}{s} \quad \text{for } k \geq k_{\lambda_s}, \;\; s = 1, 2, \ldots. \qquad \square$$

To prove the most important lemma in this section, we need the following characterization of the solutions of a convex quadratic program by Mangasarian [25].

LEMMA 2.2. *Suppose that $f(x) := \frac{1}{2}x^TBx + c^Tx$ is a convex quadratic function, x^* in X^*, and $\bar{x} \in X$. Then the following statements are equivalent:*

(1) $\bar{x} \in X^*$;

(2) $f'(x^*)^T(\bar{x} - x^*) \le 0$ *and* $(\bar{x} - x^*)^TB(\bar{x} - x^*) \le 0$;

(3) $f'(\bar{x}) = f'(x^*)$ *and* $B(\bar{x} - x^*) = 0$;

(4) $f'(\bar{x})^T(x - \bar{x}) \ge 0$ *for any* $x \in X$.

LEMMA 2.3. *Suppose that $f(x)$ is a convex quadratic function. Then, for any constant $\delta > 0$, there exists a positive constant γ such that*

$$\operatorname{dist}(x, X^*) \le \gamma\sqrt{f(x) - f_{\min}} \quad \textit{for } x \in X \text{ with } f(x) - f_{\min} \le \delta. \tag{2.3}$$

Proof. Let $x \in X$ and $\bar{x}, x^* \in X^*$. Then

$$\begin{aligned} f(x) - f(\bar{x}) &= f'(\bar{x})^T(x - \bar{x}) + \frac{1}{2}(x - \bar{x})^TB(x - \bar{x}) \\ &\ge \max\left\{f'(\bar{x})^T(x - \bar{x}), \frac{1}{2}(x - \bar{x})^TB(x - \bar{x})\right\} \\ &= \max\left\{f'(x^*)^T(x - \bar{x}), \frac{1}{2}(x - \bar{x})^TB(x - \bar{x})\right\}, \end{aligned} \tag{2.4}$$

where the first equality is the Taylor expansion of f at $\bar{x}$, the inequality derived from $f'(\bar{x})^T(x - \bar{x}) \ge 0$ (cf. Lemma 2.2 (4)) and $(x - \bar{x})^TB(x - \bar{x}) \ge 0$, and the second equality follows from Lemma 2.2 (3).

Assume the contrary, that there exists a sequence $\{x^k\} \subset X$ such that $f(x) - f_{\min} \le \delta$ and

$$\lim_{k\to\infty} \frac{\sqrt{f(x^k) - f_{\min}}}{\operatorname{dist}(x^k, X^*)} = 0. \tag{2.5}$$

Let $\bar{x}^k \in X^*$ be such that $\|x^k - \bar{x}^k\| = \operatorname{dist}(x^k, X^*)$ and set $z^k := \frac{x^k - \bar{x}^k}{\|x^k - \bar{x}^k\|}$. Then, by (2.4),

$$\frac{f(x^k) - f(\bar{x}^k)}{\|x^k - \bar{x}^k\|^2} \ge \frac{1}{2}(z^k)^TBz^k. \tag{2.6}$$

Without loss of generality, we may assume that $z^k \to z$ as $k \to \infty$. Then it follows from (2.5) and (2.6) that $z^TBz \le 0$. If $\operatorname{dist}(x^k, X^*) \le \delta$ for infinitely many k's, then, by selecting a subsequence, we may assume that $\operatorname{dist}(x^k, X^*) \le \delta$ for all k. In this case, by (2.4),

$$\frac{f(x^k) - f(\bar{x}^k)}{\|x^k - \bar{x}^k\|^2} \ge \frac{1}{\|x^k - \bar{x}^k\|}f'(x^*)^Tz^k \ge \frac{1}{\delta}f'(x^*)^Tz^k. \tag{2.7}$$

By (2.5) and (2.7), we obtain that $f'(x^*)^Tz \le 0$. If $\operatorname{dist}(x^k, X^*) \le \delta$ for finitely many k's, then, by selecting a subsequence, we may assume that $\operatorname{dist}(x^k, X^*) > \delta > 0$ for all k. In this case, since $f(x) - f_{\min} \le \delta$, (2.5) is equivalent to

$$\lim_{k\to\infty} \frac{f(x^k) - f(\bar{x}^k)}{\|x^k - \bar{x}^k\|} = 0. \tag{2.8}$$

By (2.4),

$$\frac{f(x^k)-f(\bar{x}^k)}{\|x^k-\bar{x}^k\|} \geq f'(x^*)^T z^k. \tag{2.9}$$

Again, we get $f'(x^*)^T z \leq 0$ from (2.8) and (2.9).

On the other hand, by Lemma 2.1, there exist positive constants α_k such that $y^k := x^k + \alpha_k z \in X$ and

$$\lim_{k\to\infty} \frac{\|x^k-y^k\|}{\|x^k-\bar{x}^k\|} = 0. \tag{2.10}$$

Since $\|x^k-\bar{x}^k\| = \operatorname{dist}(x^k, X^*)$, (2.10) implies $y^k \notin X^*$ when k is large enough. It follows from Lemma 2.2 (2) and (3) that either

$$f'(x^*)^T z = \frac{1}{\alpha_k} f'(\bar{x}^k)^T (y^k - \bar{x}^k) > 0$$

or

$$z^T B z = \frac{1}{\alpha_k^2}(y^k-\bar{x}^k)^T B (y^k - \bar{x}^k) > 0,$$

a contradiction to $f'(x^*)^T z \leq 0$ and $z^T B z \leq 0$. □

Remark. In contrast to (2.3), Ferris and Mangasarian [5] performed a detailed analysis on characterization conditions which guarantee the global error estimate $\operatorname{dist}(x, X^*) \leq \gamma(f(x) - f_{\min})$ for $x \in X$.

Now, to extend the above estimate for piecewise convex quadratic programs, we need the following theorem by Frank and Wolfe about the solvability of a quadratic program [6].

LEMMA 2.4. *Suppose that f is a quadratic function. If $f_{\min} > -\infty$, then $X^* \neq \emptyset$.*

Using Lemma 2.4 we can easily show (2.3) still holds if f is a piecewise convex quadratic function.

THEOREM 2.5. *If f is a piecewise convex quadratic function, then there exist positive constants δ and γ such that*

$$\operatorname{dist}(x, X^*) \leq \gamma\sqrt{f(x) - f_{\min}} \quad \text{for } x \in X \text{ with } f(x) - f_{\min} \leq \delta. \tag{2.11}$$

Proof. Let $X = \bigcup_{i=1}^{s} X_i$, where X_i's are closed convex polyhedral subsets of X such that f is a convex quadratic function on every X_i. Define $\delta := \min\left\{\frac{\delta_i}{2} : \delta_i > 0, 1 \leq i \leq s\right\} > 0$, where $\delta_i := \min_{x\in X_i} f(x) - f_{\min}$. If $X_i \cap X^* \neq \emptyset$, by Lemma 2.3, there exists $\gamma_i > 0$ such that

$$\operatorname{dist}(x, X^*) \leq \operatorname{dist}(x, X^* \cap X_i) \leq \gamma_i \sqrt{f(x) - f_{\min}} \quad \text{for } x \in X_i \text{ with } f(x) - f_{\min} \leq \delta. \tag{2.12}$$

Let $\gamma := \max\{\gamma_i : X_i \cap X^* \neq \emptyset\}$. By the definition of δ, if $x \in X$ with $f(x) - f_{\min} \leq \delta$, then $x \in X_i$ with $\min_{x\in X_i} f(x) = f_{\min}$. By Lemma 2.4, $X^* \cap X_i \neq \emptyset$. Therefore, (2.12) holds, which implies (2.11). □

The estimate (2.11) is called a local error estimate in the sense that the estimate (2.11) only holds for x near the solution set X^* (cf. Corollary 2.9). Next we give an error estimate for feasible solutions away from the solution set X^* (cf. Corollary 2.10).

THEOREM 2.6. *Suppose that f is a convex piecewise quadratic function. Then, for any constant $\delta > 0$, there exists a positive constant γ such that*

$$\operatorname{dist}(x, X^*) \le \gamma\,(f(x) - f_{\min}) \quad \textit{for } x \in X \textit{ with } f(x) - f_{\min} \ge \delta. \tag{2.13}$$

Proof. Assume the contrary, that there exists a sequence $\{x^k\} \subset X$ such that $f(x) - f_{\min} \ge \delta$ and

$$\lim_{k\to\infty} \frac{f(x^k) - f(\bar{x}^k)}{\|x^k - \bar{x}^k\|} = 0, \tag{2.14}$$

where $\bar{x}^k \in X^*$ with $\|x^k - \bar{x}^k\| = \operatorname{dist}(x^k, X^*)$. We may assume $\|x^k - \bar{x}^k\| > \delta$ for all $k \ge 1$. Then $y^k := \bar{x}^k + \lambda_k(x^k - \bar{x}^k) \in X$ with $\lambda_k := \frac{\delta}{\|x^k - \bar{x}^k\|}$. For any $\bar{x} \in X^*$,

$$\|x^k - \bar{x}^k\| \le \|x^k - \bar{x}\| \le \|x^k - y^k\| + \|y^k - \bar{x}\| = (1 - \lambda_k)\|x^k - \bar{x}^k\| + \|y^k - \bar{x}\|,$$

which implies $\|y^k - \bar{x}\| \ge \delta = \|y^k - \bar{x}^k\|$ for $\bar{x} \in X^*$. Therefore, $\operatorname{dist}(y^k, X^*) = \|y^k - \bar{x}^k\| = \delta$. On the other hand, by the convexity of f,

$$f(y^k) \le (1 - \lambda_k) f(\bar{x}^k) + \lambda_k f(x^k) = f_{\min} + \delta \cdot \frac{f(x^k) - f(\bar{x}^k)}{\|x^k - \bar{x}^k\|} \to f_{\min}.$$

By Theorem 2.5, $\operatorname{dist}(y^k, X^*) \to 0$ as $k \to \infty$. The contradiction proves (2.13). □

If f is only a piecewise convex quadratic function, then Theorem 2.6 is not true. In general, (2.11) does not hold for any $\delta \ge 0$ (cf. Mangasarian and Shiau's example [29]). However, for a convex piecewise quadratic function, the restriction on δ can be removed.

THEOREM 2.7. *Suppose that f is a convex piecewise quadratic function. Then, for any constant $\delta \ge 0$, there exists a positive constant γ such that*

$$\operatorname{dist}(x, X^*) \le \gamma\sqrt{f(x) - f_{\min}} \quad \textit{for } x \in X \textit{ with } f(x) - f_{\min} \le \delta. \tag{2.15}$$

Proof. By Theorem 2.5, there exist positive constants δ_0 and γ_1 such that

$$\operatorname{dist}(x, X^*) \le \gamma_1\sqrt{f(x) - f_{\min}} \quad \text{for } x \in X \text{ with } f(x) - f_{\min} \le \delta_0. \tag{2.16}$$

By Theorem 2.6, there exists a constant $\gamma_2 > 0$ such that

$$\operatorname{dist}(x, X^*) \le \gamma_2(f(x) - f_{\min}) \quad \text{for } x \in X \text{ with } f(x) - f_{\min} \ge \delta_0. \tag{2.17}$$

If $\delta_0 \le f(x) - f_{\min} \le \delta$, then $f(x) - f_{\min} \le \delta \le \frac{\delta}{\sqrt{\delta_0}}\sqrt{f(x) - f_{\min}}$. Therefore, it follows from (2.16) and (2.17) that (2.15) holds with $\gamma := \max\{\gamma_1, \frac{\gamma_2\delta}{\sqrt{\delta_0}}\}$. □

An easy consequence of Theorems 2.5 and 2.6 is the following global error estimate for feasible approximate solutions to a convex piecewise quadratic program.

COROLLARY 2.8. *Suppose that f is a convex piecewise quadratic function. Then there exists a positive constant γ such that*

$$\operatorname{dist}(x, X^*) \le \gamma\left(f(x) - f_{\min} + \sqrt{f(x) - f_{\min}}\right) \quad \textit{for } x \in X. \tag{2.18}$$

Remark. Recently, Luo and Pang [17] studied error estimates for approximate solutions of nonlinear feasibility problems:

$$\begin{aligned} &f_i(x) = 0, \qquad 1 \le i \le m, \\ &Ax \ge b. \end{aligned} \tag{2.19}$$

Let X^* denote the solution set of (2.19). Then they established various bounds for $\operatorname{dist}(x, X^*)$. In particular, they proved [17] that

$$\begin{aligned} \operatorname{dist}(x, X^*) \le& \gamma \left(\sqrt{\|(Ax-b)_+\|} + \|(Ax-b)_+\| + \sum_{i=1}^{m} (\sqrt{|f_i(x)|} + |f_i(x)|) \right) \\ &\text{for } x \in \mathbb{R}^n, \end{aligned} \tag{2.20}$$

if $f_i(x)$ are nonnegative on the polyhedral set $\{x : Ax \ge b\}$ and $f_i(x)$ are convex quadratic functions. If $f(x)$ is a convex quadratic function, then, with $m = 1$ and $f_1(x) = f(x) - f_{\min}$, (2.18) follows from Luo and Pang's global estimate (2.20). It is unclear whether (2.20) can be used to derive the following estimate for any approximate solution (not necessarily feasible) of (1.1):

$$\begin{aligned} \operatorname{dist}(x, X^*) \le& \gamma \left(\sqrt{\|(Ax-b)_+\|} + \|(Ax-b)_+\| + \sqrt{(f(x)-f_{\min})_+} + (f(x)-f_{\min})_+ \right) \\ &\text{for } x \in \mathbb{R}^n, \end{aligned} \tag{2.21}$$

where $f(x)$ is a convex piecewise quadratic function.

The following corollary shows directly why the estimate (2.11) is an error estimate for approximate feasible solutions near the solution set X^*.

COROLLARY 2.9. *Suppose that there exist positive constants δ and γ such that*

$$\operatorname{dist}(x, X^*) \le \gamma \sqrt{f(x) - f_{\min}} \quad \textit{for } x \in X \textit{ with } f(x) - f_{\min} \le \delta. \tag{2.22}$$

Then, for any $\beta > 0$, there exists a positive constant α such that

$$\operatorname{dist}(x, X^*) \le \alpha \sqrt{f(x) - f_{\min}} \quad \textit{for } x \in X \textit{ with } \operatorname{dist}(x, X^*) \le \beta. \tag{2.23}$$

Proof. Let $\alpha := \max\left\{\gamma, \frac{\beta}{\sqrt{\delta}}\right\}$. Let $x \in X$ with $\operatorname{dist}(x, X^*) \le \beta$. If $f(x) - f_{\min} \le \delta$, then (2.23) follows from (2.22); otherwise, $\sqrt{f(x) - f_{\min}} \ge \sqrt{\delta} \ge \frac{\sqrt{\delta}}{\beta} \operatorname{dist}(x, X^*)$. □

Similarly, we can recast Theorem 2.6 as the error estimate for approximate feasible solutions away from the solution set X^*.

COROLLARY 2.10. *Suppose that f is a convex piecewise quadratic function. Then, for any constant $\delta > 0$, there exists a positive constant γ such that*

$$\operatorname{dist}(x, X^*) \le \gamma \sqrt{f(x) - f_{\min}} \quad \textit{for } x \in X \textit{ with } \operatorname{dist}(x, X^*) \le \delta, \tag{2.24}$$

$$\operatorname{dist}(x, X^*) \le \gamma\,(f(x) - f_{\min}) \quad \textit{for } x \in X \textit{ with } \operatorname{dist}(x, X^*) \ge \delta. \tag{2.25}$$

Proof. By Theorem 2.7, there exists a positive constant γ_1 such that

$$\operatorname{dist}(x, X^*) \le \gamma_1 \sqrt{f(x) - f_{\min}} \quad \text{for } x \in X \text{ with } f(x) - f_{\min} \le \delta. \tag{2.26}$$

By Theorem 2.6, there exists a positive constant γ_2 such that

$$\operatorname{dist}(x, X^*) \leq \gamma_2 \left(f(x) - f_{\min}\right) \quad \text{for } x \in X \text{ with } f(x) - f_{\min} \geq \min\left\{\delta, \left(\frac{\delta}{\gamma_1}\right)^2\right\}. \tag{2.27}$$

Let $\gamma := \max\left\{\gamma_1, \gamma_2, 1, \sqrt{\delta}\right\}$. Assume $x \in X$ with $\operatorname{dist}(x, X^*) \leq \delta$. If $f(x) - f_{\min} \leq \delta$, then, by (2.26), $\operatorname{dist}(x, X^*) \leq \gamma\sqrt{f(x) - f_{\min}}$; otherwise,

$$\operatorname{dist}(x, X^*) \leq \delta \leq \sqrt{\delta}\sqrt{f(x) - f_{\min}} \leq \gamma\sqrt{f(x) - f_{\min}}.$$

Hence, (2.24) holds. Now let $x \in X$ with $\operatorname{dist}(x, X^*) \geq \delta$. If $f(x) - f_{\min} \geq \delta$, then, by (2.27), $\operatorname{dist}(x, X^*) \leq \gamma\left(f(x) - f_{\min}\right)$; otherwise, by (2.26), $f(x) - f_{\min} \geq \left(\frac{\delta}{\gamma_1}\right)^2$. In the latter case, by (2.27), we also have $\operatorname{dist}(x, X^*) \leq \gamma\left(f(x) - f_{\min}\right)$. Therefore, (2.25) also holds. □

Remark. In general, we cannot have the estimate $\operatorname{dist}(x, X^*) \leq \gamma(f(x) - f_{\min})$ for $x \in X$ (cf. [5]). It would be interesting to know what additional conditions are needed to guarantee such an estimate.

3. Error bounds for linear programs, linear complementarity problems, convex quadratic programs, and affine variational inequality problems.

Using the local and global error estimates given in the previous section, we can recover the weak sharp minimum property of a linear program by Mangasarian and Meyer [27] and the local error estimate for approximate solutions of an affine variational inequality problem by Luo and Tseng [21]. The unified approach shows that the estimates depend on how the original problem can be reformulated as a piecewise convex quadratic program. Therefore, it becomes clear that different reformulations of a given problem as piecewise convex quadratic programs yield different error estimates for approximate solutions of the given problem. This idea leads to a new global error estimate (cf. Theorem 3.1) for monotone linear complementarity problems which is better than the one given by Mangasarian and Shiau [29]. Moreover, using different reformulations of a convex quadratic program, we have new local and global error estimates for approximate solutions of a convex quadratic program (cf. Theorems 3.2, 3.3, and 3.4).

Consider the linear programming problem

$$c_{\min} := \min_{x \in X} c^T x. \tag{3.1}$$

Let X^* be the solution set of (3.1). Then X^* is also a solution of the following convex quadratic programming problem:

$$\min_{x \in X} (c^T x - c_{\min})^2. \tag{3.2}$$

Applying Theorem 2.7 to (3.2), we obtain that there exists a positive constant γ_1 such that

$$\operatorname{dist}(x, X^*) \leq \gamma_1\sqrt{(c^T x - c_{\min})^2} = \gamma_1(c^T x - c_{\min}) \quad \text{for } x \in X, (c^T x - c_{\min})^2 \leq 1. \tag{3.3}$$

However, (3.1) is also a degenerate convex quadratic programming problem. By Theorem 2.6, there exists a positive constant γ_2 such that

$$\operatorname{dist}(x, X^*) \leq \gamma_2(c^T x - c_{\min}) \quad \text{for } x \in X, (c^T x - c_{\min})^2 \geq 1. \tag{3.4}$$

From (3.3) and (3.4) we can recover the following weak sharp minimum property of the solutions of a linear program, which was proved by Mangasarian and Meyer [27]:

$$\mathrm{dist}(x, X^*) \leq \gamma(c^T x - c_{\min}) \quad \text{for } x \in X,$$

where $\gamma = \max\{\gamma_1, \gamma_2\}$.

An affine variational inequality problem, associated with X, an $n \times n$ matrix M, and a vector $q \in \mathbb{R}^n$, is to find a vector $x \in X$ such that

$$(3.5) \qquad (y - x)^T(q + Mx) \geq 0 \quad \text{for } y \in X.$$

The affine variational inequality problem (3.5) is equivalent to the following system of piecewise linear equations:

$$(3.6) \qquad x - \Pi_X(x - (q + Mx)) = 0,$$

where $\Pi_X(z)$ denotes the orthogonal projection of z onto X. It is known that $\Pi_X(z)$ is a piecewise linear mapping of z if X is a convex polyhedral set. Let $f(x) := \|x - \Pi_X(x - (q + MX))\|^2$. Then $f(x)$ is a piecewise convex quadratic function and (3.6) is equivalent to the following piecewise convex quadratic program:

$$(3.7) \qquad \min_{x \in \mathbb{R}^n} f(x).$$

It follows from Theorem 2.5 and $f_{\min} = 0$ that there exist positive constants γ and δ such that

$$(3.8) \quad \mathrm{dist}(x, X^*) \leq \gamma\|x - \Pi_X(x - (q + Mx))\| \quad \text{for } \|x - \Pi_X(x - (q + Mx))\| \leq \delta,$$

which is the local error estimate derived by Luo and Tseng [21].

Next, consider a special case of the affine variational inequality problem—the monotone linear complementarity problem

$$(3.9) \qquad x^T(q + Mx) = 0, \ x \geq 0, \ q + Mx \geq 0,$$

where M is positive semidefinite. Similarly, the solution set X^* of the monotone linear complementarity problem is the solution set of (3.7) with $f(x) := \|x - (x - (q + Mx))_+\|^2$, where z_+ denotes the vector whose ith component is $\max\{z_i, 0\}$. In order to obtain a global error estimate for approximate solutions to (3.9), we need to use Corollaries 2.9 and 2.10. Since $f_{\min} = 0$ in this case, by Corollary 2.9, there exist positive constants δ_1 and γ_1 such that

$$(3.10) \qquad \mathrm{dist}(x, X^*) \leq \gamma_1\|x - (x - (q + Mx))_+\| \quad \text{for } \mathrm{dist}(x, X^*) \leq \delta_1.$$

On the other hand, by the proof of Lemma 2.5 in [29], there exists a positive constant σ such that

$$(3.11) \qquad x^T(Mx + q) + \frac{\sigma}{2}(\|(-q - Mx)_+\|_1 + \|(-x)_+\|_1) \geq 0 \quad \text{for } x \in \mathbb{R}^n,$$

where $\|z\|_1 := \sum_{i=1}^n |z_i|$. Let $g(x) := x^T(Mx + q) + \sigma(\|(-q - Mx)_+\|_1 + \|(-x)_+\|_1)$. Then it is easy to verify that $g(x)$ is a convex piecewise quadratic function, since M is positive semidefinite. By (3.11), $g(x) \geq 0$ and $g(x) = 0$ if and only if $\|(-q - Mx)_+\|_1 +$

$\|(-x)_+\|_1 = 0$ and $x^T(Mx+q) = 0$. Therefore, the solution set X^* of the monotone linear complementarity problem is the solution set of the following convex piecewise quadratic program:

$$\min_{x \in \mathbb{R}^n} g(x). \tag{3.12}$$

By Corollary 2.10, for the constant $\delta_1 > 0$, there exists a positive constant γ_2 such that, for $\operatorname{dist}(x, X^*) \geq \delta_1$,

$$\begin{aligned} \operatorname{dist}(x, X^*) &\leq \gamma_2(g(x) - g_{\min}) \\ &= \gamma_2(x^T(Mx+q) + \sigma(\|(-q-Mx)_+\|_1 + \|(-x)_+\|_1)) \\ &\leq \gamma_2(\sigma+1)((x^T(Mx+q))_+ + \|(-q-Mx)_+\|_1 + \|(-x)_+\|_1). \end{aligned} \tag{3.13}$$

It is easy to verify (cf. [29]) that there exists a positive constant γ_3 such that

$$\|(-q-Mx)_+\|_1 + \|(-x)_+\|_1 \leq \gamma_3\|x - (x-(Mx+q))_+\| \quad \text{for } x \in \mathbb{R}^n. \tag{3.14}$$

From (3.10), (3.13), and (3.14) we get the following global error estimate for approximate solutions of a monotone linear complementarity problem.

THEOREM 3.1. *Suppose that M is positive semidefinite and X^* is the solution set of the monotone linear complementarity problem* (3.9). *Then there exists a positive constant γ such that*

$$\operatorname{dist}(x, X^*) \leq \gamma((x^T(Mx+q))_+ + \|x - (x-(Mx+q))_+\|) \quad \textit{for } x \in \mathbb{R}^n. \tag{3.15}$$

Remark. Recently, Mangasarian and Ren [28] also obtained an estimate similar to (3.15).

Now, let us compare the error estimate (3.15) with Mangasarian and Shiau's global error estimate for approximate solutions of a monotone linear complementarity problem.

Applying Corollary 2.8 to the convex piecewise quadratic program (3.12), we obtain that there exists a constant γ_0 such that

$$\operatorname{dist}(x, X^*) \leq \gamma_0(g(x) + \sqrt{g(x)}) \quad \text{for } x \in \mathbb{R}^n. \tag{3.16}$$

Since $g(x) \leq (\sigma+1)((x^T(Mx+q))_+ + \|(-q-Mx)_+\|_1 + \|(-x)_+\|_1)$, using Mangasarian and Shiau's notation $\|(x^T(Mx+q), -q-Mx, -x)_+\|^2 := (x^T(Mx+q))_+^2 + \|(-q-Mx)_+\|^2 + \|(-x)_+\|^2$ and the fact that any two norms in $\mathbb{R}^n$ are equivalent, we can derive from (3.16) the following error estimate:

$$\begin{aligned} \operatorname{dist}(x, X^*) \leq \gamma\Big(&\|(x^T(Mx+q), -q-Mx, -x)_+\| \\ &+ \sqrt{\|(x^T(Mx+q), -q-Mx, -x)_+\|}\Big) \quad \text{for } x \in \mathbb{R}^n, \end{aligned} \tag{3.17}$$

which is an equivalent form of Mangasarian and Shiau's estimate (2.6) in [29]. Note that they also have an explicit expression for γ with appropriate norms. However, it is interesting to see that, theoretically, their estimate can be derived as a special case of Corollary 2.8.

In the following analysis, we show that there exists a positive constant κ such that, for $x \in \mathbb{R}^n$,

$$(3.18)\quad \begin{aligned}&(x^T(Mx+q))_+ + \|x-(x-(Mx+q))_+\| \\ &\quad \le \kappa\left(\|(x^T(Mx+q), -q-Mx, -x)_+\| + \sqrt{\|(x^T(Mx+q), -q-Mx, -x)_+\|}\right),\end{aligned}$$

provided (3.17) holds.

Since $\|x-(x-(Mx+q))_+\|_1$ is a piecewise linear function, there exists a positive constant β such that

$$\big|\|x-(x-(Mx+q))_+\|_1 - \|y-(y-(My+q))_+\|_1\big| \le \beta\|x-y\|.$$

For any x, let $y \in X^*$ be such that $\mathrm{dist}(x, X^*) = \|x-y\|$. Then

$$\begin{aligned}\|x-(x-(Mx+q))_+\|_1 &= \|x-(x-(Mx+q))_+\|_1 - \|y-(y-(My+q))_+\|_1 \\ &\le \beta\cdot\|x-y\| = \beta\cdot\mathrm{dist}(x, X^*).\end{aligned}$$

Since any two norms in $\mathbb{R}^n$ are equivalent, there exists a positive constant η such that

$$(3.19)\qquad \|x-(x-(Mx+q))_+\| \le \eta\cdot\mathrm{dist}(x, X^*).$$

By (3.17) and (3.19), we get (3.18) with $\kappa := \gamma\eta + 1$. The inequality (3.18) and the following example show that (3.15) is a better global error estimate than (3.17).

Example. Let $M = \left(\begin{smallmatrix}1 & 1\\ 0 & 0\end{smallmatrix}\right)$ and $q = \left(\begin{smallmatrix}0\\ 0\end{smallmatrix}\right)$. Then it is easy to verify that

$$X^* = \left\{x \equiv \begin{pmatrix}x_1\\ x_2\end{pmatrix} : x_1 = 0, x_2 \ge 0\right\}.$$

Hence, $\mathrm{dist}(x, X^*)^2 = x_1^2 + (-x_2)_+^2$. Next we prove that (3.17) holds and then compare (3.17) with (3.15).

For any $x \in \mathbb{R}^n$,

$$s(x)^2 := \|(x^T(Mx+q), -x, -Mx-q)_+\|^2 = (-x_1)_+^2 + (-x_2)_+^2 + (-x_1-x_2)_+^2 + (x_1^2+x_1x_2)_+^2.$$

For $x_1 \le 0$,

$$s(x)^2 \ge (-x_1)_+^2 + (-x_2)_+^2 = x_1^2 + (-x_2)_+^2 = \mathrm{dist}(x, X^*)^2.$$

If $x_1 > 0$ and $x_2 \ge 0$, then

$$s(x)^2 + s(x) \ge (-x_2)_+^2 + (x_1^2 + x_1x_2)_+ \ge x_1^2 + (-x_2)_+^2 = \mathrm{dist}(x, X^*)^2.$$

When $x_1 > 0$ and $-\frac{x_1}{2} \le x_2 \le 0$,

$$s(x)^2 + s(x) \ge (-x_2)_+^2 + (x_1^2 + x_1x_2)_+ \ge \frac{x_1^2}{2} + (-x_2)_+^2 \ge \frac{1}{2}\mathrm{dist}(x, X^*)^2.$$

When $x_1 > 0$ and $-\frac{x_1}{2} \ge x_2$,

$$s(x)^2 \ge (-x_2)_+^2 \ge \frac{1}{2}(-x_2)_+^2 + \frac{1}{8}x_1^2 \ge \frac{1}{8}\mathrm{dist}(x, X^*)^2.$$

Therefore,

$$\operatorname{dist}(x, X^*) \leq 3(s(x) + \sqrt{s(x)}) \qquad \text{for } x \in \mathbb{R}^n.$$

By (3.18), (3.15) also holds. Actually, we have the following better estimate:

$$\begin{aligned} \|x - (x - (Mx+q))_+\|^2 &= (x_1 - (-x_2)_+)^2 + (-x_2)_+^2 \\ &= x_1^2 - 2x_1(-x_2)_+ + 2(-x_2)_+^2 \\ &= x_1^2 - 2\left(\sqrt{\frac{2}{3}}x_1\right)\left(\sqrt{\frac{3}{2}}(-x_2)_+\right) + 2(-x_2)_+^2 \\ &\geq x_1^2 - \left(\frac{2}{3}x_1^2\frac{3}{2}(-x_2)_+^2\right) + 2(-x_2)_+^2 \\ &\geq \frac{1}{3}(x_1^2 + (-x_2)_+^2). \end{aligned}$$

Thus, for $\lambda = \sqrt{3}$,

$$\operatorname{dist}(x, X^*) \leq \lambda \|x - (x - (Mx+q))_+\| \qquad \text{for } x \in \mathbb{R}^n. \tag{3.20}$$

In the following analysis, we show that (3.20) is better than (3.15), which is better than (3.17).

Let

$$x_\theta = \begin{pmatrix} \theta^2 \\ \theta^{-1} \end{pmatrix}$$

with $0 < \theta < 1$. Then

$$\operatorname{dist}(x_\theta, X^*) = \theta^2, \quad \|x_\theta - (x_\theta - (Mx_\theta + q))_+\| = \theta^2, \quad (x_\theta^T(Mx_\theta + q))_+ = \theta^4 + \theta.$$

Therefore,

$$\lim_{\theta \to 0+} \frac{(x_\theta^T(Mx_\theta + q))_+ + \|x_\theta - (x_\theta - (Mx_\theta + q))_+\|}{\sqrt{\|(x_\theta^T(Mx_\theta + q), -q - Mx_\theta, -x_\theta)_+\|}} = 0, \tag{3.21}$$

$$\lim_{\theta \to 0+} \frac{\|x_\theta - (x_\theta - (Mx_\theta + q))_+\|}{(x_\theta^T(Mx_\theta + q))_+} = 0. \tag{3.22}$$

Remark. An analysis given by Mangasarian and Ren [28] also shows that (3.15) holds whenever (3.17) holds. They give a counterexample showing that (3.15) does not imply (3.17). As a consequence, (3.15) holds for a wider class of linear complementarity problems than (3.17). The above analysis shows that, in the case where $s(x) + \sqrt{s(x)}$ can be used as a global error estimate, one should still use $\|x - (x - (Mx + q))_+\| + (x^T(Mx + q))_+$ as a global error estimate due to (3.18).

Finally, we use the error bounds derived in §2 to establish new error estimates for approximate solutions of a quadratic program.

If f is a convex quadratic function, the estimates given in §2 are only for feasible solutions of (1.1). However, in some cases, we can have global error estimates for any approximate solution of a convex quadratic program. The new global error estimates are based on the unconstrained reformulation of a convex quadratic program and the estimates given in §2. Consider the following convex quadratic program:

$$\min_{l\leq Ax\leq u} \frac{1}{2}x^T Mx - b^T x, \tag{3.23}$$

where M is an $n \times n$ symmetric positive semidefinite matrix, $b \in \mathbb{R}^n$, A is an $m \times n$ matrix, and $l, u \in \mathbb{R}^m$. When M is also nonsingular, x^* is a solution to (3.23) if and only if $x = M^{-1}(A^T w + b)$, where w is a solution of the following piecewise linear equation [15], [16]:

$$\varphi(w) := (\alpha I - B)w + AM^{-1}b - (AM^{-1}b - Bw)_l^u = 0. \tag{3.24}$$

Here α is any positive constant, I is the identity matrix, $B = \alpha I - AM^{-1}A^T$, and $(z)_l^u$ denotes the vector whose ith component is $\min\{u_i, \max\{z_i, l_i\}\}$. When $\alpha > \|AM^{-1}A^T\|$, the piecewise linear equation is equivalent to the following unconstrained minimization problem [15], [16]:

$$\Phi_{\text{min}} := \min_{w\in\mathbb{R}^m} \Phi(w), \tag{3.25}$$

where

$$\Phi(w) := \frac{\alpha}{2}w^T Bw - \frac{1}{2}\|(Bw + l - AM^{-1}b)_+\|^2 - \frac{1}{2}\|(AM^{-1}b - u - Bw)_+\|^2 \tag{3.26}$$

is a convex quadratic spline. That is, w^* is a solution of $\varphi(w) = 0$ (or $\|\varphi(w)\|^2 = 0$) if and only if w^* is a minimizer of $\Phi(w)$. Let W^* be the solution set of (3.24) (or (3.25)). Then W^* is the set of the Lagrange multipliers of the solution to (3.23). By Corollaries 2.9 and 2.10, we can use $\|\varphi(w)\|$ or $\sqrt{\Phi(w) - \Phi_{\text{min}}}$ as an error estimate for w near the solution set W^* and use $\Phi(w) - \Phi_{\text{min}}$ as an error estimate for w away from the solution set W^*. Moreover, if x^* is the solution to (3.23), then $x^* \equiv M^{-1}(A^T w^* + b)$ for all $w^* \in W^*$, which implies $\|M^{-1}(A^T w + b) - x^*\| = \|M^{-1}A^T(w - w^*)\|$ for all $w^* \in W^*$. Thus,

$$\|M^{-1}(A^T w + b) - x^*\| \leq \|M^{-1}A^T\| \cdot \text{dist}(w, W^*),$$

where $\|M^{-1}A^T\|$ denotes the 2-norm of the matrix $M^{-1}A^T$. Therefore, we have the following global estimate for Lagrange multipliers and approximate solution of (3.23).

THEOREM 3.2. *If M is symmetric positive definite and $\alpha > \|AM^{-1}A^T\|$, then there exists a positive constant γ such that, for any w and $x := M^{-1}(A^T w + b)$,*

$$\|x - x^*\| + \text{dist}(w, W^*) \leq \gamma(\min\{\|\varphi(w)\|, \sqrt{\Phi(w) - \Phi_{\text{min}}}\} + (\Phi(w) - \Phi_{\text{min}})). \tag{3.27}$$

If M is singular but A is nonsingular, then we can still have a global error estimate for approximate solutions of (3.23). Let $E := I - \alpha(A^{-1})^T MA^{-1}$ and $\bar{b} = \alpha(A^T)^{-1}b$. Define

$$\psi(x) := Ax - (EAx + \bar{b})_l^u. \tag{3.28}$$

Then, for any $\alpha > 0$, x is a solution to (3.23) if and only if $\psi(x) = 0$ (or $\|\psi(x)\|^2 = 0$). Since $\|\psi(x)\|^2$ is a piecewise convex quadratic function, by Theorem 2.5, we have the following local error estimate for approximate solutions of (3.23).

THEOREM 3.3. *Suppose that A is nonsingular and X^* is the solution set of* (3.23). *Then there exist positive constants γ and δ such that*

$$\text{dist}(x, X^*) \le \gamma \|Ax - (EAx + \bar{b})_l^u\| \quad \textit{for } \|Ax - (EAx + \bar{b})_l^u\| \le \delta. \tag{3.29}$$

When $0 < \alpha < \|(A^{-1})^T M A^{-1}\|^{-1}$, $x \in X^*$ if and only if x is a solution to the following unconstrained minimization problem:

$$\min_{x \in \mathbb{R}^n} \Psi(x), \tag{3.30}$$

where

$$\Psi(x) := \frac{1}{2} x^T A^T (E - E^2) A x + \frac{1}{2} \|(l - (EAx + \bar{b}))_+\|^2 + \frac{1}{2} \|((EAx + \bar{b}) - u)_+\|^2$$

is a convex quadratic spline [15], [16]. We have the following global error estimate for approximate solutions of (3.23).

THEOREM 3.4. *Let X^* be the solution set of* (3.30). *If A is nonsingular and $0 < \alpha < \|(A^{-1})^T M A^{-1}\|^{-1}$, then there exists a positive constant γ such that*

$$\text{dist}(x, X^*) \le \gamma(\min\{\|\psi(x)\|, \sqrt{\Psi(x) - \Psi_{\min}}\} + (\Psi(x) - \Psi_{\min})) \quad \text{for } x \in \mathbb{R}^n. \tag{3.31}$$

Remark. The most interesting case of nonsingular A's is $A = I$. Then (3.23) becomes the convex quadratic program with simple bound constraints. The estimate (3.27) (or (3.31)) involve the unknown value $\Phi_{\min}$ (or $\Psi_{\min}$), which is not desirable as a general error estimate for approximate solutions of (3.23). However, such estimates are useful in convergence analysis of descent methods for solving (3.25) or (3.30) (cf. [13], [16]). Moreover, it is interesting to know when $\|\varphi(w)\|$ (or $\|\psi(x)\|$) can be used as a global error estimate for $\text{dist}(w, W^*)$ (or $\text{dist}(x, X^*)$). Some related global error estimates can be found in [23] and [31].

4. Error estimates for iterates of the proximal point algorithm. Consider the following proximal point algorithm for solving (1.1):

$$x^{k+1} := \arg\min_{x \in X} f(x) + \frac{1}{2\epsilon_k} \|x - x^k\|_D^2 \quad k = 0, 1, \ldots, \tag{4.1}$$

where $\|x\|_D := (x^T D x)^{1/2}$ denotes the norm induced by a positive definite matrix D, f is a convex function, x^0 is any given vector, and $\epsilon_k > 0$ for all k.

The proximal point algorithm was first applied to solving convex programs by Martinet [30]. It is well known that $\{x^k\}$ converges to a solution to (1.1) if $\lim_{k \to \infty} \inf \epsilon_k = \epsilon > 0$. The convergence rate of iterates were studied by Rockafellar [33] and Luque [24], and sufficient conditions for the finite convergence of the proximal point algorithm were given by Rockafellar [33], Ferris [4], and Lefebvre and C. Michelot [9]. See [10] for a survey on the proximal point algorithm and see [7], [1]–[3], [8], [39] for the recent development on the proximal point algorithm.

Using the error bound for feasible approximate solutions to a convex piecewise quadratic program, we can derive new estimates for the distance from x^k to the solution set X^*.

The following result is a special case of Güler's key inequality in convergence analysis of the proximal point algorithm in a Hilbert space setting [7].

LEMMA 4.1. *For any $x \in X$,*

$$f(x^{k+1}) - f(x) \leq \frac{1}{2\epsilon_k}\left(\|x - x^k\|_D^2 - \|x - x^{k+1}\|_D^2 - \|x^{k+1} - x^k\|_D^2\right). \tag{4.2}$$

THEOREM 4.2. *Suppose that there exist positive constants δ, γ such that*

$$\operatorname{dist}(x, X^*) \leq \gamma\sqrt{f(x) - f_{\min}} \quad \textit{for } x \in X \textit{ with } f(x) - f_{\min} \leq \delta. \tag{4.3}$$

Then there exists a positive constant η (depending only on f, X, D, and $\operatorname{dist}(x^0, X^*)_D$*) such that*

$$\|x^{k+1} - x^k\|_D \leq \operatorname{dist}(x^k, X^*)_D \leq \sqrt{\frac{1}{1 + \eta\epsilon_{k-1}}} \cdot \operatorname{dist}(x^{k-1}, X^*)_D \quad \textit{for } k \geq 1, \tag{4.4}$$

where $\operatorname{dist}(x, X^*)_D := \min_{\bar{x} \in X^*} \|x - \bar{x}\|_D$.

Proof. Since D is positive definite, there exists a positive constant β such that

$$\|x\|_D \leq \beta\|x\| \quad \text{for } x, y \in \mathbb{R}^n,$$

which implies $\operatorname{dist}(x, X^*)_D \leq \beta \cdot \operatorname{dist}(x, X^*)$. Therefore, by (4.3) and Corollary 2.9, there exists a positive constant η (depending only on f, X, D, and $\operatorname{dist}(x^0, X^*)_D$) such that

$$\operatorname{dist}(x, X^*)_D \leq \sqrt{\frac{2}{\eta}}\sqrt{f(x) - f_{\min}} \quad \text{for } x \in X \text{ with } \operatorname{dist}(x, X^*)_D \leq \operatorname{dist}(x^0, X^*)_D. \tag{4.5}$$

Let $\bar{x}^k \in X^*$ be such that $\operatorname{dist}(x^k, X^*)_D = \|x^k - \bar{x}^k\|_D$. By (4.5),

$$f(x^{k+1}) - f(\bar{x}^{k+1}) \geq \frac{\eta}{2}\|x^{k+1} - \bar{x}^{k+1}\|_D^2. \tag{4.6}$$

Substituting x by $\bar{x}^k$ in (4.2), we obtain

$$\begin{aligned} f(x^{k+1}) - f(\bar{x}^{k+1}) &= f(x^{k+1}) - f(\bar{x}^k) \\ &\leq \frac{1}{2\epsilon_k}\left(\|x^k - \bar{x}^k\|_D^2 - \|x^{k+1} - \bar{x}^k\|_D^2 - \|x^{k+1} - x^k\|_D^2\right) \\ &\leq \frac{1}{2\epsilon_k}\left(\|x^k - \bar{x}^k\|_D^2 - \|x^{k+1} - \bar{x}^k\|_D^2\right) \\ &\leq \frac{1}{2\epsilon_k}\left(\|x^k - \bar{x}^k\|_D^2 - \|x^{k+1} - \bar{x}^{k+1}\|_D^2\right). \end{aligned} \tag{4.7}$$

It follows from (4.6) and (4.7) that

$$\|x^{k+1} - \bar{x}^{k+1}\|_D^2 \leq \frac{1}{1 + \eta\epsilon_k}\|x^k - \bar{x}^k\|_D^2 \quad \text{for } k = 0, 1, \ldots. \tag{4.8}$$

Substituting x by $\bar{x}^k$ in (4.2), we get

$$\|x^k - x^{k+1}\|_D^2 \leq \|x^k - \bar{x}^k\|_D^2 - \|x^{k+1} - \bar{x}^k\|_D^2 - 2\epsilon_k(f(x^{k+1}) - f(\bar{x}^k)) \leq \|x^k - \bar{x}^k\|_D^2.$$

□

Remark. It follows from (4.4) that

$$\|x^k - x^{k+1}\|_D \le \operatorname{dist}(x^k, X^*)_D \le \operatorname{dist}(x^0, X^*)_D \left(\prod_{i=0}^{k-1} \frac{1}{1+\eta\epsilon_i}\right)^{\frac{1}{2}} \quad \text{for } k \ge 1. \tag{4.9}$$

Note that $\prod_{i=0}^{\infty} \frac{1}{1+\eta\epsilon_i} = 0$ if and only if $\prod_{i=0}^{\infty}(1+\eta\epsilon_i) = \infty$, which is equivalent to

$$\sum_{i=0}^{\infty} \eta\epsilon_i = \sum_{i=1}^{\infty} ((1+\eta\epsilon_i) - 1) = \infty. \tag{4.10}$$

Therefore, by (4.9), any accumulation point of $\{x^k\}$ is a solution of (1.1) whenever $\sum_{i=0}^{\infty} \epsilon_i = \infty$. As a consequence, for any sequence of positive constants $\{\epsilon_k\}$ with $\sum_{i=0}^{\infty} \epsilon_i = \infty$, the iterates $\{x^k\}$ generated by (4.1) converge to a solution to (1.1). This recovers Güler's convergence result for the proximal point algorithm [7] under the assumption (4.3). Moreover, the error estimate (4.9) is interesting in its own right. If $\epsilon_k \ge c > 0$, by (4.9), the iterates $\{x^k\}$ generated by (4.1) converge linearly (in the root sense) to a solution to (1.1). When $\lim_{k\to\infty} \epsilon_k = \infty$, by (4.9), $\{x^k\}$ converges superlinearly (in the root sense) to a solution to (1.1).

By Theorem 2.5, (4.4) and (4.9) hold if f is a convex piecewise quadratic function. However, in this case, the subdifferential mapping ∂f of f is polyhedral [38] and is upper Lipschitz continuous [32]. Thus, if f is a convex piecewise quadratic function, (4.4) also follows from the proof of Theorem 2.1 in [24], which is an extension of Rockafellar's linear convergence result of the proximal point algorithm [33]. Of course, one can also discuss the linear convergence of inexact proximal point algorithms under the assumption (4.3) (cf. [33], [24]).

Somehow, (4.9) provides a link between Rockafellar and Luque's linear convergence results and Güler's convergence result and shows how $\{\epsilon_k\}$ affect the convergence rate of iterates $\{x^k\}$.

The proximal point algorithm can also be formulated as an algorithm for solving the equivalent generalized equation of (1.1). In this case, Luque studied the relationship between convergence rates of iterates and error estimates of approximate solutions in terms of the normal equation. Similar results can be established for the relationship between convergence rates of iterates and error estimates of feasible solutions in terms of the objective function.

Since η in Theorem 4.2 depends on $\operatorname{dist}(x^0, X^*)_D$, (4.9) might not be considered as a "global" error estimate for x^k, even though it holds for all k. However, as an application of global error estimates for feasible solutions of convex piecewise quadratic programs, we can have a "global" error estimate for x^k when f is a convex piecewise quadratic function.

PROPOSITION 4.3. *Suppose that f is a convex piecewise quadratic function. Then there exists a positive constant β (depending only on f, X, and D) such that*

$$\operatorname{dist}(x^{k+1}, X^*) \le \sqrt{\frac{1}{1+\beta\epsilon_k}} \cdot \operatorname{dist}(x^k, X^*)_D \quad \textit{for } \operatorname{dist}(x^{k+1}, X^*)_D \le 1,$$

$$\operatorname{dist}(x^{k+1}, X^*) \le \frac{\operatorname{dist}(x^0, X^*)_D}{\operatorname{dist}(x^0, X^*)_D + \beta\epsilon_k} \cdot \operatorname{dist}(x^k, X^*)_D \quad \textit{for } \operatorname{dist}(x^{k+1}, X^*)_D \ge 1.$$

Proof. Due to the equivalence of any two norms on $\mathbb{R}^n$, by Theorems 2.6 and 2.7, there exists a positive constant β (depending only on f, X, and D) such that

$$\min\{\operatorname{dist}(x, X^*)_D^2, \operatorname{dist}(x, X^*)_D\} \le \frac{1}{\beta}(f(x) - f_{\min}) \quad \text{for } x \in X. \tag{4.11}$$

Let $\bar{x}^k \in X^*$ be such that $\operatorname{dist}(x^k, X^*)_D = \|x^k - \bar{x}^k\|_D$. If $\|x^{k+1} - \bar{x}^{k+1}\|_D \le 1$, then $\|x^{k+1} - \bar{x}^{k+1}\|_D^2 \le \frac{1}{\beta}(f(x^{k+1}) - f(\bar{x}^{k+1}))$ and it follows from the proof of (4.4) that

$$\operatorname{dist}(x^{k+1}, X^*)_D^2 \le \frac{1}{1+2\beta\epsilon_k}\operatorname{dist}(x^k, X^*)_D^2 \le \frac{1}{1+\beta\epsilon_k}\operatorname{dist}(x^k, X^*)_D^2. \tag{4.12}$$

Otherwise,

$$\|x^{k+1} - \bar{x}^{k+1}\|_D \le \frac{1}{\beta}(f(x^{k+1}) - f(\bar{x}^{k+1})). \tag{4.13}$$

By (4.7) and (4.13), we get

$$\|x^{k+1} - \bar{x}^{k+1}\|_D^2 + 2\beta\epsilon_k\|x^{k+1} - \bar{x}^{k+1}\|_D \le \|x^k - \bar{x}^k\|_D^2.$$

Solving the above inequality for $\|x^{k+1} - \bar{x}^{k+1}\|_D$, we obtain

$$\begin{aligned} \|x^{k+1} - \bar{x}^{k+1}\|_D &\le \frac{1}{2}(\sqrt{(2\beta\epsilon_k)^2 + 4\|x^k - \bar{x}^k\|_D^2} - 2\beta\epsilon_k) \\ &= (\sqrt{(\beta\epsilon_k)^2 + \|x^k - \bar{x}^k\|_D^2} - \beta\epsilon_k) \\ &= \frac{\|x^k - \bar{x}^k\|_D^2}{\sqrt{(\beta\epsilon_k)^2 + \|x^k - \bar{x}^k\|_D^2} + \beta\epsilon_k} \\ &\le \frac{\|x^k - \bar{x}^k\|_D^2}{\|x^k - \bar{x}^k\|_D + \beta\epsilon_k}. \end{aligned} \tag{4.14}$$

Since $\frac{t}{t+\beta\epsilon_k}$ is a monotone increasing function for $t \ge 0$ and $\|x^k - \bar{x}^k\|_D \le \operatorname{dist}(x^0, X^*)_D$, it follows from (4.14) that

$$\operatorname{dist}(x^{k+1}, X^*)_D \le \frac{\operatorname{dist}(x^0, X^*)_D}{\operatorname{dist}(x^0, X^*)_D + \beta\epsilon_k} \cdot \operatorname{dist}(x^k, X^*)_D. \tag{4.15}$$

The proposition follows from (4.14) and (4.15). □

Remark. The above estimate suggests that, initially, the convergence rate of the proximal point algorithm might be affected not only by the error bound constant $\frac{1}{\beta}$ for f but also by the distance of x^0 to the solution set X^*.

5. Comments. In this paper, we have established a local error estimate for feasible solutions of a piecewise convex quadratic program and a global error estimate for feasible solutions of a convex piecewise quadratic program. These error estimates provide a unified approach for deriving many old and new error estimates for linear programs, linear complementarity problems, convex quadratic programs, and affine variational inequality problems. The approach reveals the fact that each error estimate is a consequence of some reformulation of the original problem as a piecewise convex quadratic program or a convex piecewise quadratic program.

As an application, we give new (global) error estimates for iterates of the proximal point algorithm for solving a convex piecewise quadratic program, which provide additional insight on the convergence behavior of the iterates.

For those who are interested in general convex programming problems, similar results might be possible when the objective function is a composition of a strongly convex smooth function and a piecewise linear mapping. That is, if $h'(y)$ is Lipschitz continuous and strongly monotone, $g(x)$ is a piecewise linear mapping, and $f(x) := h(g(x))$, then the main results given in §2 might still be true. See [20], [19], [11] for related results. Also it seems natural to conjecture that Theorem 2.5 is still true if f is only a piecewise quadratic function.

As the conclusion of this paper, we want to show that the estimate (2.11) is actually a generalization of Robinson's result on the upper Lipschitz continuity of a polyhedral mapping [32]. Recall that a set-valued mapping $\Gamma(y)$ from $\mathbb{R}^n$ to subsets of $\mathbb{R}^n$ is said to be a polyhedral mapping if its graph $\{(x,y) : x \in \Gamma(y)\}$ is a union of finitely many closed convex polyhedral sets. Therefore, for a given polyhedral mapping $\Gamma(y)$, there exist matrices $\{A_i, B_i\}_{i=1}^r$ and vectors $\{c_i\}_{i=1}^r$ such that

$$\{(x,y) : x \in \Gamma(y)\} = \bigcup_{i=1}^{r} \{(x,y) : A_i x + B_i y \geq c_i\}.$$

Robinson proved that, if $\Gamma(z)$ is not empty, then there exist positive constants δ and γ such that

$$\text{dist}(x, \Gamma(z)) \leq \gamma \|y - z\| \quad \text{for } x \in \Gamma(y) \text{ with } \|z - y\| \leq \delta. \tag{5.1}$$

Define

$$f_z(x) := \left(\min_{1 \leq i \leq r} \|(A_i x + B_i z - c_i)_+\|_1 \right)^2 .$$

Since $\min\{t, s\} = t - (t - s)_+$, one can easily verify that $\sqrt{f_z(x)}$ is a piecewise linear function of both x and z; hence, for a fixed z, $f_z(x)$ is a piecewise convex quadratic function of x. Obviously, $\Gamma(z)$ is the solution set of $(0 =) \min_{x \in \mathbb{R}^n} f_z(x)$. By Theorem 2.5, there exist positive constants β and γ such that

$$\text{dist}(x, \Gamma(z)) \leq \gamma \sqrt{f_z(x)} \quad \text{for } x \in \mathbb{R}^n \text{ with } f_z(x) \leq \beta. \tag{5.2}$$

For $x \in \Gamma(y)$, we have $f_y(x) = 0$. By the piecewise linearity of $\sqrt{f_z(x)}$, there exists a positive constant α such that

$$\sqrt{f_z(x)} = \sqrt{f_z(x)} - \sqrt{f_y(x)} \leq \alpha \cdot \|z - y\|. \tag{5.3}$$

Let $\delta := \frac{\alpha}{\sqrt{\beta}}$ Then (5.1) follows from (5.2) and (5.3). The above discussion shows that (5.1) can be considered as a consequence of the error estimate (5.2) for approximate solutions of the piecewise convex quadratic function $f_z(x)$.

REFERENCES

[1] Y. Censor and A. Zenios, *Proximal minimization algorithm with D-functions*, J. Optim. Theory. Appl., 73 (1992), pp. 451–464.

[2] G. Chen and M. Teboulle, *Convergence Analysis of a Proximal-Like Minimization Algorithm using Bregman Functions*, Research Report 90-23, Dept. of Math and Statistics, Univ. of Maryland, Catonsville, MD, 1990.

[3] ———, *A Proximal-Based Decomposition Method for Convex Minimization Problems*, Research Report 92-11, Dept. of Math and Statistics, Univ. of Maryland, Catonsville, MD, 1992.

[4] M.C. Ferris, *Finite termination of the proximal point algorithm*, Math. Programming, 50 (1991), pp. 359–366.

[5] M.C. Ferris and O. L. Mangasarian, *Minimum principle sufficiency*, Math. Programming, 57 (1992), pp. 1–14.

[6] M. Frank and P. Wolfe, *An algorithm for quadratic programming*, Naval Res. Logist., 3 (1956), pp. 95–110.

[7] O. Güler, *On the convergence of the proximal point algorithm for convex minimization*, SIAM J. Control Optim., 29 (1991), pp. 403–419.

[8] A. N. Iusemi, B. F. Svaiter, and M. Teboulle, *Entropy-Like Proximal Methods in Convex Programming*, Research Report 92-09, Dept. of Math and Statistics, Univ. of Maryland, Catonsville, MD, 1992.

[9] O. Lefebvre and C. Michelot, *About the finite convergence of the proximal point algorithm*, in Trends in Mathematical Optimization, International Series of Numerical Mathematics, Vol. 84, Birkhäuser Verlag, Basel, 1988, pp. 153–161.

[10] B. Lemaire, *The proximal algorithm*, in International Series of Numerical Mathematics, Vol. 87, J. P. Penot ed., Birkhäuser Verlag, Basel, 1989, pp. 73–87.

[11] W. Li, *Error Bounds for solutions of parametric convex-concave minimax problems*, in Parametric Optimization and Related Topics III, J. Guddat, H. Th. Jongen, B. Kummer, and F. Nozicka, eds., Approximation & Optimization, Vol. 3, Verlag Peter Lang, Frankfurt am Main, 1993, pp. 373–394.

[12] ———, *Remarks on convergence of matrix splitting algorithm for the symmetric linear complementarity problem*, SIAM J. Optim., 3 (1993), pp. 155–163.

[13] ———, *Linearly convergent descent methods for unconstrained minimization of a convex quadratic spline*, J. Optim. Theory Appl., to appear.

[14] W. Li, P. Pardalos, and C. G. Han, *Gauss-Seidel method for least distance problems*, J. Optim. Theory Appl., 75 (1992), pp. 487–500.

[15] W. Li and J. Swetits, *A Newton method for convex regression, data smoothing, and quadratic programming with bounded constraints*, SIAM J. Optim., 3 (1993), pp. 466–488.

[16] ———, *A New Algorithm for Solving Strictly Convex Quadratic Programs*, Tech. Report TR92-14, Dept. of Math and Statistics, Old Dominion Univ., Norfolk, VA 23529, 1992; SIAM J. Optim, submitted.

[17] Z.-Q. Luo and J.-S. Pang, *Error Bounds for Analytic Systems and Their Applications*, Dept. of Mathematical Sciences, The Johns Hopkins University, Baltimore, MD, 1993, preprint.

[18] Z.-Q. Luo and P. Tseng, *On the convergence of a matrix splitting algorithm for the symmetric monotone linear complementarity problem*, SIAM J. Control Optim., 29 (1991) pp. 1037–1060.

[19] ———, *On the convergence rate of dual ascent methods for linearly constrained convex minimization*. Math. Oper. Res., 18 (1993), pp. 846–867.

[20] ———, *On the linear convergence of descent methods for convex essentially smooth minimization*, SIAM J. Control Optim., 30 (1992) pp. 408–425.

[21] ———, *Error bound and convergence analysis of matrix splitting algorithms for the affine variational inequality problem*, SIAM J. Optim., 2 (1992), pp. 43-54.

[22] ———, *On the convergence of the coordinate descent method for convex differentiable minimization*, J. Optim. Theory Appl., 72 (1992), pp. 7–35.

[23] ———, *On global error bound for a class of monotone* AVI *problems*, Oper. Res. Lett., 11 (1992), pp. 159–165.

[24] F. J. Luque, *Asymptotic convergence analysis of the proximal point algorithm*, SIAM J. Control Optim., 22 (1984), pp. 277–293.

[25] O. L. Mangasarian, *A simple characterization of solution sets of convex programs*, Oper. Res. Lett., 7 (1988), pp. 21–26.

[26] ———, *Convergence of iterates of an inexact matrix splitting algorithm for the symmetric monotone linear complementarity problem*, SIAM J. Optim., 1 (1991), pp. 114–122.

[27] O. L. Mangasarian and R. R. Meyer, *Nonlinear perturbation of linear programs*, SIAM J. Control Optim., 17 (1979), pp. 745–752.

[28] O. L. Mangasarian and J. Ren, *New Error Bounds for the Linear Complementarity Problem*, Tech. Report No. 1156, Computer Sciences Dept., Univ. of Wisconsin, Madison, WI, 1993.

[29] O.L. Mangasarian and T. H. Shiau, *Error bounds for monotone linear complementarity problems*, Math. Programming, 36 (1986), pp. 81–89.

[30] B. Martinet, *Régularisation d'inéquations variationnelles par approximations successives*, Rev. Francaise Inf. Rech. Oper., 1970, pp. 154–159.

[31] R. Mathias and J. S. Pang, *Error bounds for the linear complementarity problem with a P-matrix*, Linear Algebra Appl., 132 (1990), pp. 123–136.

[32] S. M. Robinson, *Some continuity properties of polyhedral multifunctions*, Math. Programming Stud., 14 (1981), pp. 206–214.

[33] R. T. Rockafellar, *Monotone operators and the proximal point algorithm*, SIAM J. Control. Optim., 14 (1976), pp. 877–898.

[34] ———, *Linear-quadratic programming and optimal control*, SIAM J. Control. Optim., 25 (1987), pp. 781–814.

[35] ———, *Large-scale extended linear-quadratic programming and multistage optimization*, in Advances in Numerical Partial Differential Equations and Optimization, S. Gomez, J.-P. Hennart, and R. Tapia, eds., Society for Industrial and Applied Mathematics, Philadelphia, PA, 1991, pp. 247–261.

[36] R. T. Rockafellar and R. J. B. Wets, *Generalized linear-quadratic problems of deterministic and stochastic optimal control in discrete time*, SIAM J. Control. Optim., 28 (1990), pp. 810–822.

[37] J. Sun, *On the structure of convex piecewise quadratic functions*, J. Optim. Theory Appl., 72 (1992), pp. 499–510.

[38] ———, On Monotropic Piecewise Quadratic Programming, Ph.D. thesis, University of Washington, Seattle, WA, 1986.

[39] M. Teboulle, *Entropic proximal mappings with applications to nonlinear programming*, Math. Oper. Res., 17 (1992), pp. 670–690.

SIAM J. CONTROL OPTIM.
Vol. 33, No. 5, pp. 1530–1559, September 1995

© 1995 Society for Industrial and Applied Mathematics

011

A PERSPECTIVE THEORY FOR MOTION AND SHAPE ESTIMATION IN MACHINE VISION[*]

B. K. GHOSH[†] AND E. P. LOUCKS[†]

Abstract. In this paper, we consider the problem of motion and shape estimation of a moving body with the aid of a monocular camera. We show that the estimation problem reduces to a specific parameter estimation of a perspective dynamical system. Surprisingly, the above reduction is independent of whether the data measured is the brightness pattern which the object produces on the image plane or whether the data observed are points, lines, or curves on the image plane produced as a result of discontinuities in the brightness pattern. Many cases of the perspective parameter estimation problem have been analyzed in this paper. These cases include a fairly complete analysis of a planar textured surface undergoing a rigid flow and an affine flow. These two cases have been analyzed for orthographic, pseudo-orthographic, and image-centered projections. The basic procedure introduced for parameter estimation is to subdivide the problem into two modules, one for "spatial averaging" and the other for "time averaging." The estimation procedure is carried out with the aid of a new "realization theory for perspective systems" introduced for systems described in discrete time and in continuous time. Finally, much of our analysis has been substantiated by computer simulation of specific algorithms developed in order to explicitly compute the parameters. Detailed simulation that would answer the perspective realizability question is a subject of future research.

Key words. perspective, vision, parameter identification

AMS subject classifications. 93B30, 93C10, 93C15, 93C60

1. Introduction.

The problem that we consider in this paper is described as follows.

PROBLEM 1. *We have a textured surface which is moving in continuous time following a certain vector field where we assume that both the shape of the surface and the vector field are unknown. Assume that a camera produces a perfect image of the textured surface in continuous time. The problem of interest is to estimate the shape and motion parameters of the surface from the observed time-varying image produced by the camera.*

Two important assumptions regarding the surface being observed, the camera, and its imaging mechanism need to be emphasized. First, we assume that the surface is constantly under focus, i.e., there is no blurring of the image as a result of imperfect focusing. Second, we assume that the photometric effects on the image due to the light source and the physical properties of the surface are negligible and can be ignored. Thus, the process of image formation is such that the intensity corresponding to each pixel on the surface is transferred to the image plane unattenuated via the projection process.

The existing approaches to the estimation problem in the literature can be divided broadly into two categories depending upon what is assumed to be measured from the scene. If the data observed is assumed to be the brightness pattern which the object produces on the image plane, a well-known approach in the literature is based on analyzing the optical flow field (see [1], [32], [33]). For a system theoretic treatment [2] of the subject we refer the reader to [47]. On the other hand, if the data observed are assumed to be the discontinuity curves in the brightness pattern on the image

[*] Received by the editors April 2, 1993; accepted for publication (in revised form) May 6, 1994. This research was supported in part by the Department of Energy under grant DE-FG02-90ER14140.

[†] Department of Systems Science and Mathematics, Washington University, Campus Box 1040, One Brookings Drive, St. Louis, Missouri 63130.

plane, a well-known feature-based approach is to identify the correspondence of various features such as points, lines, and curves between one frame and the next (see [3]–[6], [8], [10], [12], [13], [40]). The former approach assumes that the image intensity is a smooth function and restricts attention to the smooth part of the image plane only. The latter approach assumes that the image intensity is a piecewise smooth function and restricts attention to the region of the image plane wherein the image intensity is separated by a discontinuity curve. Of course for each of the two approaches, there are various projection models that one might want to consider. The two projection models well known in the literature are called "orthographic" and "perspective."

There are also other projection models (see [11]) that generalize orthographic and perspective projections. They are described as "image centered projection" and "viewer-centered projection." There are still other projection models in the literature [48] not considered in this paper. In this paper, we consider a model of projection (see equation (3.1)) that generalizes the various projection models considered in the literature. The generalized projection degenerates to orthographic, pseudo-orthographic, and perspective projection under various limiting cases. The corresponding estimates of the parameters also degenerate and these have been studied in detail in this paper. Before we describe the main contribution of this paper, we survey some of the important contributions in the field of motion parameter estimation.

The problem of estimating the motion parameters in computer vision has a long history, initiated by the early works of Ullman [9]. The problem was tested subsequently with real images by Roach and Aggarwal [16]. Finally Nagel [17] reduced the problem to solving a single nonlinear equation. A fairly complete analytical solution for eight feature points was given independently by Longuet-Higgins [18] and Tsai and Huang [21]. Zhuang [23], [24] proposed a simplified eight-point algorithm and discussed the uniqueness issue. On the question of uniqueness, Netravali et. al. [25] introduced a numerical technique called the homotopy method and showed the existence of 10 solutions. Using projective geometry, Faugeras and Maybank [7] showed that at most 10 solutions can be obtained from 5 feature points. Using the quaternion representation of three-dimesnional (3-D) rotation, Jerian and Jain [26] reduced the problem to solving the resultant of degree 16 of a pair of polynomials of degree 4 in 2 variables. Jerian and Jain [27] also compared known algorithms exhaustively and compared their performances with noisy data.

Many algorithms in the literature are known to perform poorly under noisy data. A robust algorithm was introduced by Weng, Huang, and Ahuja [28] and by Spetsakis and Aloimonos [14], [15]. They used optimization-based methods to compute "epipolar equations." Grzywacz and Hildreth [29] have also indicated that the effects of image noise on reconstruction from image velocities are severe in some cases. Jerian and Jain [26] and Murray and Buxton [30] proposed various schemes toward a stable reconstruction algorithm. The particular estimation problem has been summarized in two books by Maybank [31] and by Kanatani [11]. In fact, one of the reconstruction algorithms described in this paper has been initiated by Kanatani [11]. For some other related books and references we refer the reader to [39], [41], [45], [42], [43].

In this paper, we consider in detail the problem of estimating motion and shape parameters of a planar surface undergoing an affine motion. The proposed affine motion generalizes the rigid motion already considered in the literature (see [3], [17], [19]–[22]). While preserving the shape of the surface being observed, an affine motion adequately models many other nonrigid deformations. We also consider a generalized projection which includes as a special case both "image-centered projection"

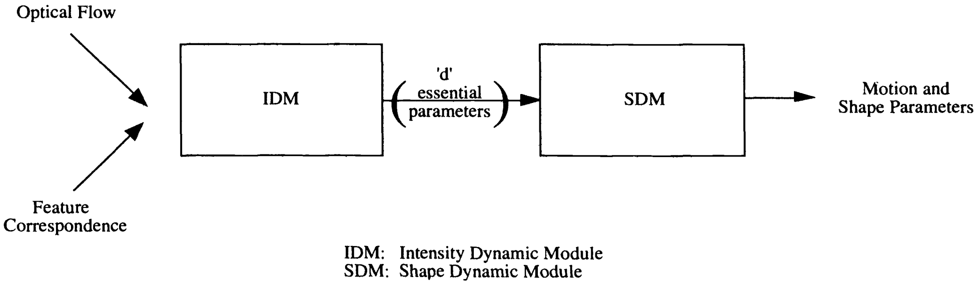

Optical Flow
Feature Correspondence
IDM
'd' essential parameters
SDM
Motion and Shape Parameters
IDM: Intensity Dynamic Module
SDM: Shape Dynamic Module

FIG. 1.1. *A two-module approach to parameter identification.*

and "viewer-centered projection," together with orthographic and perspective projections. Finally, we consider both the "optical flow analysis" (see [6], [32], [33]) and the "feature-based analysis" (see [35], [34], [44], [46], [40]) and show as the main contribution of this paper that irrespective of what is assumed to be the nature of the data observed (within the class of data considered), and regardless of what is assumed to be the projection model (within the chosen class of models), the problem of motion and shape estimation for a moving textured surface can always be analyzed as a specific parameter estimation problem of a perspective system. The specific form of the perspective system depends on how the surface and the motion field have been parameterized. It may be recalled that perspective systems have already been introduced in [36] in order to study feature-based estimation of motion parameters. Roughly speaking, a perspective system is a linear system with a homogeneous observation function (see [36]).

The details about the estimation scheme proposed in this paper are explained as follows. As shown in Fig. 1.1, the estimation problem is broken up into two modules, known as the Intensity Dynamic Module (IDM) and the Shape Dynamic Module (SDM). Data from the observed surface are first processed in the IDM in order to estimate a set of "essential parameters." Effectively, IDM performs a "spatial averaging" throughout the entire image plane from either the observed sequence of features or the optical flow data.

The essential parameters are functions of motion and shape parameters. The shape-dynamic module views them as an observation function corresponding to the "shape dynamics" introduced in this paper. The shape-dynamical system together with the essential parameters (viewed as an output) can be regarded as an example of a perspective system introduced in [36]. By observing the essential parameters over time, the SDM obtains an estimate of the motion and shape parameters.

Thus, via a dynamical systems approach, we characterize a complete set of identifiable parameters or functions of parameters for a planar surface undergoing an affine motion. Such a characterization is done both for a generalized projection (3.1) and for an orthographic projection (3.2). As a special case we consider the case when the motion is restricted to a rigid flow and recover many known results in the literature.

In summary, this paper introduces a new unified treatment of the estimation problem.

2. Shape dynamics of a surface patch. We assume throughout this paper that we have a textured surface patch which faces a camera without any occlusion. Futhermore, we assume that every point on the surface moves according to a certain differential equation. As a result of the motion of the individual points, the shape of the surface undergoes deformation while the surface moves in time. In this section, we write down a differential equation that describes the motion of the surface. We also specialize the equation to a planar surface patch undergoing affine motion and subsequently to a planar surface patch undergoing rigid motion.

Let us assume that (X, Y, Z) is the world coordinate frame wherein we have a surface defined by the equation

$$Z = S(X, Y, t). \tag{2.1}$$

We assume that S is smooth enough so that its derivatives with respect to each of the variables are defined everywhere. We now assume that the motion field is given by the equation

$$\dot{X} = f(X, Y, Z), \quad \dot{Y} = g(X, Y, Z), \quad \dot{Z} = h(X, Y, Z). \tag{2.2}$$

We now describe how the surface (2.1) moves as points on the surface move following the motion field (2.2). This is given by

$$\frac{\partial S}{\partial t} + f(X, Y, S)\frac{\partial S}{\partial X} + g(X, Y, S)\frac{\partial S}{\partial Y} = h(X, Y, S). \tag{2.3}$$

The equation (2.3) is a quasilinear partial differential equation and is called the "shape dynamics." We consider the initial condition

$$S(X, Y, 0) = \phi(X, Y). \tag{2.4}$$

The pair (2.3), (2.4) constitutes an example of a Riccati partial differential equation introduced in [38]. In this paper, we shall assume that the surface (2.1) is a plane described as

$$Z = pX + qY + r, \tag{2.5}$$

where p, q, r are shape parameters that are changing in time as a result of the motion field (2.2). Furthermore we shall also assume that the motion field (2.2) is affine and is given by

$$\dot{\mathcal{X}} = A\mathcal{X} + b, \tag{2.6}$$

where

$$A = [a_{ij}], b = \mathrm{col}[b_1, b_2, b_3] \tag{2.7}$$

are respectively a 3×3 matrix and a 3×1 vector and where $\mathcal{X} = \mathrm{col}\,[X, Y, Z]$. Thus in this paper, we do not assume that the shape undergoes any deformation as a result of the motion field. We now construct a differential equation that describes the motion of the shape parameters p, q, r. This is done as follows. Let us homogenize the vector (X, Y, Z) as $X = \bar{X}/\bar{W}$, $Y = \bar{Y}/\bar{W}$, $Z = \bar{Z}/\bar{W}$ and the vector (p, q, r) as

$$p = \bar{p}/\bar{s}, \quad q = \bar{q}/\bar{s}, \quad r = \bar{r}/\bar{s}. \tag{2.8}$$

We rewrite (2.5) as $\begin{pmatrix} \bar{p}, & \bar{q}, & -\bar{s}, & \bar{r} \end{pmatrix} \bar{\mathcal{X}} = 0$ and (2.6) as $\dot{\bar{\mathcal{X}}} = -\mathcal{A}^T \bar{\mathcal{X}}$ where $\bar{\mathcal{X}} = \begin{pmatrix} \bar{X}, & \bar{Y}, & \bar{Z}, & \bar{W} \end{pmatrix}^T$ and

$$-\mathcal{A}^T = \begin{pmatrix} A & b \\ 0 & 0 \end{pmatrix}. \tag{2.9}$$

It follows that

$$\frac{d}{dt}\left(\ \bar{p},\ \ \bar{q},\ \ -\bar{s},\ \ \bar{r}\ \right)^T = \mathcal{A}\left(\ \bar{p},\ \ \bar{q},\ \ -\bar{s},\ \ \bar{r}\ \right)^T, \tag{2.10}$$

where $\mathcal{A}$ is the 4×4 matrix in (2.9) and is defined up to addition by a scalar multiple of the identity matrix. If we assume initial condition to be $\bar{s}(0) = 1$, $\bar{p}(0) = p(0)$, $\bar{q}(0) = q(0)$, $\bar{r}(0) = r(0)$, it may be concluded that the dynamical system (2.10) describes the motion of the shape parameters p, q, r. In fact, from (2.8) and (2.10) the dynamics of p, q, r can be written as the following Riccati equation:

$$\begin{aligned} \dot{p} &= (a_{33} - a_{11})p - a_{21}q + a_{31} - a_{13}p^2 - a_{23}pq, \\ \dot{q} &= (a_{33} - a_{22})q - a_{12}p + a_{32} - a_{13}pq - a_{23}q^2, \\ \dot{r} &= -(a_{33} + a_{23}q + a_{13}p)r + (b_3 - b_2q - b_1p). \end{aligned} \tag{2.11}$$

In general, Riccati equation (2.3) or (2.11) propagates in time the relationship between coordinates X, Y, and Z expressed via the surface (2.1) or the plane (2.5). Note that the equation (2.11) is parameterized by 12 motion parameters and 3 initial conditions on shape parameters. Thus there is a total of 15 parameters describing the shape dynamics (2.10) for the affine motion.

An important special case of the affine motion (2.6) is the case when A is a skew symmetric matrix given by

$$\begin{pmatrix} 0 & \omega_1 & \omega_2 \\ -\omega_1 & 0 & \omega_3 \\ -\omega_2 & -\omega_3 & 0 \end{pmatrix} \triangleq \Omega. \tag{2.12}$$

Under this assumption, the motion field (2.6) describes a rigid motion. The shape dynamics (2.10) can be written as

$$\frac{d}{dt}\begin{pmatrix} \bar{p} \\ \bar{q} \\ -\bar{s} \\ \bar{r} \end{pmatrix} = \begin{pmatrix} \Omega & 0 \\ -b^T & 0 \end{pmatrix}\begin{pmatrix} \bar{p} \\ \bar{q} \\ -\bar{s} \\ \bar{r} \end{pmatrix}. \tag{2.13}$$

Note that the shape dynamics (2.11) reduces to $\dot{p} = -\omega_2(1 + p^2) + \omega_1 q - \omega_3 pq$, $\dot{q} = -\omega_3(1 + q^2) - \omega_1 p - \omega_2 pq$, and $\dot{r} = b_3 - b_1 p - b_2 q - r(\omega_3 q + \omega_2 p)$ which is parameterized by a total of six motion parameters and three initial conditions on shape parameters. Thus there is a total of nine parameters describing the shape dynamics (2.13) for the rigid motion.

3. Intensity dynamics of a moving textured surface. Assume that the surface described by (2.1) is textured, i.e., the intensity $E(X, Y, Z, t)$ of a point (X, Y, Z) on the surface at time t does not change along the integral curves of (2.2). We also assume that the camera is perfectly focused on the object surface, i.e., intensity from a surface on the object to the image plane is transferred unattenuated under the camera correspondence. The above two assumptions together imply that the intensity on the image plane does not change along the projection of the integral curves of (2.2). In this paper we consider the projection to be described as follows.

Let (x, y) be the coordinates of the image plane obtained under the projection of a point (X, Y, Z) on the surface of the object. We define

$$x = \frac{fX}{Z + \delta}, \quad y = \frac{fY}{Z + \delta}, \tag{3.1}$$

where $\delta \in [0, f]$ and f is the focal length of the camera. Note that if $\delta = 0$ we obtain a viewer-centered projection. If $\delta = f$ we obtain an image-centered projection. These two projections have been described in [11]. Finally note that if $\delta = f$ and $f \rightarrow \infty$ we obtain

$$x = X, \quad y = Y \tag{3.2}$$

which is known in the literature [11] as the "orthographic projection."

In an orthographic projection, a point (X, Y, Z) is projected by dropping the Z coordinate information. In order to motivate the image-centered and viewer-centered projections, assume that the image plane is perpendicular to the Z axis and passes through the point $Z = a$. Assume that the optical axis is the Z axis and a point (X, Y, Z) is projected onto the image plane via the center of the camera located at $Z = -Z_0$. In order to derive the projected point, one computes the line l passing through the points (X, Y, Z) and $(0, 0, -Z_0)$ and computes the intersection of l with the image plane. The projection of the point (X, Y, Z) is this intersection. If the center of the camera is the origin of the coordinate axis, i.e., if $Z_0 = 0$, we obtain a viewer-centered projection. On the other hand, if we assume that the image plane passes through the origin of the coordinate axis, i.e., if $a = 0$, we obtain an image-centered projection.

For a given fixed value of f, δ we have a new set of coordinates (x, y, Z). We now rewrite the shape equation (2.1) and the restriction of the motion field (2.2) on the image plane in the new set of coordinates as

$$Z = \tilde{S}(x, y, t) \tag{3.3}$$

and

$$\dot{x} = \tilde{f}(x, y, \tilde{S}(x, y, t)), \quad \dot{y} = \tilde{g}(x, y, \tilde{S}(x, y, t)) \tag{3.4}$$

for some suitable functions $\tilde{S}$, $\tilde{f}$, $\tilde{g}$.

The integral curves of (3.4) are exactly the projection of the integral curves of the motion field under the generalized projection (3.1). The vector field described by (3.4) has been described in the literature (see Horn [1]) as "optical flow." Note in particular that the optical flow is in general a time-varying dynamical system described via the coordinates of the image plane. The time variation of the optical flow is a result of the motion of the surface (2.1), or equivalently (3.3).

Let $e(x, y, t)$ be the intensity of a point (x, y) on the image plane at time instant t. Since $e(x, y, t)$ does not change along the integral curves of (3.4), it follows that $e(x, y, t)$ satisfies the partial differential equation given by

$$\frac{\partial e}{\partial t} + \tilde{f}(x, y, \tilde{S}(x, y, t))\frac{\partial e}{\partial x} + \tilde{g}(x, y, \tilde{S}(x, y, t))\frac{\partial e}{\partial y} = 0. \tag{3.5}$$

We shall call the dynamical system (3.5) as "intensity dynamics." Let us now assume that the initial condition is given by

$$e(x, y, 0) = \Psi(x, y). \tag{3.6}$$

We shall call the function $\Psi(x, y)$ the "texture function." The above pair (3.5), (3.6) is a linear partial differential equation, which describes the dynamics of the intensity function on the image plane.

Let us now restrict our attention to a planar surface (2.5) with affine motion (2.6) and assume a generalized projection (3.1). The "optical flow" equation for this special case can be written as follows:

$$\begin{aligned} \dot{x} &= d_1 + d_3x + d_4y + \frac{1}{f}\left(d_7x^2 + d_8xy\right), \\ \dot{y} &= d_2 + d_6y + d_5x + \frac{1}{f}\left(d_8y^2 + d_7xy\right), \end{aligned} \tag{3.7}$$

where

$$\begin{aligned} &d_1 = f(a_{13} + c_1), d_2 = f(a_{23} + c_2), d_3 = (a_{11} - a_{33}) - (c_3 + pc_1), \\ &d_4 = a_{12} - qc_1, d_5 = a_{21} - pc_2, \\ &d_6 = (a_{22} - a_{33}) - (c_3 + qc_2), d_7 = pc_3 - a_{31}, d_8 = qc_3 - a_{32} \end{aligned} \tag{3.8}$$

and where

$$c_i = (b_i - a_{i3}\delta)/(r + \delta), i = 1, 2, 3. \tag{3.9}$$

Various limits of the optical flow equation have been considered in the literature. They all pertain to analyzing what happens when f tends to ∞, assuming $f = \delta$. In the process of taking the limit, one would approximate the coefficients of the optical flow equation (3.7) up to order $\frac{1}{f}$, while neglecting the higher-order terms. If we define

$$h_j = \lim_{f\to\infty} d_j; \quad j = 1, 2, \ldots, 8 \tag{3.10}$$

we obtain the following:

$$\begin{aligned} &h_1 = a_{13}r + b_1, h_2 = a_{23}r + b_2, \\ &h_3 = a_{11} + a_{13}p, h_4 = a_{12} + a_{13}q, \\ &h_5 = a_{21} + a_{23}p, h_6 = a_{22} + a_{23}q, \\ &h_7 = -a_{31} - a_{33}p, h_8 = -a_{32} - a_{33}q. \end{aligned} \tag{3.11}$$

Thus when $f \to \infty$ and $f = \delta$, the optical flow equation can be approximated up to order $\frac{1}{f}$ by

$$\begin{aligned} \dot{x} &= h_1 + h_3x + h_4y + \frac{1}{f}(h_7x^2 + h_8xy), \\ \dot{y} &= h_2 + h_5x + h_6y + \frac{1}{f}(h_8y^2 + h_7xy). \end{aligned} \tag{3.12}$$

Of course if the focal length of the camera is fixed at ∞, one observes the optical flow equation as

$$\dot{x} = h_1 + h_3x + h_4y, \dot{y} = h_2 + h_5x + h_6y. \tag{3.13}$$

The projection which produces the optical flow given by (3.13) is known as "orthographic projection." Such a projection described by (3.2) does not give any information about the quadratic component d_7 and d_8 of the optical flow (3.7) in general. The optical flow equation (3.12), on the other hand, is an approximation of (3.7) up to order $\frac{1}{f}$ assuming f is approaching ∞. Thus if the focal length of a camera can be varied, one can obtain the asymptotic values of d_7 and d_8 for large f and use this information to compute h_7 and h_8. We shall call (3.12) the optical flow under "orthographic approximation," as opposed to (3.13), which is the optical flow under "orthographic projection."

We also introduce a "pseudo-orthographic approximation" of (3.7) originally introduced by Kanatani [11]. This is described as follows:

$$\begin{aligned} \dot{x} &= d_1 + d_3 x + d_4 y + \frac{1}{f}\left(h_7 x^2 + h_8 xy\right), \\ \dot{y} &= d_2 + d_6 y + d_5 x + \frac{1}{f}\left(h_8 y^2 + h_7 xy\right). \end{aligned} \tag{3.14}$$

"Orthographic approximation" and "pseudo-orthographic approximation" to the optical flow equation (3.7) is useful in the process of reconstructing the motion and shape parameters from the coefficients of the optical flow equation. The reconstruction algorithm has been described in §5 using an approach described by Kanatani [11].

4. Estimation of essential parameters based on intensity and feature measurements. Assume as in §3 that we have a moving textured plane which produces a time-varying intensity profile on the image plane. In this section we consider the intensity dynamic module problem described as follows.

PROBLEM 2 (intensity dynamic module problem). *Assume that the intensity function $e(x, y, t)$ is measured in a given region of the image plane over a given interval of time. The problem is to estimate the vector $(d_1, \ldots, d_8)$ from this data.*

In subsequent sections, we shall see that the vector $(d_1, \ldots, d_8)$ is of paramount importance in estimating the motion and shape parameters. For this reason we shall call the vector $(d_1, \ldots, d_8)$ the "vector of essential parameters."

4.1. Estimation based on intensity measurements. Assume that the intensity function is smooth so that all its partial derivatives exist and can be computed. If the motion field is affine given by (2.6), it follows from (3.5), (3.7) that the intensity dynamics is given by

$$\frac{\partial e}{\partial t} + F(x, y)\frac{\partial e}{\partial x} + G(x, y)\frac{\partial e}{\partial y} = 0, \tag{4.1}$$

where $e(x, y, t)$ is the observed intensity function on the image plane and

$$\begin{aligned} F(x, y) &= d_1 + d_3 x + d_4 y + \frac{1}{f}\left(d_7 x^2 + d_8 xy\right), \\ G(x, y) &= d_2 + d_6 y + d_5 x + \frac{1}{f}\left(d_8 y^2 + d_7 xy\right). \end{aligned} \tag{4.2}$$

The parameters $d_1, \ldots, d_8$ can be defined from (3.8). Combining (4.1) and (4.2), we now write

$$v^T d = -\frac{\partial e}{\partial t}, \tag{4.3}$$

where

$$v^T = \left(e_x, e_y, xe_x, ye_x, xe_y, ye_y, \frac{1}{f}\left(x^2 e_x + xye_y\right), \frac{1}{f}\left(xye_x + y^2 e_y\right)\right) \tag{4.4}$$

and

$$d = (d_1, \dots, d_8)^T . \tag{4.5}$$

In order to compute an estimate of the coefficient vector d, we proceed as follows. Choose $n \geq 8$ points on the image plane denoted by $(x_i, y_i), i = 1, \dots, n$. From the observed data $e(x, y, t)$ we now form the matrices

$$V = \begin{pmatrix} v(x_1, y_1) & v(x_2, y_2) & . & . & . & v(x_n, y_n) \end{pmatrix} \tag{4.6}$$

and

$$u = \begin{pmatrix} -e_t(x_1, y_1) & -e_t(x_2, y_2) & . & . & . & -e_t(x_n, y_n) \end{pmatrix}^T \tag{4.7}$$

From (4.3) it follows that $V^T d = u$. If the points (x_i, y_i) are chosen in such a way that rank $V = 8$, we compute

$$\hat{d} = (VV^T)^{-1} V u \tag{4.8}$$

as an estimate of d. We therefore have the following theorem.

THEOREM 4.1. *Assume that the function* $e(x, y, t)$ *is such that all its partial derivatives are available and can be measured. Assume furthermore that the points* $(x_i, y_i), i = 1, \dots, n$ *are such that rank* $V = 8$, *where* V *is given by* (4.6). *It is possible to obtain a unique estimate of* d.

4.2. Estimation based on feature measurements: Curve correspondence. By the word "feature" we shall mean points or curves of discontinuity for the intensity function $e(x, y, t)$. We shall assume that, via edge detection, these features can be observed in real time. We shall assume that the moving textured surface produces a time-varying intensity function on the screen. The moving intensity function in turn would make the features move on the screen. The dynamical system which describes such a motion is called "feature dynamics." The main result of this section is to see that the coefficients of the feature dynamics are exactly the essential parameters introduced in (3.8). Thus under an appropriate technical condition, the essential parameters can be estimated from the feature dynamics as well, as was the case for intensity dynamics. In order to describe the feature dynamics we proceed as follows.

Let

$$y = \mathcal{I}(x, t) \tag{4.9}$$

be the curve along which the function $e(x, y, t)$ is discontinuous. We want to study how the feature curve (4.9) changes in time. Differentiating (4.9) with respect to time, we obtain

$$\dot{y} = \frac{\partial \mathcal{I}}{\partial x} \dot{x} + \frac{\partial \mathcal{I}}{\partial t}. \tag{4.10}$$

Recall that

$$\begin{aligned} \dot{x} &= F(x, y), \\ \dot{y} &= G(x, y), \end{aligned} \tag{4.11}$$

where $F(x,y), G(x,y)$ are given in (4.2). It follows that

$$\begin{aligned}\frac{\partial \mathcal{I}}{\partial t} + \frac{\partial \mathcal{I}}{\partial x}\left[d_1 + d_3x + d_4\mathcal{I}(x,t) + \frac{1}{f}(d_7x^2 + d_8x\mathcal{I}(x,t))\right] \\ = d_2 + d_6\mathcal{I}(x,t) + d_5x + \frac{1}{f}\left(d_8\mathcal{I}(x,t)^2 + d_7x\mathcal{I}(x,t)\right. .\end{aligned} \tag{4.12}$$

The above equation (4.12) is referred to as the feature dynamics, which can be rewritten as

$$v^Td = -\frac{\partial \mathcal{I}}{\partial t}, \tag{4.13}$$

where d is defined as in (4.5) to be the vector of essential parameters. The vector v^T is given by

$$v^T = \left(\mathcal{I}_x, -1, x\mathcal{I}_x, \mathcal{I}\mathcal{I}_x, -x, -\mathcal{I}, \frac{1}{f}\left(x^2\mathcal{I}_x - x\mathcal{I}\right), \frac{1}{f}\left(-\mathcal{I}^2 + x\mathcal{I}\mathcal{I}_x\right)\right). \tag{4.14}$$

We now choose $n \geq 8$ points on the curve (4.9) denoted by $(x_i, y_i), i = 1, \ldots, 8$. As in (4.6), (4.7) we construct the matrix V and vector u and obtain an estimate $\hat{d}$ of d given by (4.8), provided of course rank $V = 8$.

In order for the matrix V to have rank 8, the curve (4.9) has to be of sufficiently high order. In fact, if (4.9) is a polynomial, it cannot be of degree < 4. On the other hand, if

$$\mathcal{I}(x,t) = a_0 + a_1x + a_2x^2 + a_3x^3 + a_4x^4, a_4 \neq 0 \tag{4.15}$$

in order for rank $V = 8$, one must have

$$a_3^2 \neq \frac{8}{3}a_2a_4. \tag{4.16}$$

Thus we have essentially proved the following theorem.

THEOREM 4.2. *Assume that the observed feature is a polynomial discontinuity curve* (4.9) *of degree* 4 *given by* (4.15). *It is possible to estimate* $\hat{d}$ *given by* (4.8) *iff* (4.16) *is satisfied.*

If the observed discontinuity curve is of degree < 4, we shall see that one needs to observe a larger number of features in order for rank $V = 8$. Two cases of interest are when the observed feature is a line and when it is a point. These two subcases are now considered.

4.3. Estimation based on line correspondence.

Let

$$y = ax + b \tag{4.17}$$

be the line along which the function $e(x,y,t)$ is discontinuous. Assume furthermore that the line (4.17) is generated as a result of a discontinuity in the texture of the surface (2.5). We also assume that changes in x, y are given by (3.7). Thus, the feature dynamics is given by (4.12) or (4.13) where

$$\frac{\partial \mathcal{I}}{\partial t} = \dot{a}x + \dot{b} \tag{4.18}$$

and

$$v^T = \left(a, -1, ax, a(ax+b), -x, -(ax+b), -\frac{b}{f}x, -\frac{b}{f}(ax+b)\right). \tag{4.19}$$

The vector d of essential parameters (see (4.5)) satisfies the ordinary differential equation

$$\begin{pmatrix} -a & 1 & 0 & -ab & 0 & b & 0 & \frac{b^2}{f} \\ 0 & 0 & -a & -a^2 & 1 & a & \frac{b}{f} & \frac{ab}{f} \end{pmatrix} d = \begin{pmatrix} \dot{a} \\ \dot{b} \end{pmatrix}. \tag{4.20}$$

If we assume that the motion of the line (4.17) is observed, we might infer that in (4.20), a, b, $\dot{a}$, $\dot{b}$ is observed. Thus (4.20) represents a pair of equations in eight variables, the variables being the eight-parameter d vector. Choosing a set of four lines on the surface being observed and assuming that these four lines define a set of eight independent conditions on the d vector, one can obtain an unique estimate of the d vector. The procedure is similar to that outlined in §4.1 and described by (4.8). We now state the following theorem.

THEOREM 4.3. *Assume that the observed feature is a set of four lines on the image plane given by the equation*

$$y = a_i x + b_i, \quad i = 1, \ldots, 4, \tag{4.21}$$

where the lines (4.21) *are generated as a result of discontinuity in the texture of the surface* (2.5). *Define*

$$\phi_i = \begin{pmatrix} -a_i & 1 & 0 & -a_i b_i & 0 & b_i & 0 & \frac{b_i^2}{f} \\ 0 & 0 & -a_i & -a_i^2 & 1 & a_i & \frac{b_i}{f} & \frac{a_i b_i}{f} \end{pmatrix} \tag{4.22}$$

$i = 1, \ldots, 4$ *and the* 8×8 *matrix* $\phi = \left(\phi_1^T \; \phi_2^T \; \phi_3^T \; \phi_4^T\right)^T$. *It is possible to estimate the vector* d *uniquely given by*

$$\hat{d} = \left(\phi^T \phi\right)^{-1} \phi^T \left(\dot{a}_1 \quad \dot{b}_1 \quad \dot{a}_2 \quad \dot{b}_2 \quad \dot{a}_3 \quad \dot{b}_3 \quad \dot{a}_4 \quad \dot{b}_4 \right)^T \tag{4.23}$$

iff rank $\phi = 8$.

4.4. Estimation based on point correspondence. If we assume that the texture function is discontinuous at a single point, one would observe this point as a discontinuity in the function $e(x, y, t)$. Tracking the discontinuity in real time would amount to tracking the projection of the feature point on the image plane. Thus we rewrite the optical flow (3.7) as

$$\begin{pmatrix} 1 & 0 & x & y & 0 & 0 & \frac{x^2}{f} & \frac{xy}{f} \\ 0 & 1 & 0 & 0 & x & y & \frac{xy}{f} & \frac{y^2}{f} \end{pmatrix} d = \begin{pmatrix} \dot{x} \\ \dot{y} \end{pmatrix}, \tag{4.24}$$

where d is once again the vector of essential parameters given by (4.5). The point (x, y) is the projection of the feature point on the image plane. Assuming that we are able to observe $x, y, \dot{x}, \dot{y}$ in real time, it follows that equation (4.24) represents a pair of equations in eight variables, the variables being the eight-parameter d vector of essential parameters. As in §4.3, if we choose a set of four feature points on the image plane that are projections of points of discontinuity in the texture of the surface (2.5), and assume that they define a set of eight independent conditions on the vector d, it follows that one can uniquely obtain an estimate of the vector d. Thus we have the following theorem.

THEOREM 4.4. *Assume that the observed feature is a set of four points on the image plane given by* (x_i, y_i), $i = 1, \ldots, 4$, *where we assume that the points are generated as a result of discontinuity in the texture of the surface* (2.5). *Assume furthermore that the* 8×8 *matrix*

$$\psi = \left(\psi_1^T \ \psi_2^T \ \psi_3^T \psi_4^T\right)^T \tag{4.25}$$

is nonsingular, where

$$\psi_i = \begin{pmatrix} 1 & 0 & x_i & y_i & 0 & 0 & \frac{x_i^2}{f} & \frac{x_i y_i}{f} \\ 0 & 1 & 0 & 0 & x_i & y_i & \frac{x_i y_i}{f} & \frac{y_i^2}{f} \end{pmatrix}, \tag{4.26}$$

$i = 1, 2, 3, 4$. *It is possible to estimate the vector* d *uniquely given by*

$$\hat{d} = \left(\psi^T \psi\right)^{-1} \psi^T \begin{pmatrix} \dot{x}_1 & \dot{y}_1 & \dot{x}_2 & \dot{y}_2 & \dot{x}_3 & \dot{y}_3 & \dot{x}_4 & \dot{y}_4 \end{pmatrix}^T. \tag{4.27}$$

To summarize the main results of this section, we show that the vector d of essential parameters can be estimated from intensity and feature measurements. The task of the IDM is to estimate the vector d. It may be noted that the IDM requires information only at a given instant of time and performs "spatial averaging."

5. Estimating motion and shape parameters from the recovery equation. In this section we shall assume that the essential parameter vector d has already been estimated by the intensity dynamic module. The problem that we would like to consider is to solve (3.8) for the motion and shape parameters. We would also like to study how the solution degenerates for $f = \delta$ as $f \to \infty$, i.e., when the projection model degenerates to that produced by orthographic projection. Some portion of our analysis in this section is an adaptation of earlier work due to Kanatani [11].

5.1. Estimation under general projection. We assume that we have a planar surface (2.5) undergoing a rigid motion (2.13). The essential parameter vector d given by (3.8) for this case is given as follows:

$$\begin{aligned} &d_1 = f(\omega_2 + c_1), \quad d_2 = f(\omega_3 + c_2), \quad d_3 = -(c_3 + pc_1), \quad d_4 = \omega_1 - qc_1, \\ &d_5 = -\omega_1 - pc_2, \quad d_6 = -(c_3 + qc_2), \quad d_7 = (\omega_2 + pc_3), \quad d_8 = (\omega_3 + qc_3), \end{aligned} \tag{5.1}$$

where

$$c_1 = (b_1 - \omega_2 \delta)/(r + \delta), \quad c_2 = (b_2 - \omega_3 \delta)/(r + \delta), \quad c_3 = b_3/(r + \delta). \tag{5.2}$$

The problem that we consider is described as follows.

PROBLEM 3. *Assume that we are given* $(d_1, \ldots, d_8)$. *Using the algebraic equation* (5.1), (5.2), *solve for the parameters* $c_1, c_2, c_3, \omega_1, \omega_2, \omega_3, p, q$.

It may be noted that (5.1) describes exactly a set of eight nonlinear equations in eight parameters. This particular set of equations is known as the "recovery equation." The following result is quite surprising, however.

THEOREM 5.1. *Assume* $c_3 \neq 0$; *then* (5.1) *can be solved for exactly two real solutions. If*

$$(c_1, c_2, c_3, \omega_1, \omega_2, \omega_3, p, q) \tag{5.3}$$

is one solution, then the other solution is given by

$$(-c_3p,\ -c_3q,\ c_3,\ \omega_1 - c_1q + c_2p,\ \omega_2 + c_1 + c_3p,\ \omega_3 + c_2 + c_3q,\ -c_1/c_3,\ -c_2/c_3)\,. \tag{5.4}$$

It may be remarked that the existence of two solutions to the recovery equation (5.1) and described by Theorem 5.1 has been reported earlier in the literature by Waxman and Ullman [8] and by Kanatani [11]. In [8] the analytical steps leading up to the two solutions have not been documented. In [11] the analytical formula (5.4) of the two solutions has not been presented. The purpose of stating and proving Theorem 5.1 is therefore tutorial.

Before we prove Theorem 5.1, we proceed to solve the set of equations (5.1). Let us define

$$\begin{aligned} &T = d_3 + d_6, && R = d_5 - d_4, && U_0 = d_1 + id_2,\\ &K = \frac{1}{f}(d_7 + id_8), && S = d_3 - d_6 + i(d_4 + d_5), && \end{aligned} \tag{5.5}$$

and

$$P = p + iq, \quad V = c_1 + ic_2, \quad W = \omega_3 - i\omega_2, \quad L = fK - \frac{1}{f}U_0. \tag{5.6}$$

The equations (5.1) can be written as

$$\begin{aligned} U_0 &= f(V + iW),\\ S &= -PV, L = c_3P - V,\\ -iPV^* &= R + 2\omega_1 + i(T + 2c_3). \end{aligned} \tag{5.7}$$

Note that (5.7) is a set of four equations in complex variables that needs to be solved. From (5.7) we have

$$V^2 + LV + c_3S = 0. \tag{5.8}$$

Solving (5.8) for V and then using (5.7) for P we have

$$V = \frac{-L \pm \sqrt{L^2 - 4c_3S}}{2}, \tag{5.9}$$

$$P = \frac{L \pm \sqrt{L^2 - 4c_3S}}{2c_3}. \tag{5.10}$$

From (5.7) we have

$$\omega_1 = [Im(PV^*) - R]/2, \tag{5.11}$$

$$T + 2c_3 = -Re(PV^*). \tag{5.12}$$

From (5.9) and (5.10) we have

$$Re(PV^*) = \frac{-|L|^2 + \sqrt{(L^2 - 4c_3S)(L^2 - 4c_3S)^*}}{4c_3}. \tag{5.13}$$

Combining (5.12) and (5.13) we have

$$|L|^2 - 4Tc_3 - 8c_3^2 = \sqrt{|L|^4 + 16c_3^2|S|^2 - 8c_3Re(L^2S^*)}. \tag{5.14}$$

Note that (5.14) as an equation in c_3 has two solutions. One solution is at $c_3 = 0$ and the other solution is at $c_3 = c_3^*$. Squaring (5.14) on both sides, we conclude that c_3^* is the middle root of the cubic equation

$$c_3^3 + Tc_3^2 + \frac{1}{4}(T^2 - |L|^2 - |S|^2)c_3 + \frac{1}{8}(Re(L^2S^*) - T|L|^2) = 0 \tag{5.15}$$

Using c_3, one can solve for a pair of solutions for P and V from (5.9) and (5.10). Finally, from (5.7) we have

$$W = i\left(V - \frac{1}{f}U_0\right) \tag{5.16}$$

and from (5.11) one can solve for ω_1. Thus the set of equations (5.7) can be solved for exactly two distinct solutions if $c_3 \neq 0$. If (5.1) is solved, these are exactly the two solutions that one would obtain.

Proof of Theorem 5.1. It can be easily checked that if (5.3) is one solution of (5.1), then the other solution is given by (5.4). However, since (5.1) has exactly two solutions, these are the only solutions. Moreover the solutions are obtained by solving the cubic polynomial equation (5.15) outlined as above.

From the two solutions to the recovery equation (5.7), it is easy to see what happens when $f \to \infty$. Note that

$$\lim_{f\to\infty} c_1 = -\omega_2, \quad \lim_{f\to\infty} c_2 = -\omega_3, \quad \lim_{f\to\infty} c_3 = 0. \tag{5.17}$$

It follows that one of the two solutions $(c_1, c_2, c_3, \omega_1, \omega_2, \omega_3, p, q)$ approaches the vector

$$(-\omega_2,\ -\omega_3,\ 0,\ \omega_1,\ \omega_2,\ \omega_3,\ p,\ q), \tag{5.18}$$

the first six components of the other solution approach the vector

$$(0,\ 0,\ 0,\ \omega_1 + \omega_2 q - \omega_3 p,\ 0,\ 0) \tag{5.19}$$

and the last two components of the other solution approach ∞ asymptotically along the line

$$p/q = \omega_2/\omega_3. \tag{5.20}$$

The parameters b_1, b_2, b_3, r are never recovered exactly. In fact, from the definition of d_1, d_2, c_3 we have, for a given f, the straight line

$$\omega_2 r + b_1 = \left(1 + \tfrac{r}{f}\right) d_1, \quad \omega_3 r + b_2 = \left(1 + \tfrac{r}{f}\right) d_2, \quad b_3 = c_3(r + f) \tag{5.21}$$

described in the (b_1, b_2, b_3, r) space corresponding to the solution $(c_1, c_2, c_3, \omega_1, \omega_2, \omega_3, p, q)$. On the other hand, corresponding to the other solution we have the straight line

$$\begin{aligned}(\omega_2 + c_1 + c_3 p)r + b_1 &= \left(1 + \frac{r}{f}\right) d_1,\\ (\omega_3 + c_2 + c_3 q)r + b_2 &= \left(1 + \frac{r}{f}\right) d_2,\\ b_3 &= c_3(r + f).\end{aligned} \tag{5.22}$$

As $f \to \infty$, the straight line (5.21) tends to the straight line

$$\omega_2 r + b_1 = h_1, \omega_3 r + b_2 = h_2, b_3 = b_3^*, \tag{5.23}$$

where b_3^* is an arbitrary constant. To see (5.23) we need the following lemma.

LEMMA 5.2. *In the (b_3, r) space the straight line $b_3 = c_3(r+f)$ converges to the line $b_3 = b_3^*$ as $f \to \infty$, where b_3^* is an arbitrary constant.*

Proof. Recall that $d_3 = -c_3 - pc_1$, i.e.,

$$(d_3 + pc_1)r + b_3 = -(d_3 + pc_1)f. \tag{5.24}$$

As $f \to \infty$, we have $(d_3 + pc_1) \to 0$ and $(b_3 + (d_3 + pc_1)f) \to 0$. At a given f, the line (5.24) passes through the point $(0, -f)$ and $(-(d_3 + pc_1)f, 0)$. For large f, the line passes closely through the points $(0, -f)$ and $(b_3^*, 0)$ where b_3^* is a fixed constant, which is also the true value of b_3. Thus as $f \to \infty$ the line (5.24) approaches the line $b_3 = b_3^*$. □

The above calculation can be summarized via the following theorem.

THEOREM 5.3. *Consider the solution vector $(\omega_1, \omega_2, \omega_3, p, q)$ for the recovery equation* (5.7). *For a given fixed f there are exactly two solutions, one of which remains unchanged as $f \to \infty$ and the other of which goes off to infinity as described by* (5.19), (5.20). *For the parameter vector (b_1, b_2, b_3, r), the recovery equation specifies these parameters up to a choice of two straight lines* (5.21) *and* (5.22). *The line* (5.21) *corresponds to the parameter vector $(\omega_1, \omega_2, \omega_3, p, q)$, which does not change with f. Moreover as $f \to \infty$, the line* (5.21) *changes with f and approaches the limit* (5.23).

Remark. It follows from Theorem 5.3 that for large f one recovers (b_1, b_2, r) up to a line given by (5.23) and b_3 exactly.

5.2. Estimation under pseudo-orthographic approximation. Under the pseudo-orthographic approximation, the equation we need to solve for instead of (5.1) is given by

$$\begin{array}{llll} d_1 = f(\omega_2 + c_1), & d_2 = f(\omega_3 + c_2), & d_3 = -(c_3 + pc_1), & d_4 = \omega_1 - qc_1, \\ d_5 = -\omega_1 - pc_2, & d_6 = -(c_3 + qc_2), & h_7 = \omega_2, & h_8 = \omega_3. \end{array} \tag{5.25}$$

Let us define T, R, U_0, S as in (5.5) and replace K by K_1 given by $K_1 = \frac{1}{f}(h_7 + ih_8)$. Furthermore let us define P, V, W as in (5.6) and replace L by L_1 given by $L_1 = fK_1 - \frac{1}{f}U_0$. The recovery equation (5.25) can be written as $U_0 = f(V + iW)$, $S = -PV$, $L_1 = -V$, and $-iPV^* = (R + 2\omega_1) + i(T + 2c_3)$, which can be easily solved (see [11]) and the solution is given by

$$\begin{array}{rl} V = & -L_1, \\ P = & S/L_1, \\ \omega_1 = & -[Im(SL_1^*/L_1) + R]/2, \\ W = & i\left(V - \dfrac{1}{f}U_0\right). \end{array} \tag{5.26}$$

The following theorem describes an important property of the pseudo-orthographic approximation.

THEOREM 5.4. *The solution* (5.26) *of the pseudo-orthographic approximation, converges as $f \to \infty$ to one of the solution of the recovery equation* (5.7), *described by* (5.9), (5.10), (5.11), *and* (5.16). *The solution to which* (5.26) *converges to is exactly the one which does not change with f.*

Proof of Theorem 5.4. It is easy to see from (5.9), (5.10) that

$$\lim_{c_3 \to 0} \frac{-L - \sqrt{L^2 - 4c_3 S}}{2} = -L,$$

$$\lim_{c_3 \to 0} \frac{L - \sqrt{L^2 - 4c_3 S}}{2c_3} = S/L.$$

If $f \to \infty$ it follows that $c_3 \to 0$. Thus it may be concluded that if $f \to \infty$, the solution (5.26) approaches one of the two solutions of the recovery equation (5.7). Finally note that as $f \to \infty$, (5.26) remains finite. To see this we compute

$$\tag{5.27} \begin{aligned} &\bar{V} = -\lim_{f\to\infty} L_1 = -(h_7 + ih_8), \\ &\bar{P} = \lim_{f\to\infty} S/L_1 = \frac{(h_3 - h_6) + i(h_4 + h_5)}{h_7 + ih_8} \\ &\bar{\omega}_1 = [Im(\bar{P}\bar{V}^*) - (h_5 - h_4)]/2, \\ &\bar{W} = i\bar{V}. \end{aligned}$$

Thus the solution (5.26) to the pseudo-orthographic approximation remains finite and approaches one of the two solutions to the recovery equation (5.7). It follows that it must approach the one which does not change with f because the other solution does not remain finite. □

Remark. The limiting solution (5.27) is exactly the solution to the recovery equation under orthographic approximation. Such an equation will be given by $h_1 = \omega_2 r + b_1$, $h_2 = \omega_3 r + b_2$, $h_3 = \omega_2 p$, $h_4 = \omega_1 + \omega_2 q$, $h_5 = -\omega_1 + \omega_3 p$, $h_6 = \omega_3 q$, $h_7 = \omega_2$, and $h_8 = \omega_3$. Verification of this fact is straightforward.

Remark. The advantage of using pseudo-orthographic approximation as opposed to solving the recovery equation (5.7) is that one needs to solve only linear equations in the former whereas one needs to solve a cubic equation in the latter.

6. Identifiability condition of a planar surface undergoing affine motion. We consider a planar surface undergoing an affine motion and note that the motion of the shape parameters is given by (2.10). In this section we shall consider identifying parameters of (2.10) by considering an output equation given by (3.8). However, since (3.8) is nonlinear in the parameters, we would like to homogenize the vector $(d_1, \ldots, d_8)^T$ as follows. Let us define

$$\tag{6.1} d_j = \frac{y_j}{y_9}, \quad j = 1, \ldots, 8$$

so that the vector

$$\tag{6.2} (y_1, \ldots, y_9)$$

is a homogenization of the essential parameters. Equation (3.8) can be written as

$$\tag{6.3} \begin{pmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \\ y_5 \\ y_6 \\ y_7 \\ y_8 \\ y_9 \end{pmatrix} = \begin{pmatrix} 0 & 0 & -fb_1 & fa_{13} \\ 0 & 0 & -fb_2 & fa_{23} \\ -b_1' & 0 & b_3 - \delta a_{11} & a_{11} - a_{33} \\ 0 & -b_1' & -\delta a_{12} & a_{12} \\ -b_2' & 0 & -\delta a_{21} & a_{21} \\ 0 & -b_2' & b_3 - \delta a_{22} & a_{22} - a_{33} \\ -b_3' & 0 & -\delta a_{31} & a_{31} \\ 0 & -b_3' & -\delta a_{32} & a_{32} \\ 0 & 0 & -\delta & 1 \end{pmatrix} \begin{pmatrix} \bar{p} \\ \bar{q} \\ -\bar{s} \\ \bar{r} \end{pmatrix},$$

where $\bar{p}$, $\bar{q}$, $\bar{s}$, and $\bar{r}$ have been defined as given by (2.8) and

$$b' = (b_1 - a_{13}\delta \; b_2 - a_{23}\delta \; b_3 - a_{33}\delta) \overset{\Delta}{=} (b'_1 \; b'_2 \; b'_3). \tag{6.4}$$

We now consider the shape dynamic module problem described as follows.

Shape dynamic module problem. Consider a dynamical system (2.10) together with the output function (6.3). The problem is to identify the parameters A, b given by (2.7) and the initial conditions $\bar{p}(0), \bar{q}(0), \bar{s}(0), \bar{r}(0)$ to the extent possible.

The main result of this section is to derive a complete answer to the shape dynamic module problem. Note in particular that the perspective system (2.10), (6.3) is parameterized by a set of 12 motion parameters A, b and a set of 3 shape parameters p, q, r. We shall show that not all 15 parameters are identifiable, i.e., there is a nonunique choice of parameters for which the observation described by (6.3) is the same. The main result of this section is described as follows.

THEOREM 6.1. *Under a suitable generic condition on the set of* 15 *parameters of the perspective system* (2.10), (6.3), *the following parameters or functions of parameters are identifiable. They are*

$$(A, p, q, c_1, c_2, c_3)\,, \tag{6.5}$$

where c_1, c_2, c_3 *is defined in* (3.9).

Thus 14 parameters or functions of the parameters out of a total 15 free parameters are identifiable. The method of solving a set of recovery equations presented in §5 cannot be used to identify these 14 parameters. This is because the output equation (6.3) describes only 8 equations in 15 unknowns. In order to identify 14 parameters, one needs to use the dynamical system (2.10) together with the output equation (6.3). The parameter identification has been carried out via a new "realization theory for perspective systems" described in this section (see also [37]). An important corollary of Theorem 6.1 is now described.

COROLLARY 6.2. *Consider the perspective system* (2.13) (6.3) *parameterized by a set of nine parameters.* (*Here we assume that in* (6.3) *the parameters* a_{ij} *have been replaced by* ω_{ij} *as given by* (2.12)). *Under a suitable generic condition on the set of nine parameters, the following parameters or functions of parameters are identifiable. They are*

$$(\omega_1, \omega_2, \omega_3, p, q, c_1, c_2, c_3)\,, \tag{6.6}$$

where c_1, c_2, c_3 *is defined in* (5.2).

Thus for the perspective system (2.13), (6.3), eight functions of the nine parameters are identifiable. In §5, we have shown that the eight functions (6.6) can be identified, up to a choice of two alternative solutions, by solving the output equation (6.3) alone. Thus use of the dynamical system (2.13) results only in recovering the correct alternative.

In order to prove Theorem 6.1 we need the following notation. Define

$$\mathcal{P} = (\bar{p}\;\bar{q}\; -\bar{s}\;\bar{r})^T\,, \tag{6.7}$$

$$\mathcal{Y} = (y_1\; y_2 \cdots y_9)^T\,, \tag{6.8}$$

$$\mathcal{A} = \begin{pmatrix} -A^T & 0 \\ -b^T & 0 \end{pmatrix}, \tag{6.9}$$

$$\Delta = (\text{the } 9 \times 4 \text{ matrix in } (4.12))\,. \tag{6.10}$$

From (2.10), (6.3) it follows that

$$\mathcal{Y} = \Delta e^{\mathcal{A}t}\mathcal{P}(0), \tag{6.11}$$

where the vector $\mathcal{Y}$ is observed up to a homogeneous line. We shall denote this line by $[\mathcal{Y}]$. As has been described in Ghosh, Jankovic, and Wu [36], the nonuniqueness in $\Delta, \mathcal{A}, \mathcal{P}(0)$, which produces the same $[\mathcal{Y}]$, is given by the orbits of the following group action. They are described as follows:

1. $P \in GL(4)$ acting on $(\Delta, \mathcal{A}, \mathcal{P}(0))$ as follows:

$$(\Delta, \mathcal{A}, \mathcal{P}(0)) \mapsto \left(\Delta P, P^{-1}\mathcal{A}P, P^{-1}\mathcal{P}(0)\right). \tag{6.12}$$

2. $\mu \in \mathbb{R}$ acting on $(\Delta, \mathcal{A}, \mathcal{P}(0))$ as follows:

$$(\Delta, \mathcal{A}, \mathcal{P}(0)) \mapsto (\Delta, \mu I + \mathcal{A}, \mathcal{P}(0)). \tag{6.13}$$

3. $\lambda_1, \lambda_2 \in \mathbb{R} - \{0\}$ acting on $(\Delta, \mathcal{A}, \mathcal{P}(0))$ as follows:

$$(\Delta, \mathcal{A}, \mathcal{P}(0)) \mapsto (\lambda_1\Delta, \mathcal{A}, \lambda_2\mathcal{P}(0)). \tag{6.14}$$

The collective actions (6.12), (6.13), (6.14) will be referred to as the action due to the perspective group $\mathcal{G}$. It is easy to see that the parameters in the orbit of the group $\mathcal{G}$ produce the same output $[\mathcal{Y}]$ and hence cannot be identified. The following proposition shows that under an appropriate generic condition on the parameters of the perspective system (2.13), (6.3), two orbits of the group $\mathcal{G}$ indeed produce a different output $[\mathcal{Y}]$. Hence the orbits of the group $\mathcal{G}$ can indeed be identified.

PROPOSITION 6.3. *Consider a perspective system in continuous time given by*

$$\begin{aligned} \dot{x} &= Ax, \\ z &= [Cx], \end{aligned} \tag{6.15}$$

where we assume that the triplet (C, A, x_0) is minimal. The set of all minimal triplets which produce the same output z is given precisely by the orbits of the $\mathcal{G}$ action.

Proof of Proposition 6.3. Note that the vector function $y(t) = Ce^{At}x_0$ is observed for each t up to a homogeneous line. Assume that there is a scaling function $r(t)$ such that $r(t)y(t)$ is the output of a linear system of degree n, where we assume that $r(0) = 1$. Discretizing the system (6.15) at discrete interval $T, 2T, \ldots$, where T has been chosen to be sufficiently small, it follows from [37] that $r(jT) = r(T)^j$. Since T is arbitrary, it follows that the function $r(t)$ is such that $r(jTt) = r(Tt)^j$, for all $t \in \mathbb{R}, \mathrm{j} = 0, 1, \ldots$. If $r(t)$ is a differentiable function at $t = 0$, it follows that

$$\frac{r(t+\Delta t)}{r(t)} = r(\Delta t) = r(0) + r'(0)\Delta t.$$

One therefore concludes that

$$r'(t) = r'(0)r(t).$$

Thus the scaling function $r(t)$ is an exponential function given by

$$r(t) = e^{r'(0)t}.$$

Thus the scaling of C, A, x_0 is such that C, x_0 is scaled by a scalar multiple. The matrix A is scaled as

$$A \mapsto r'(0)I + A. \qquad \square$$

In general the $GL(4)$ action on the triplet $(\Delta, \mathcal{A}, \mathcal{P}(0))$ changes the structure of the matrix Δ and $\mathcal{A}$. The subgroup of $GL(4)$ which preserves the structure is now described.

THEOREM 6.4. *Define*

$$b_1' = b_1 - a_{13}\delta, \quad b_2' = b_2 - a_{23}\delta, \quad b_3' = b_3 - a_{33}\delta. \tag{6.16}$$

Assume that

$$b_1' a_{23} - b_2' a_{13} \neq 0, \quad b_2' a_{31} - b_3' a_{21} \neq 0, \quad b_1' a_{32} - b_3' a_{12} \neq 0, \tag{6.17}$$

$$(b_1' \; b_2' \; b_3') \neq 0, \tag{6.18}$$

$$\begin{pmatrix} A \\ b'^T \end{pmatrix} \text{ has rank } 3, \tag{6.19}$$

where $b'^T = (b_1' \; b_2' \; b_3')$. *Under the generic assumption* (6.17), (6.18), (6.19), *the only subgroup of* $GL(4)$ *which preserves the structure of* $(\Delta, \mathcal{A})$ *under the action* (6.12) *is given by*

$$\bar{P} = \begin{pmatrix} \alpha_{11} & 0 & 0 & 0 \\ 0 & \alpha_{11} & 0 & 0 \\ 0 & 0 & \alpha_{11} & 0 \\ 0 & 0 & \delta\alpha_{11} & \alpha_{44} \end{pmatrix}, \tag{6.20}$$

where $\alpha_{11} \neq 0$, $\alpha_{44} \neq 0$.

Proof of Theorem 6.4. Let

$$Q = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & \delta & 1 \end{pmatrix}. \tag{6.21}$$

It is easy to see that

$$\mathcal{A}_1 \triangleq Q^{-1}\mathcal{A}Q = \begin{pmatrix} -A^T & 0 \\ -b'^T & 0 \end{pmatrix}, \tag{6.22}$$

$$\Delta_1 \triangleq \Delta Q = \begin{pmatrix} 0 & 0 & -fb_1' & fa_{13} \\ 0 & 0 & -fb_2' & fa_{23} \\ -b_1' & 0 & b_3' & a_{11} - a_{33} \\ 0 & -b_1' & 0 & a_{12} \\ -b_2' & 0 & 0 & a_{21} \\ 0 & -b_2' & b_3' & a_{22} - a_{33} \\ -b_3' & 0 & 0 & a_{31} \\ 0 & -b_3' & 0 & a_{32} \\ 0 & 0 & 0 & 1 \end{pmatrix}. \tag{6.23}$$

Let $Q_1 = (\alpha_{ij})$ be a nonsingular 4×4 matrix. Under the generic condition (6.17), (6.18) it may be concluded that $\Delta_1 Q_1$ has the same structure as Δ_1 if Q_1 has the form

$$Q_1 = \begin{pmatrix} \alpha_{11} & 0 & 0 & \alpha_{14} \\ 0 & \alpha_{11} & 0 & \alpha_{24} \\ 0 & 0 & \alpha_{11} & \alpha_{34} \\ 0 & 0 & 0 & \alpha_{44} \end{pmatrix} \triangleq \begin{pmatrix} \alpha_{11} I & \Theta \\ 0 & \alpha_{44} \end{pmatrix}. \tag{6.24}$$

Computing $Q_1^{-1}\mathcal{A}_1 Q_1$ we have

$$Q_1^{-1}\mathcal{A}_1 Q_1 = \begin{pmatrix} -A^T + \frac{1}{\alpha_{44}}\Theta b'^T & -\frac{1}{\alpha_{11}}A^T\Theta + \frac{1}{\alpha_{11}\alpha_{44}}\Theta b'^T\Theta \\ -\frac{\alpha_{11}}{\alpha_{44}}b'^T & -\frac{1}{\alpha_{44}}b'^T\Theta \end{pmatrix}. \tag{6.25}$$

In order for $Q_1^{-1}\mathcal{A}_1 Q_1$ to have the same structure as $\mathcal{A}_1$ we must have

$$b'^T\Theta = 0 \quad \text{and} \quad A^T\Theta = 0.$$

Under the generic condition (6.19), it follows that $\Theta = 0$. Thus Q_1 is of the form

$$Q_1 = \begin{pmatrix} \alpha_{11} & 0 & 0 & 0 \\ 0 & \alpha_{11} & 0 & 0 \\ 0 & 0 & \alpha_{11} & 0 \\ 0 & 0 & 0 & \alpha_{44} \end{pmatrix}. \tag{6.26}$$

The structure (6.20) of the $\bar{P}$ matrix is obtained by defining $\bar{P} = QQ_1$, where Q, Q_1 are given by (6.21), (6.26), respectively. □

Note that $\bar{P}^{-1}\mathcal{A}\bar{P}$ takes up the form

$$\begin{pmatrix} -A^T & 0 \\ -\frac{\alpha_{11}}{\alpha_{44}}b'^T & 0 \end{pmatrix}. \tag{6.27}$$

On the other hand, $\bar{P}^{-1}\mathcal{P}$ is given by

$$\begin{pmatrix} \frac{\bar{p}}{\alpha_{11}} & \frac{\bar{q}}{\alpha_{11}} & -\frac{\bar{s}}{\alpha_{11}} & \frac{\bar{r}+\delta\bar{s}}{\alpha_{44}} \end{pmatrix}^T \tag{6.28}$$

and $\Delta\bar{P}$ is given by

$$\begin{pmatrix} 0 & 0 & -\alpha_{11}fb'_1 & \alpha_{44}fa_{13} \\ 0 & 0 & -\alpha_{11}fb'_2 & \alpha_{44}fa_{23} \\ -\alpha_{11}b'_1 & 0 & \alpha_{11}b'_3 & \alpha_{44}(a_{11}-a_{33}) \\ 0 & -\alpha_{11}b'_1 & 0 & \alpha_{44}a_{12} \\ -\alpha_{11}b'_2 & 0 & 0 & \alpha_{44}a_{21} \\ 0 & \alpha_{11}b'_2 & \alpha_{11}b'_3 & \alpha_{44}(a_{22}-a_{33}) \\ -\alpha_{11}b'_3 & 0 & 0 & \alpha_{44}a_{31} \\ 0 & -\alpha_{11}b'_3 & 0 & \alpha_{44}a_{32} \\ 0 & 0 & 0 & \alpha_{44} \end{pmatrix}. \tag{6.29}$$

In order to get the last row of $\Delta\bar{P}$ to be a unit vector we apply the group in (6.14) to (6.29) with $\lambda_1 = \frac{1}{\alpha_{44}}$. From (6.27), (6.28), (6.29) it can be concluded that the subgroup (6.20) essentially scales the vector b' by the scalar $\frac{\alpha_{11}}{\alpha_{44}}$. Likewise it scales

$r+\delta$ by $\frac{\alpha_{11}}{\alpha_{44}}$. Hence the function $\frac{b'}{(r+\delta)}$ remains invariant in the orbit of the subgroup (6.20) action.

The subgroup (6.13) essentially changes the diagonal of the matrix $\mathcal{A}$. Since the diagonal of the matrix $\mathcal{A}$ is given by $(-a_{11} \;\; -a_{22} \;\; -a_{33} \;\; 0)$ it follows that the subgroup which preserves the structure is given by $\mu = 0$ and the parameters a_{11}, a_{22}, a_{33} remain invariant in the orbit of this subgroup action.

Proof of Theorem 6.1. Note that under the generic conditions (6.17), (6.18), (6.19) the functions (6.5) remain invariant under the action of the perspective group $\mathcal{G}$, i.e., they remain constant in the orbits of the $\mathcal{G}$ action. In Proposition 6.3 we show that additionally if $(\Delta, \mathcal{A}, \mathcal{P}(0))$ is a minimal triplet then no two orbits of the $\mathcal{G}$ action produce the same output $[\mathcal{Y}]$. □

7. Identification of parameters based on the orthographic projection.

The orthographic projection occurs as a special case of the generalized projection (3.1) when we assume $\delta = f$ and let $f \to \infty$. In this case, the parameters d_7, d_8 of the output equation (6.3) or (3.9) are forced to zero or equivalently, the quadratic term in (3.7) or (3.12) drops out. Thus, the optical flow equation is given by (3.13) and the recovery equation is given by the first six components $h_1, \ldots, h_6$ of (3.11).

7.1. Solution to the recovery equation for the rigid motion.

We begin this section by considering a plane undergoing a rigid motion given by (2.13). The corresponding recovery equation is given by

$$\begin{aligned} d_1 &= \omega_2 r + b_1, & d_2 &= \omega_3 r + b_2, & d_3 &= \omega_2 p, \\ d_4 &= \omega_1 + \omega_2 q, & d_5 &= -\omega_1 + \omega_3 p, & d_6 &= \omega_3 q, \end{aligned} \tag{7.1}$$

where we shall assume that the vector $(d_1, \ldots, d_6)$ is estimated by the IDM. Kanatani [11] has considered the problem of solving (7.1) for the parameters $\omega_1, \omega_2, \omega_3, p, q, r, b_1, b_2$. The parameter b_3 does not enter (7.1) and is therefore not recoverable from the equation (7.1). Moreover since we have six equations in eight unknowns we do not expect to recover the parameters even up to finitely many alternatives. In fact, it is already known (see Kanatani [11]) that the recovery equation (7.1) can be solved in the following way.

Let us define

$$\begin{aligned} V &= d_1 + i d_2, & T &= d_3 + d_6, & R &= d_5 - d_4, \\ S &= d_3 - d_6 + i(d_4 + d_5), & P &= p + iq, & W &= -\omega_3 + i\omega_2. \end{aligned} \tag{7.2}$$

From (7.1) we obtain the following:

$$\begin{aligned} PW &= iS, \\ PW^* &= -(2\omega_1 + R) - iT. \end{aligned} \tag{7.3}$$

The equation (7.3) can be solved as follows.

$$\begin{aligned} &\omega_1 = -\tfrac{1}{2}\left(R \pm \sqrt{SS^* - T^2}\right), \\ &W = k \exp\left[i\left\{\tfrac{\pi}{4} + \tfrac{1}{2}\arg S - \tfrac{1}{2}\arg(-2\omega_1 - R - iT)\right\}\right], \\ &P = i\frac{S}{W}, \end{aligned} \tag{7.4}$$

where k is an arbitrary constant. The parameters b_1, b_2 and r are given as $B - iWr = V$ where $B = b_1 + ib_2$. Thus we have the following theorem essentially described by Kanatani [11].

THEOREM 7.1. *The recovery equation* (7.1) *can be solved up to two parameters* k_1, k_2 *and up to a choice of sign as follows:*

$$\begin{aligned}
&\omega_1 = -\tfrac{1}{2}\left(R \pm \sqrt{SS^* - T^2}\right),\\
&W = k_1 \exp\left[i\left\{\tfrac{\pi}{4} + \tfrac{1}{2}\arg S - \tfrac{1}{2}\arg\left(-2\omega_1 - R - iT\right)\right\}\right],\\
&P = i\tfrac{S}{W},\\
&B = V + ik_2W,\\
&r = k_2.
\end{aligned} \tag{7.5}$$

The proof of Theorem 7.1 is clear from the above discussion. Note that the solution to the recovery equation (7.2) is ambiguous up to a sign and is obtained up to a pair of parameters k_1 and k_2, out of a total of eight parameters, which excludes the parameter b_3.

7.2. Identification of a planar surface undergoing affine motion. Let us now homogenize the output equation (7.1) described as follows:

$$\begin{pmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \\ y_5 \\ y_6 \\ y_9 \end{pmatrix} = \begin{pmatrix} 0 & 0 & -b_1 & a_{13} \\ 0 & 0 & -b_2 & a_{23} \\ a_{13} & 0 & -a_{11} & 0 \\ 0 & a_{13} & -a_{12} & 0 \\ a_{23} & 0 & -a_{21} & 0 \\ 0 & a_{23} & -a_{22} & 0 \\ 0 & 0 & -1 & 0 \end{pmatrix} \begin{pmatrix} \bar{p} \\ \bar{q} \\ -\bar{s} \\ \bar{r} \end{pmatrix}. \tag{7.6}$$

We now proceed to consider the technique described in §7 for the perspective system (2.10), (7.6). Note that (7.6) is the homogeneous version of the recovery equation (3.8). Denote the 7 × 4 matrix in (7.6) by π. The main result of this section is described as follows.

THEOREM 7.2. *Consider the perspective system* (2.10), (7.6) *parameterized by a set of* 15 *parameters. Assume furthermore that the parameters satisfy the generic condition*

$$a_{13} \neq 0, \quad b_1a_{23} - b_2a_{13} \neq 0, \quad a_{12}a_{23} - a_{13}a_{22} \neq 0, \quad a_{11}a_{23} - a_{13}a_{21} \neq 0 \tag{7.7}$$

and the triplet $(\pi, \mathcal{A}, \mathcal{P}(0))$ *is a minimal triplet. The functions of the parameters that can be identified are given by*

$$\begin{gathered}
\tfrac{a_{23}}{a_{13}}, \quad b_2 - b_1\tfrac{a_{23}}{a_{13}}, \quad a_{21} - a_{11}\tfrac{a_{23}}{a_{13}}, \quad a_{22} - a_{12}\tfrac{a_{23}}{a_{13}}, \quad -a_{13}p - a_{11}, \quad -a_{13}q - a_{12},\\
-a_{13}r - b_1, \xi, \quad \left(a_{11}^2 + a_{12}a_{21} + a_{13}a_{31}\right) - a_{11}\xi,\\
\left(a_{11}a_{12} + a_{12}a_{22} + a_{13}a_{32}\right) - a_{12}\xi, \left(a_{11}b_1 + a_{12}b_2 + a_{13}b_3\right) - b_1\xi,
\end{gathered} \tag{7.8}$$

where ξ *is defined to be*

$$\xi = \tfrac{1}{a_{13}}\left(a_{11}a_{13} + a_{12}a_{23} + a_{13}a_{33}\right).$$

Remark. Thus there is a total of 11 functions of motion and shape parameters that can be identified.

Proof of Theorem 7.2. Let P be a nonsingular 4 × 4 matrix. Under the generic condition (7.7) it can be shown that πP has the same structure as that of π provided P is of the form

$$P = \begin{pmatrix} \alpha_{11} & 0 & \alpha_{13} & 0 \\ 0 & \alpha_{11} & \alpha_{23} & 0 \\ 0 & 0 & \alpha_{33} & 0 \\ 0 & 0 & \alpha_{43} & \alpha_{11} \end{pmatrix} \tag{7.9}$$

and $\frac{1}{\alpha_{33}}\pi P$ is of the form

$$\begin{pmatrix} 0 & 0 & -b_1 + \frac{\alpha_{43}}{\alpha_{33}}a_{13} & \frac{\alpha_{11}}{\alpha_{33}}a_{13} \\ 0 & 0 & -b_2 + \frac{\alpha_{43}}{\alpha_{23}}a_{23} & \frac{\alpha_{11}}{\alpha_{33}}a_{23} \\ \frac{\alpha_{11}}{\alpha_{33}}a_{13} & 0 & -a_{11} + \frac{\alpha_{13}}{\alpha_{33}}a_{13} & 0 \\ 0 & \frac{\alpha_{11}}{\alpha_{33}}a_{13} & -a_{12} + \frac{\alpha_{23}}{\alpha_{33}}a_{13} & 0 \\ \frac{\alpha_{11}}{\alpha_{33}}a_{23} & 0 & -a_{21} + \frac{\alpha_{13}}{\alpha_{33}}a_{23} & 0 \\ 0 & \frac{\alpha_{11}}{\alpha_{33}}a_{23} & -a_{22} + \frac{\alpha_{23}}{\alpha_{33}}a_{23} & 0 \\ 0 & 0 & -1 & 0 \end{pmatrix}. \tag{7.10}$$

Likewise $P^{-1}\mathcal{A}P$ is of the form.

$$\begin{pmatrix} -a_{11} + \frac{\alpha_{13}}{\alpha_{33}}a_{13} & -a_{21} + \frac{\alpha_{13}}{\alpha_{33}}a_{23} & \Theta_1 & 0 \\ -a_{12} + \frac{\alpha_{23}}{\alpha_{33}}a_{13} & -a_{22} + \frac{\alpha_{23}}{\alpha_{33}}a_{23} & \Theta_2 & 0 \\ -\frac{\alpha_{11}}{\alpha_{33}}a_{13} & -\frac{\alpha_{11}}{\alpha_{33}}a_{23} & \Theta_3 & 0 \\ -b_1 + \frac{\alpha_{43}}{\alpha_{33}}a_{13} & -b_2 + \frac{\alpha_{43}}{\alpha_{33}}a_{23} & \Theta_4 & 0 \end{pmatrix}, \tag{7.11}$$

where

$$\Theta_1 = -\frac{1}{\alpha_{11}}\left(\alpha_{13}a_{11} + \alpha_{23}a_{21} + \alpha_{33}a_{31}\right) + \frac{\alpha_{13}}{\alpha_{11}\alpha_{33}}\left(\alpha_{13}a_{13} + \alpha_{23}a_{23} + \alpha_{33}a_{33}\right),$$

$$\Theta_2 = -\frac{1}{\alpha_{11}}\left(\alpha_{13}a_{12} + \alpha_{23}a_{22} + \alpha_{33}a_{32}\right) + \frac{\alpha_{23}}{\alpha_{11}\alpha_{33}}\left(\alpha_{13}a_{13} + \alpha_{23}a_{23} + \alpha_{33}a_{33}\right),$$

$$\Theta_3 = -\frac{1}{\alpha_{33}}\left(\alpha_{13}a_{13} + \alpha_{23}a_{23} + \alpha_{33}a_{33}\right),$$

$$\Theta_4 = -\frac{1}{\alpha_{11}}\left(\alpha_{13}b_1 + \alpha_{23}b_2 + \alpha_{33}b_3\right) + \frac{\alpha_{43}}{\alpha_{11}\alpha_{33}}\left(\alpha_{13}a_{13} + \alpha_{23}a_{23} + \alpha_{33}a_{33}\right).$$

Of course the matrices (7.10), (7.11) are the new structures of the matrix π and $\mathcal{A}$ respectively after transformation.

It follows that the set of parameters that would produce the same output (7.6) is given by

$$\begin{aligned}
&a_{11} \mapsto a_{11} - \pi_2 a_{13}, \\
&a_{21} \mapsto a_{21} - \pi_2 a_{23}, \\
&a_{12} \mapsto a_{12} - \pi_3 a_{13}, \\
&a_{22} \mapsto a_{22} - \pi_3 a_{23}, \\
&a_{13} \mapsto \pi_1 a_{13}, \\
&a_{23} \mapsto \pi_1 a_{23}, \\
&b_1 \mapsto b_1 - \pi_4 a_{13}, \\
&b_2 \mapsto b_2 - \pi_4 a_{23}, \\
&a_{31} \mapsto \frac{1}{\pi_1}\left(\pi_2 a_{11} + \pi_3 a_{21} + a_{31}\right) - \frac{\pi_2}{\pi_1}\left(\pi_2 a_{13} + \pi_3 a_{23} + a_{33}\right), \\
&a_{32} \mapsto \frac{1}{\pi_1}\left(\pi_2 a_{12} + \pi_3 a_{22} + a_{32}\right) - \frac{\pi_3}{\pi_1}\left(\pi_2 a_{13} + \pi_3 a_{23} + a_{33}\right), \\
&b_3 \mapsto \frac{1}{\pi_1}\left(\pi_2 b_1 + \pi_3 b_2 + b_3\right) - \frac{\pi_4}{\pi_1}\left(\pi_2 a_{13} + \pi_3 a_{23} + a_{33}\right), \\
&a_{33} \mapsto \pi_2 a_{13} + \pi_3 a_{23} + a_{33}, \\
&p \mapsto \frac{1}{\pi_1}(p + \pi_2), \\
&q \mapsto \frac{1}{\pi_1}(q + \pi_3), \\
&r \mapsto \frac{1}{\pi_1}(r + \pi_4),
\end{aligned} \tag{7.12}$$

where

$$\pi_1 = \frac{\alpha_{11}}{\alpha_{33}}, \quad \pi_2 = \frac{\alpha_{13}}{\alpha_{33}}, \quad \pi_3 = \frac{\alpha_{23}}{\alpha_{33}}, \quad \pi_4 = \frac{\alpha_{43}}{\alpha_{33}}.$$

In fact (7.12) describes the orbit in the parameter space corresponding to the subgroup (7.9). The orbit is parameterized by 4 parameters $\pi_1, \pi_2, \pi_3, \pi_4$.

From the results in Proposition 6.3 it can be inferred that parameters can be identified up to the orbit described in (7.12). Finally the functions (7.8) are derived by choosing $\pi_1, \pi_2, \pi_3, \pi_4$ by restricting $a_{11} - \pi_2 a_{13} = 0$, $b_1 - \pi_4 a_{13} = 0$, $\pi_1 a_{13} = 1$, and $a_{12} - \pi_3 a_{13} = 0$. □

7.3. Identification of a planar surface undergoing rigid motion. If we assume that the matrix A is skew symmetric, one needs to restrict the following in (7.12):

$$a_{11} = a_{22} = a_{33} = 0, \quad a_{12} = -a_{21},\ a_{13} = -a_{31},\ a_{23} = -a_{32}. \tag{7.13}$$

It follows that $\pi_2 = 0$, $\pi_3 = 0$, implying that $\alpha_{13} = 0$, $\alpha_{23} = 0$. Furthermore $\pi_1 a_{13} = -\frac{1}{\pi_1} a_{31}$, i.e., $\pi_1 = \pm 1$ or $\alpha_{11} = \pm\alpha_{33}$. Thus the subgroup P described by (7.9) is further restricted to

$$P_1 = \begin{pmatrix} \alpha_{11} & 0 & 0 & 0 \\ 0 & \alpha_{11} & 0 & 0 \\ 0 & 0 & \pm\alpha_{11} & 0 \\ 0 & 0 & \alpha_{43} & \alpha_{11} \end{pmatrix}. \tag{7.14}$$

The orbit (7.12) under the new subgroup action (7.14) is given by $a_{12} \mapsto a_{12}$, $a_{13} \mapsto \pm a_{13}$, $a_{23} \mapsto \pm a_{23}$, $b_1 \mapsto b_1 \mp \frac{\alpha_{43}}{\alpha_{11}} a_{13}$, $b_2 \mapsto b_2 \mp \frac{\alpha_{43}}{\alpha_{11}} a_{23}$, $b_3 \mapsto \pm b_3$, $p \mapsto \pm p$, $q \mapsto \pm q$, and $r \mapsto \pm(r + \frac{\alpha_{43}}{\alpha_{11}})$.

Thus we have the following theorem regarding the condition of identifiability for a perspective system (2.13), (7.6), parameterized by a set of nine parameters.

THEOREM 7.3. *Consider the perspective system* (2.13), (7.6), *parameterized by* $\omega_1, \omega_2, \omega_3$, b_1, b_2, b_3, p, q, r. *Assume that the parameters satisfy the generic condition* $\omega_2 \neq 0$, $\omega_3 \neq 0$, $\omega_1 \neq 0$, $b_1\omega_3 - b_2\omega_2 \neq 0$. *Assume furthermore that the triplet*

$$\begin{pmatrix} 0 & 0 & -b_1 & \omega_2 \\ 0 & 0 & -b_2 & \omega_3 \\ \omega_2 & 0 & 0 & 0 \\ 0 & \omega_2 & -\omega_1 & 0 \\ \omega_3 & 0 & \omega_1 & 0 \\ 0 & \omega_3 & 0 & 0 \\ 0 & 0 & -1 & 0 \end{pmatrix}, \begin{pmatrix} 0 & \omega_1 & \omega_2 & 0 \\ -\omega_1 & 0 & \omega_3 & 0 \\ -\omega_2 & -\omega_3 & 0 & 0 \\ -b_1 & -b_2 & -b_3 & 0 \end{pmatrix}, \begin{pmatrix} \bar{p} \\ \bar{q} \\ -\bar{s} \\ \bar{r} \end{pmatrix}$$

is minimal. The functions of the parameters that can be identified are given by ω_1, $\pm\omega_2$, $\pm\omega_3$, $b_1 \mp \frac{\alpha_{43}}{\alpha_{11}}\omega_2$, $b_2 \mp \frac{\alpha_{43}}{\alpha_{11}}\omega_3$, $\pm b_3$, $\pm p$, $\pm q$, $\pm(r + \frac{\alpha_{43}}{\alpha_{11}})$. *The ratio* $\frac{\alpha_{43}}{\alpha_{11}}$ *is to be thought of as a single parameter.*

Remark. In Theorem 7.3 the parameters that can be identified are ambiguous up to a sign and up to one parameter. It may be verified that the functions $b_1 + \omega_2 r$ and $b_2 + \omega_3 r$ remain constant regardless of the choice of the sign. Thus one concludes that the parameters b_1, b_2, r can be recovered up to a line in the space (b_1, b_2, r)

THEOREM 7.4. *Let us consider the perspective system* (2.13), (6.3). *Under generic condition, the set of parameters or function of parameters that can be identified in the set* $\omega_1, \omega_2, \omega_3, b_1, b_2, b_3, p, q, r$ *approaches the set of parameters or function of parameters* (*up to possibly a sign ambiguity*) *that can be identified for the perspective system* (2.13), (7.6) *as* $f \to \infty$. *The parameters that can be identified as* $f \to \infty$ *are given precisely by* $\omega_1, \omega_2, \omega_3, p, q, b_3, b_1 + \omega_2 r, b_2 + \omega_3 r$.

Remark. The result of Theorem 7.4 is actually quite surprising. It says that for the projection (3.1), if $f = \delta$ and $f \to \infty$, as the generalized projection (3.1) approaches orthographic projection, the line in the parameter space (b_1, b_2, b_3, r) that can be identified at any given f indeed approaches (modulo sign) the corresponding line in the parameter space (b_1, b_2, b_3, r) that can be identified under orthographic projection. This continuity continues to hold even though under orthographic projection one measures only the vector $(d_1, \ldots, d_6)$, i.e., the measurements d_7 and d_8 are completely lost.

The following theorem generalizes the result stated in the Theorem 7.4.

THEOREM 7.5. *Let us consider the perspective system* (2.10), (6.3). *Under generic condition, the functions of parameters that can be identified as $f \to \infty$ are given precisely by A, p, q, $b_1 + a_{13}r$, $b_2 + a_{23}r$, and $b_3 + a_{33}r$. Thus parameters are recovered up to a one-parameter ambiguity even when $f \to \infty$. Moreover this one-parameter ambiguity curve is a subset of the four-parameter orbit described by* (7.12).

Proof of Theorems 7.4 *and* 7.5. At a given value of f, the parameters that can be identified have been already described by Theorem 6.1 and Corollary 6.2. As $f \to \infty$, the essential parameter d_1 approaches h_1 and d_2 approaches h_2. Hence in the limit one observes $b_1 + a_{13}r$ and $b_2 + a_{23}r$. At a given value of f, the parameters b_3 and r are known only up to the line given by

$$(d_3 + pc_1)r - (a_{11} - a_{33})r + b_3 = (a_{11} - (d_3 + pc_1))f.$$

As $f \to \infty$ the above line converges to the line $b_3 + a_{33}r =$ constant. Hence in the limit one also observes the function $b_3 + a_{33}r$. Finally note that in the orbit described by (7.12), if we assume that $\pi_2 = 0$, $\pi_3 = 0$ and $\pi_1 = 1$, we obtain a one-parameter orbit in which A, p, q, $b_1 + a_{13}r$, $b_2 + a_{23}r$, and $b_3 + a_{33}r$ are all invariants. This completes the proof. □

Remark. The proof of Theorems 7.4 essentially follows from Theorem 5.3.

8. Simulation results. Extensive simulations have been carried out for the methods outlined in §§4 and 5 of this paper. Simulations were performed only for the case of rigid body motion of a planar surface. First, the "intensity-dynamics" based approach was implemented to estimate the essential parameter vector d following equations (4.3)–(4.8). Simulations were performed for this approach using three different texture functions while the effect of varying the spatial and temporal sampling rates (step size) were examined. Additional algorithms were implemented to estimate the vector d using "feature-dynamics" based approaches for points (4.24)–(4.27), lines (4.17)–(4.23), and curves (4.13)–(4.15). Simulations for each of these approaches were performed to examine the effect of varying the number of points sampled and the step size. Motion parameters were estimated following equations (5.9)–(5.11) and (5.15)–(5.16). We draw the following conclusions from the results of the simulations:

1. Under the assumptions of a textured surface, perfect focus, and no noise, the methods outlined in this paper are effective for the estimation of shape and motion parameters.

2. The choice of the initial intensity function does not significantly affect the accuracy of the "intensity-dynamics" based approach. To illustrate, given the initial intensity function $e(x, y, 0) = \sin^2 x + \cos^2 y$ with a step size of $\Delta x = \Delta y = \Delta t = 10^{-8}$, we were able to estimate with a root mean square (rms) error of approximately

TABLE 8.1
Observation at multiple times removes the ambiguity of dual solutions.

	Motion and shape parameters		
	ω	c	$[p, q]^T$
Actual values (at $t = 0.0, f = 1.0$)	-4.000 5.000 1.000	3.500 1.500 1.500	0.500 -1.500
Solution no. 1 for $t = 0.0$	-4.750 0.750 7.000	-0.750 2.250 1.500	-2.333 -1.000
Solution no. 2 for $t = 0.0$	-4.000 5.000 1.000	3.500 1.500 1.500	0.500 -1.500
Solution no. 1 for $t = 0.1$	-3.932 0.083 5.758	-1.244 1.506 1.574	-2.333 -1.000
Solution no. 2 for $t = 0.1$	-4.000 5.000 1.000	3.673 1.574 1.574	0.790 -0.957

2.9×10^{-5}. The rms error for the initial intensity function $e(x, y, 0) = 1/x + 1/y$ was approximately 1.2×10^{-5} for the same step size.

3. Increasing the number of points sampled does not, in general, significantly increase accuracy. For example, in the case of the "feature-point" based approach, observation of the minimum four points yielded an rms error of 1.8×10^{-6} whereas observation of 32 points yielded an rms error of 8.6×10^{-7}. In both cases the step size was as noted before.

4. Decreasing the spatial or temporal sampling rates has a significant adverse affect on accuracy. For example, if the step size is increased to $\Delta x = \Delta y = \Delta t = 10^{-5}$, the rms increases by roughly the same factor, 10^3. This effect cannot be compensated for by increasing the number of points sampled.

Further simulations were performed to demonstrate how the ambiguity of the two solutions described in Theorem 5.1 can be resolved by sampling at either multiple times or multiple focal lengths. The use of multiple times to resolve this ambiguity has previously been suggested by Waxman and Ullman [7] and Tsai and Huang [21].

In Table 8.1 we note that the estimated values for ω do not change with time in solution no. 2 but do change with time in solution no. 1. Thus, since the ω values are constant, solution no. 2 is chosen as the correct solution. The values of c, p, and q change with time in both solutions. This is to be expected since c depends on r and p, q, and r all vary with time. Table 8.2 illustrates corresponding results for multiple focal lengths. For the correct solution in this case, the values of ω, p, and q remain constant while the values of c vary with focal length.

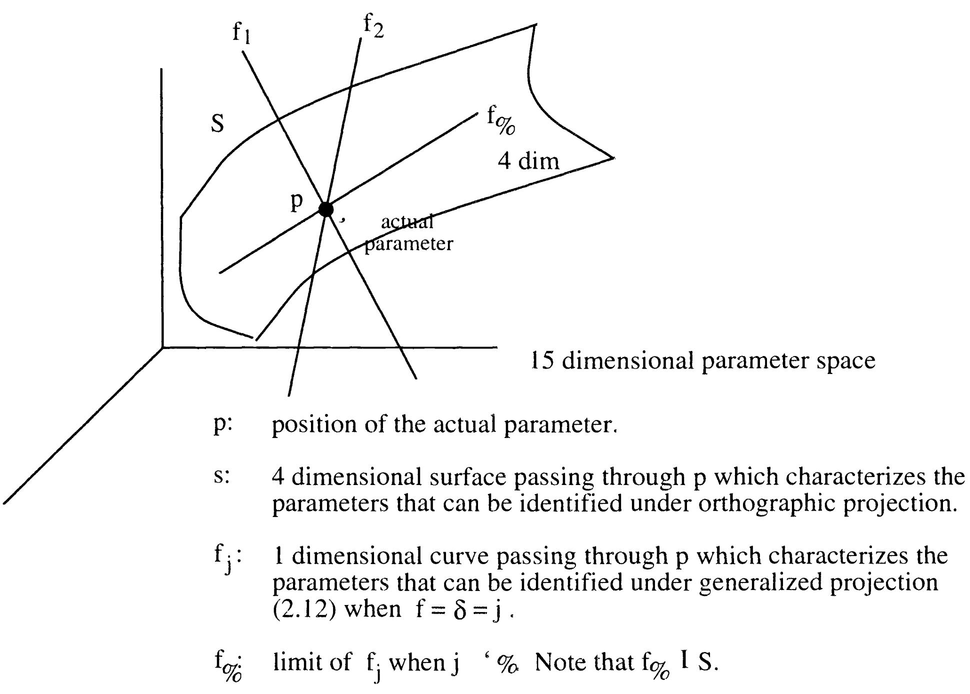
9. Summary and conclusions.

This paper introduces a two-module approach to motion and shape estimation either by observing dynamically moving intensity or shading or by observing dynamically moving feature points, lines, or curves. When restricted to a planar surface undergoing affine motion, the problem can be tackled by estimating an intermediate set of parameters known as essential parameters. We show that the essential parameter vector can be estimated, under a suitable generic condition, independent of whether the observation is the moving intensity function or the moving features on the image plane.

We introduce a new "dynamical systems" viewpoint on the motion and shape

TABLE 8.2
Observation at different focal lengths also removes the ambiguity of dual solutions.

	Motion and shape parameters		
	ω	c	$[p, q]^T$
Actual values (at $t = 0.0, f = 1.0$)	-4.000 5.000 1.000	3.500 1.500 1.500	0.500 -1.500
Solution no. 1 for $f = 1.0$	-4.750 0.750 7.000	-0.750 2.250 1.500	-2.333 -1.000
Solution no. 2 for $f = 1.0$	-4.000 5.000 1.000	3.500 1.500 1.500	0.500 -1.500
Solution no. 1 for $f = 2.0$	-3.167 0.500 8.167	-0.500 1.500 1.000	-4.000 -2.333
Solution no. 2 for $f = 2.0$	-4.000 5.000 1.000	4.000 2.333 1.000	0.500 -1.500



p: position of the actual parameter.

s: 4 dimensional surface passing through p which characterizes the parameters that can be identified under orthographic projection.

f_j: 1 dimensional curve passing through p which characterizes the parameters that can be identified under generalized projection (2.12) when f = δ = j .

$f_{\%}$: limit of f_j when j ‘ %. Note that $f_{\%}$ I S.

FIG. 9.1. *Identifiable parameters for a planar surface undergoing an affine flow.*

estimation problem and show that the dynamics of the plane, known as the shape dynamics, together with the essential parameters viewed as an output equation are an example of a perspective system. Introducing a new realization theory for perspective systems, we show that the parameters of the system can be identified up to orbits of a suitable "perspective group" action, provided of course the parameters satisfy a suitable generic condition.

Using this approach, we analyze a planar surface undergoing a rigid motion and show that the solution to the parameter estimation problem under a general projection and a pseudo-orthographic projection indeed converges to that obtained (up to choice of a sign) under orthographic projection as the general projection model converges in the limit to the orthographic projection. This conclusion is in sharp contrast to that reported by Kanatani [11], wherein only the recovery equation has been used. We also analyze a planar surface undergoing an affine motion and show that under general projection, parameters are recovered up to a one-parameter ambiguity whereas under orthographic projection parameters are recovered up to four-parameter ambiguity. In the limit when the general projection model converges to the orthographic projection, the above family of one-parameter orbits converge to a one-parameter subset of the four-parameter class.

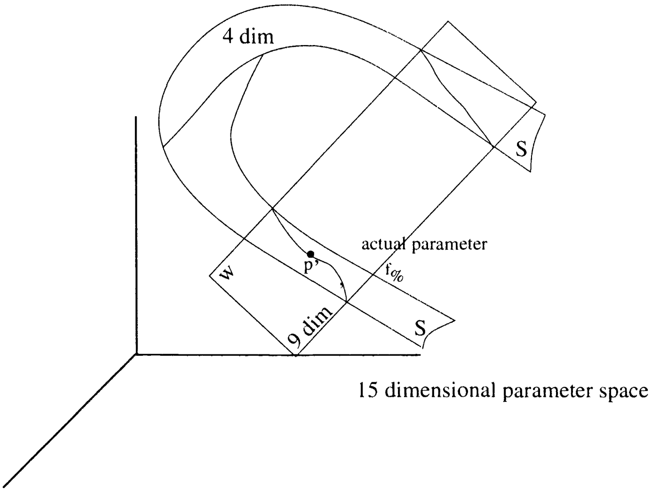


FIG. 9.2. *Identifiable parameters for a planar surface undergoing a rigid motion.*

This indicates that "one can see a nonrigid affine flow better" using a visual system with the capability of varying the focal length f all the way to infinity, as compared to a visual system with focal length f fixed at infinity. However, for a rigid flow, there is no distinction.

The above conclusion has been summarized in Figs. 9.1 and 9.2. In Fig. 9.1 we show that if p is the position of the actual parameters in $\mathbb{R}^{15}$, where $\mathbb{R}^{15}$ is the parameter space for A, b, p, q, r, under projection (3.1), if $f = \delta = j$, the curve f_j, $j = 1, 2, 3, \ldots$ indicates the set of parameters that can be identified for various values of f. In fact when $f \longrightarrow \infty$, , f_∞ denotes the limiting curve that describes the set of parameters that can be asymptotically identified. The four-dimensional surface S characterizes the parameters that can be identified under orthographic projection. In this paper we show that $f_\infty \subset S$. Thus we conclude that for an affine flow it helps to consider a visual system with a capability to vary f. For f permanently with focus at ∞, parameters are recovered up to a four-parameter ambiguity as opposed to one-parameter ambiguity in all the other cases.

In Fig. 9.2 we show a nine-dimensional subspace W of parameters describing the parameters of a planar surface undergoing a rigid flow. The subspace W intersects S in exactly two one-dimensional curves. In this paper we show that one of the two one-dimensional curves is f_∞. Thus for a planar surface undergoing a rigid flow, orthographic projection identifies parameters up to a one-parameter curve and up to a sign ambiguity. Via the process of choosing a projection (3.1) and letting $f \to \infty$, one can determine the sign. The one-parameter ambiguity remains.

Acknowledgment. We would like to acknowledge the comments of a reviewer, which improved the presentation of this paper.

REFERENCES

[1] B. K. P. HORN, *Robot Vision*, MIT Press, Cambridge, MA, 1986.

[2] T. KAILATH, *Linear Systems*, Prentice-Hall, Englewood Cliffs, NJ, 1980.

[3] Y. LIU AND T. S. HUANG, *Estimation of rigid body motion using straight line correspondence*, in Proc. Workshop on Motion: Representation and Analysis, IEEE Computer Society Press, Piscataway, NJ, 1986.

[4] ———, *A linear algorithm for motion estimation using straight line correspondences*, Computer Vision, Graphics, and Image Processing, 44 (1988), pp. 35–57.

[5] O. D. FAUGERAS, F. LUSTMAN, AND G. TOSCANI, *Motion and structure from point and line matches*, INRIA, Le Chesnay, France, 1987, preprint.

[6] O. D. FAUGERAS, *On the Motion of 3D Curves and its Relation to Optical Flow*, in Proc. First European Conference on Computer Vision, O. Faugeras, ed., Lecture Notes in Computer Science, Vol. 427, Springer-Verlag, Berlin, Heidelberg, New York, 1990, pp. 107–117.

[7] O. D. FAUGERAS AND S. MAYBANK, *Motion from point matches: Multiplicity of solutions*, Int. J. Comput. Vision, 4 (1990), pp. 225–246.

[8] A. M. WAXMAN AND S. ULLMAN, *Surface structure and 3-D motion from image flow: Kinematic analysis*, Internat. J. Robotics Research, 4 (1985), pp. 72–94.

[9] S. ULLMAN, *The Interpretation of Visual Motion*, MIT Press, Cambridge, MA, 1979.

[10] K. I. KANATANI, *Detecting motion of a planar surface by line and surface integrals*, Computer Vision, Graphics, and Image Processing, 29 (1985), pp. 13–22.

[11] ———, *Group-Theoretical Methods in Image Understanding*, Springer-Verlag, New York, 1990.

[12] A. MITICHE, S. SEIDA, AND J. K. AGGARWAL, *Line based computation of structure and motion using angular invariance*, in Proc. Workshop on Motion: Representation and Analysis, IEEE Computer Society Press, Piscataway, NJ, 1986.

[13] M. E. SPETSAKIS AND J. ALOIMONOS, *Structure from motion using line correspondences*, Internat. J. Comput. Vision, 4 (1990), pp. 171–183.

[14] ———, *Optimal computing of structure from motion using point correspondences in two frames*, in Proc. Second Int. Conf. Comput. Vision, Tampa, FL, Dec. 1988, pp. 449–453.

[15] ———, *A multiframe approach to visual motion perception*, Internat. J. Comput. Vision, 6 (1991), pp. 245–255.

[16] J. W. ROACH AND J. K. AGGARWAL, *Determining the movement of objects from a sequence of images*, IEEE Trans. Patt. Anal. Mach. Intell., 6 (1980), pp. 554–562.

[17] H. H. NAGEL, *Representation of moving rigid objects based on visual observations*, Computing, 14 (1981), pp. 29–39.

[18] H. C. LONGUET-HIGGINS, *A computer algorithm for reconstructing a scene from two projections*, Nature, 293 (1981), pp. 133–135.

[19] R. Y. TSAI AND T. S. HUANG, *Estimating three dimensional motion parameters of a rigid planar patch*, IEEE Trans. Acoustics, Speech, and Signal Processing, 29 (1981), pp. 1147–1152.

[20] R. Y. TSAI, T. S. HUANG, AND W. L. ZHU, *Estimating three dimensional motion parameters of a rigid planar patch*, II: *Singular value decomposition*, IEEE Trans. Acoustics, Speech, and Signal Processing, 30 (1982), pp. 525–534; errata, 31 (1983), p. 514.

[21] R. Y. TSAI AND T. S. HUANG, *Uniqueness and estimation of three dimensional motion parameters of rigid objects with curved surfaces*, IEEE Trans. Patt. Anal. Mach. Intell., 6 (1984), pp. 13–26.

[22] ———, *Estimating* 3*-dimensional motion parameters of a rigid planar patch,* III: *Finite point correspondences and three views problem*, IEEE Trans. Acoustics, Speech, and Signal Processing, 32 (1984), pp. 213–220.

[23] X. ZHUANG, T. S. HUANG, AND R. M. HARALICK, *Two-view motion analysis: A unified algorithm*, J. Opt. Soc. Amer., A-3 (1986), pp. 1492–1500.

[24] X. Zhuang, *A simplification to linear two-view motion algorithm*, Computer Vision Graphics and Image Processing, 46 (1989), pp. 175–178.

[25] A. N. Netravali, T. S. Huang, A. S. Krishnakumar, and R. J. Holt, *Algebraic methods in 3-D motion estimation from two-view point correspondences*, Internat. J. Imaging Systems. Tech., 1 (1989), pp. 78–99.

[26] C. Jerian and R. Jain, *Polynomial methods for structure from motion*, IEEE Trans. Patt. Anal. Mach. Intell., 12 (1990), pp. 1150–1165.

[27] ———, *Structure from motion - A critical analysis of methods*, IEEE Trans. Systems Man Cybernet., 21 (1991), pp. 572–588.

[28] J. Weng, T. S. Huang, and N. Ahuja, *Motion and structure from two perspective views: Algorithms, error analysis and error estimation*, IEEE Trans. Patt. Anal. Mach. Intell., 11 (1989), pp. 451-467.

[29] N. M. Grzywacz and E. C. Hildreth, *Incremental rigidity scheme for recovering structure from motion: position-based versus velocity-based methods*, J. Opt. Soc. Amer., 4 (1987), pp. 503–518.

[30] D. W. Murray and B. F. Buxton, *Experiments in the machine interpretation of visual motion*, MIT Press, Cambridge, MA, 1990.

[31] S. Maybank, *Theory of reconstruction from image motion*, Springer Series in Information Sciences, Vol. 28, Springer Verlag, New York, 1993.

[32] G. Adiv, *Determining three dimensional motion and structure from optical flow generated by several objects*, IEEE Trans. Pattern Anal. Mach. Intell., 7 (1985), pp. 384–401.

[33] J. Aloimonos and C. M. Brown, *Preception of structure from motion,* I: *Optical flow v.s. discrete displacements*, in Proc. IEEE Conf. Computer Vision and Pattern Recognition, June 1986.

[34] S. Amari, *Feature spaces which admit and detect invariant signal transformation*, in Proc. 4th Int'l Joint Conf. on Pattern Recognition, 1978, pp. 452–456.

[35] ———, *Invariant structures and feature spaces in pattern recognition problems*, R.A.A.G., Memoirs 4 (1968), pp. 553–556.

[36] B. K. Ghosh, M. Jankovic, and Y. T. Wu, *Perspective problems in system theory and its application to machine vision*, J. Math. Systems Estim. Control, 4 (1994), pp. 3–38.

[37] B. K. Ghosh and E. P. Loucks, *A realization theory for perspective systems with applications to parameter estimation problems in machine vision*, IEEE Trans. Automat. Control, to appear.

[38] A. Isidori and C. I. Byrnes, *Output regulation of nonlinear systems*, IEEE Trans. Automat. Control, 35 (1990), pp. 131–140.

[39] J. K. Aggarwal and N. Nandhakumar, *On the computation of motion from sequences of images - A review*, Proc. IEEE, 76 (1988), pp. 917–935.

[40] A. Mitiche and J. K. Aggarwal, *Analysis of Time Varying Imagery, Handbook of Pattern Recognition and Image Processing*, T. Y. Young and K. S. Fu, eds., Academic Press, New York, 1986, pp. 311–332.

[41] J. K. Aggarwal, *Motion and time-varying imagery–An overview*, in Proceedings Workshop on Motion: Representation and Analysis, Charleston, SC, May 1986.

[42] B. K. P. Horn and M. J. Brooks, *Shape From Shading*, MIT Press, Cambridge, MA, 1989.

[43] R. J. Schalkoff, *Digital Image Processing and Computer Vision*, John Wiley and Sons, New York, 1989.

[44] J. Canny, *A Computational Approach to Edge Detection,* IEEE Trans. Patt. Anal. Mach. Intell., 8 (1986), pp. 679–698.

[45] D. Marr, *Vision*, W. H. Freeman and Company, San Francisco, 1982.

[46] A. Rosenfeld and M. Thurston, *Edge and curve detection for visual scene analysis*, IEEE Trans. Comput., C-20 (1971), pp. 562–569.

[47] R. W. Brockett, *Gramians, generalized inverses, and the least-squares approximation of optical flow*, Journal of Visual Communication and Image Representation, 1 (1990), pp. 3–11.

[48] C. J. Poelman and T. Kanade, *A paraperspective factorization method for shape and motion recovery*, preprint.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 33, No. 5, pp. 1560–1576, September 1995

©1995 Society for Industrial and Applied Mathematics

012

UNIQUENESS FOR VISCOSITY SOLUTIONS OF NONSTATIONARY HAMILTON–JACOBI–BELLMAN EQUATIONS UNDER SOME A PRIORI CONDITIONS (WITH APPLICATIONS)*

WILLIAM M. McENEANEY†

Abstract. This paper extends the uniqueness results for viscosity solutions of nonstationary Hamilton–Jacobi–Bellman equations. The conditions for uniqueness which are obtained can involve a trade-off between the growth of the solution and the growth of the Hamiltonian. In particular, the result is valid for solutions which grow quadratically in the space variable and are associated with Hamiltonians which also grow quadratically. This particular class arises in the robust control limit of risk-sensitive stochastic control problems.

Key words. viscosity solutions, uniqueness, Hamilton–Jacobi–Bellman equations, risk-sensitive control, robust control

AMS subject classifications. 49L25, 35B37, 93E20, 90D25, 60F10

1. Introduction.

The question of uniqueness for viscosity solutions of Hamilton–Jacobi–Bellman (HJB) equations has been considered in great detail in a number of papers, particularly by Ishii [14], Crandall and Lions [6], Ishii [15], and Crandall, Ishii, and Lions [4]. This latter paper [4] summarizes and generalizes much of the work of the other papers. In it both the stationary problem and the Cauchy problem are considered. The chief assumption on each solution is that it be continuous or even, as noted in a remark, only semicontinuous. (Then, however, some additional assumptions, such as the boundedness of the difference between the solutions being compared, may be required.) Some assumptions on the Hamiltonian are then employed to obtain a comparison principle.

More recently, some problems have arisen in risk-sensitive limits and H^∞ control for which the previous assumptions on the Hamiltonian are too strict to yield a comparison result. However, in these cases, the applications were such that one could obtain a priori bounds on the behavior of the solutions. In particular, growth conditions and Lipschitz conditions could be obtained. Once the set of functions being compared is reduced in this manner, the assumptions on the Hamiltonian can be considerably weakened. Since the solutions are then continuous, one obtains a uniqueness result. The approach used here is specific to nonstationary problems. That is, we consider HJB equations of the form

$$0 = U_t + H(t, x, U, \nabla U) \tag{1}$$

on $[0, T] \times E$, where $T < \infty$ and $E \subseteq \Re^n$ has nonempty interior. A remark at the end of §3 extends the result to problems on $[0, \infty) \times E$, although we do not concentrate on that here.

A common approach has been to consider first the related stationary problem, i.e.,

$$0 = U + H(x, U, \nabla U), \tag{2}$$

*Received by the editors February 4, 1994; accepted for publication May 8, 1994.

†Department of Mathematics, Carnegie Mellon University, Pittsburgh, Pennsylvania 15213.

and to apply a maximum principle type of argument to obtain a comparison principle for the stationary problem. A modification of that approach is then applied to the nonstationary problem (1). Thus the required conditions on the Hamiltonian for problem (1) are similar to those for (2). Here a cone of dependence argument is applied to the nonstationary problem directly. This type of approach was first used in the seminal paper of Crandall and Lions [5], although with the earlier definition of a viscosity solution. In the following, this approach is used to prove uniqueness for nonstationary problems in which one does not have uniqueness for the corresponding stationary problem. This will be discussed below.

2. Assumptions. Let $Q_T = [0,T] \times \Re^n$, $Q_T^R = [0,T] \times \{x \in \Re^n : |x| \leq R\}$, and $Q_T^E = [0,T] \times E$. We restrict our attention throughout to solutions (subsolutions and supersolutions) U that are continuous and uniformly locally Lipschitz in x. That is, given $R < \infty$, there exists $K_R < \infty$ such that

$$|U(t,x) - U(t,y)| \leq K_R|x-y|$$

for all $(t,x),(t,y) \in Q_T^R$. Denote this class by C_L. By Rademacher's theorem, $U \in C_L$ implies

$$\|\nabla U(t,\cdot)\|_{L^\infty(\{|x|\leq R\})} \leq K_R \qquad \forall t \in [0,T]. \tag{3}$$

Denote the K_R corresponding to a function U by $K_R(U)$. Let W be a viscosity subsolution to (1), and let V be a viscosity supersolution to (1). Assume that $V, W \in C_L$.

The weakest assumptions under which the results will be obtained here are as follows.

$$H(\cdot,\cdot,\cdot,\cdot) \text{ is continuous.} \tag{A}$$

(B)
$$\begin{aligned}
&\exists a < \infty \text{ such that } \forall R < \infty,\ \exists k_R < \infty \text{ such that if } \exists\ (t,x) \in Q_T^R \cap Q_T^E\\
&\text{such that } W(t,x) > V(t,x), \text{ then}\\
&H(t,x,V(t,x),p) - H(t,x,W(t,x),q) \leq k_R[W(t,x) - V(t,x)] + a(1+R)|p-q|\\
&\forall p,q \in \Re^n \text{ such that } |p|,|q| \leq \max\{K_R(V),K_R(W)\}.
\end{aligned}$$

There may be weaker assumptions obtainable by this method, but they will not be attempted here.

In order to clarify the class of problems satisfying assumption (B), we consider some stronger assumptions. For any $U \in C_L$ let $M_R(U) = \max_{Q_T^R}|U(t,x)|$. We replace assumption (B) by the following stronger pair of assumptions.

$$\begin{aligned}
&\text{Given } R < \infty,\ \exists k_R < \infty \text{ such that } H(t,x,r,p) - H(t,x,s,p) \geq -k_R|r-s|\\
&\forall\ (t,x) \in Q_T^R,\ \forall\,|p| \leq \max\{K_R(V),K_R(W)\},\\
&\forall\, r,s \in \Re \text{ such that } r > s \text{ and } |r|,|s| \leq \max\{M_R(V),M_R(W)\}.
\end{aligned} \tag{C1}$$

$$\begin{aligned}
&\exists a < \infty \text{ such that } \forall R < \infty \quad |H(t,x,r,p) - H(t,x,r,q)| \leq a(1+R)|p-q|\\
&\forall\ (t,x) \in Q_T^R,\ \forall\,|r| \leq \max\{M_R(V),M_R(W)\},\\
&\forall\,|p|,|q| \leq \max\{K_R(V),K_R(W)\}.
\end{aligned} \tag{C2}$$

To obtain (B) from (C1) and (C2), consider the following. By (C2) there exists $a < \infty$ such that for all $R < \infty$

$$|H(t,x,r,p) - H(t,x,r,q)| \leq a(1+R)|p-q|$$

for all t, x, r, p, q as specified in (C2). In particular, this holds if $(t,x) \in Q_T^R$ is such that $W(t,x) > V(t,x)$ and $r = V(t,x)$. In this case we have

$$H(t,x,V(t,x),p) - H(t,x,V(t,x),q) \leq a(1+R)|p-q|. \tag{4}$$

Now, by (C1), there exists $k_R < \infty$ such that

$$H(t,x,W(t,x),q) - H(t,x,V(t,x),q) \geq -k_R|W(t,x) - V(t,x)|. \tag{5}$$

Subtracting (5) from (4), we have (B).

Note that (C2) is essentially a Lipschitz bound on H with respect to its last argument, where the Lipschitz constant may not grow faster than linearly in x. As we discuss below, this involves a trade-off between the growth of H and the growth of the solutions via the dependence of H on the solutions.

The assumptions could be made even stronger by retaining (A) and (C2) and replacing (C1) by

(D1) $\qquad H(t,x,\cdot,p)$ is nondecreasing $\forall\,(t,x,p) \in Q_T \times \Re^n$.

To obtain (C1) from (D1), simply note that (D1) implies that if $r > s$, then

$$H(t,x,r,p) - H(t,x,s,p) \geq 0 \geq -k_R|r-s|.$$

In particular, (C1) states that $H(t,x,\cdot,p)$ may not decrease faster than linearly on compact sets. However, it may still increase as fast as it likes.

We now consider some examples fitting the above assumptions.

Consider the case where

$$H(t,x,r,p) = -|p|^2$$

and $V, W \in C_L$ are such that

$$K_R(V) \leq \hat{a}(1+R),$$
$$K_R(W) \leq \hat{a}(1+R)$$

for all $R < \infty$ (i.e., V, W grow at most quadratically). Then for any $R < \infty$

$$\begin{aligned}|H(t,x,r,p) - H(t,x,r,q)| &\leq (|p| + |q|)|p-q| \\ &\leq 2\hat{a}(1+R)|p-q|\end{aligned}$$

for all $|p|, |q| \leq \max\{K_R(V), K_R(W)\}$. Thus assumption (C2) is satisfied. Assumptions (A) and (D1) are also clearly satisfied.

More generally, consider the case where for some $\gamma \in (1,\infty)$

$$H(t,x,r,p) = -|p|^\gamma \tag{6}$$

and

$$K_R(V) \leq \hat{a}(1+R)^{\frac{1}{\gamma-1}}, \tag{7}$$
$$K_R(W) \leq \hat{a}(1+R)^{\frac{1}{\gamma-1}}$$

for all $R < \infty$. Then, again, assumptions (A), (D1), and (C2) would be satisfied, and the uniqueness result below would follow. Interestingly, however, one does not have a comparison/uniqueness result for the corresponding stationary problem in this case. This was noted in Ishii [14], and the counterexample provided there is as follows. Consider the one-dimensional case $n = 1$, and let $E = \Re$. Then (2) with Hamiltonian (6) has two solutions:

$$U^1 = 0$$

and

$$U^2 = (\gamma^*)^{\gamma^*} |x|^{\gamma^*},$$

where $\gamma^* = \frac{\gamma}{\gamma-1}$. Both solutions satisfy (7). Thus one would not expect the method used to obtain a comparison principle for the stationary problem to yield the result below.

Also consider

$$H(t,x,r,p) = x^T p$$

so that for any $R < \infty$

$$|H(t,x,r,p) - H(t,x,r,q)| \leq R|p-q|$$

for all $(t,x) \in Q_T^R$. Thus (A), (D1), and (C2) would hold without a growth restriction on V or W (other than $V, W \in C_L$).

In §4, we will consider an application to the case where

$$H(t,x,p) = -\frac{1}{4\gamma^2}|p|^2 - \min_{v\in U}\left[f(t,x,v)\cdot p + L(t,x,v)\right], \tag{8}$$

where f grows at most linearly in x, L grows at most quadratically in x, and U is compact. This is related to risk-sensitive control and the robust limit. We will also consider an application to the case where

$$H(x,r,p) = -\max_{v\in U}\left[-\frac{B^2(v)}{2r}x^2p^2 + A(v)xp\right], \tag{9}$$

where A and B are continuous and U is compact. This is related to a risk-sensitive limit in a Merton porfolio problem. Neither of these applications appears to fit within the framework of existing results.

Finally, note that these extensions of the comparison/uniqueness results do not come entirely for free. It is now required that the solutions (subsolutions and supersolutions) be in C_L. This implies that one could not trivially apply these results in cases where the solutions are obtained via the Barles and Perthame procedure (see Barles and Perthame [2] and Fleming and Soner [13]), where one has only semicontinuity of the solutions. In particular, although the comparison/uniqueness results are extended to problems with more rapidly growing solutions and Hamiltonians, the price to be paid is that one must obtain the local Lipschitz bounds on the solutions. Of course, whereas the growth rates for certain problems of interest are not within our control, one may with sufficient effort be able to obtain the required local Lipschitz conditions.

3. Comparison/uniqueness result. In this section we derive the main comparison result which, under the assumptions here, directly yields a uniqueness result. Reiterating from above, we suppose that $V, W \in C_L$ are a viscosity supersolution and a viscosity subsolution, respectively, of (1) on Q_T^E. Further, we assume that

$$W \leq V \qquad \text{on} \quad (\{t=0\} \times E) \cup ([0,T] \times \partial E)\,. \tag{10}$$

(As mentioned above, E may be all of $\Re^n$, in which case ∂E is empty.)

The following two lemmas lead to the result. The statements are similar in form to the corresponding lemmas in Crandall and Lions [5].

LEMMA 1. *Let W, V be as above, and assume* (A) *and* (B). *Let $R < \infty$. Let $\Lambda \in C^1(Q_T)$, $\Lambda \geq 0$, $\Lambda(t,x) = 0$ if $|x| \geq R$, and*

$$-\Lambda_t > a(1+R)|\Lambda_x| + k_R\Lambda \qquad on \quad (\text{supp}\Lambda)^\circ \cap (Q_T^E)^\circ. \tag{11}$$

Then $W \leq V$ on $(\text{supp}\Lambda) \cap Q_T^E$.

Proof. Let $\phi \in C^\infty(\Re^n)$, $\text{supp}(\phi) \subset B(0,1)$, $\phi(0) = 1$, and $0 \leq \phi \leq 1$. Let $\eta \in C^\infty(\Re)$, $\text{supp}(\eta) \subset B(0,1)$, $\phi(0) = 1$, and $0 \leq \phi \leq 1$. Define $\phi_\alpha(x) = \phi(x/\alpha)$ and $\eta_\alpha(t) = \eta(t/\alpha)$.

Assume that

$$M_0 \equiv \max_{Q_T^E} \Lambda(t,x)[W(t,x) - V(t,x)] > 0. \tag{12}$$

(Otherwise there is nothing to prove.) Let $C > C_0$, where

$$C_0 \equiv \max_{Q_T^E \times Q_T^E} \Lambda(s,y)|W(t,x)| + \Lambda(t,x)|V(s,y)|. \tag{13}$$

Consider

$$M_\alpha \equiv \max_{Q_T^E \times Q_T^E} [\Lambda(s,y)W(t,x) - \Lambda(t,x)V(s,y) + C\eta_\alpha(t-s)\phi_\alpha(x-y)]\,, \tag{14}$$

and let the maximum occur at $(t_\alpha, x_\alpha), (s_\alpha, x_\alpha)$. By the definition of C, η_α, and ϕ_α we have

$$|t_\alpha - s_\alpha| \leq \alpha, \quad |x_\alpha - y_\alpha| \leq \alpha, \tag{15}$$

and by (13), (14)

$$\eta_\alpha(t_\alpha - s_\alpha) \geq \min\left\{1, \frac{C - C_0}{C}\right\}, \quad \phi_\alpha(x_\alpha - y_\alpha) \geq \min\left\{1, \frac{C-C_0}{C}\right\}. \tag{16}$$

By the compactness of Q_T^E, there exists a sequence $\alpha_n \downarrow 0$ such that $(t_{\alpha_n}, x_{\alpha_n})$ converges to some $(t_0, x_0) \in Q_T^E$. By (15), we also have $(s_{\alpha_n}, y_{\alpha_n}) \to (t_0, x_0)$.

It is easily shown that

$$\Lambda(t_0,x_0)[W(t_0,x_0) - V(t_0,x_0)] = M_0 > 0.$$

Therefore $(t_0, x_0) \in (\mathrm{supp}\Lambda)^\circ \cap (Q_T^E)^\circ$, where the superscript $\circ$ indicates interior. This implies that for n sufficiently large,

$$(t_{\alpha_n}, x_{\alpha_n}) \in (\mathrm{supp}\Lambda)^\circ \cap (Q_T^E)^\circ,$$
$$(s_{\alpha_n}, y_{\alpha_n}) \in (\mathrm{supp}\Lambda)^\circ \cap (Q_T^E)^\circ.$$

Therefore, for n sufficiently large,

$$(t_{\alpha_n}, x_{\alpha_n}), (s_{\alpha_n}, y_{\alpha_n})$$

is a local, unconstrained maximizer of (14).

Consider

$$(17)\quad W(t,x) - \frac{1}{\Lambda(s_{\alpha_n}, y_{\alpha_n})}[\Lambda(t,x)V(s_{\alpha_n}, y_{\alpha_n}) - C\eta_{\alpha_n}(t - s_{\alpha_n})\phi_{\alpha_n}(x - y_{\alpha_n}) - M_{\alpha_n}],$$

which has a local maximum of zero at $(t_{\alpha_n}, x_{\alpha_n})$. Since W is a subsolution of (1), we have

$$(18)\quad \begin{aligned} &\frac{\Lambda_t(t_{\alpha_n}, x_{\alpha_n})V(s_{\alpha_n}, y_{\alpha_n}) - C\eta'_{\alpha_n}(t_{\alpha_n} - s_{\alpha_n})\phi_{\alpha_n}(x_{\alpha_n} - y_{\alpha_n})}{\Lambda(s_{\alpha_n}, y_{\alpha_n})} \\ &+ H\Bigg(t_{\alpha_n}, x_{\alpha_n}, W(t_{\alpha_n}, x_{\alpha_n}), \\ &\qquad \frac{\Lambda_x(t_{\alpha_n}, x_{\alpha_n})V(s_{\alpha_n}, y_{\alpha_n}) - C\eta_{\alpha_n}(t_{\alpha_n} - s_{\alpha_n})\phi'_{\alpha_n}(x_{\alpha_n} - y_{\alpha_n})}{\Lambda(s_{\alpha_n}, y_{\alpha_n})}\Bigg) \le 0 \end{aligned}$$

where the superscript $'$ indicates differentiation.

Similarly (noting that V is a supersolution), we obtain

$$(19)\quad \begin{aligned} &\frac{\Lambda_s(s_{\alpha_n}, y_{\alpha_n})W(t_{\alpha_n}, x_{\alpha_n}) - C\eta'_{\alpha_n}(t_{\alpha_n} - s_{\alpha_n})\phi_{\alpha_n}(x_{\alpha_n} - y_{\alpha_n})}{\Lambda(t_{\alpha_n}, x_{\alpha_n})} \\ &+ H\Bigg(s_{\alpha_n}, y_{\alpha_n}, V(s_{\alpha_n}, y_{\alpha_n}), \\ &\qquad \frac{\Lambda_y(s_{\alpha_n}, y_{\alpha_n})W(t_{\alpha_n}, x_{\alpha_n}) - C\eta_{\alpha_n}(t_{\alpha_n} - s_{\alpha_n})\phi'_{\alpha_n}(x_{\alpha_n} - y_{\alpha_n})}{\Lambda(t_{\alpha_n}, x_{\alpha_n})}\Bigg) \ge 0. \end{aligned}$$

We would like to take limits as $\alpha_n \downarrow 0$; however, we must ensure that the limits exist for all terms in (18) and (19). Specifically, we will show that there exist subsequential limits for $\eta'_{\alpha_n}(t_{\alpha_n} - s_{\alpha_n})$ and $\phi'_{\alpha_n}(x_{\alpha_n} - y_{\alpha_n})$.

Consider the last term in (18), i.e.,

$$\frac{\Lambda_x(t_{\alpha_n}, x_{\alpha_n})V(s_{\alpha_n}, y_{\alpha_n}) - C\eta_{\alpha_n}(t_{\alpha_n} - s_{\alpha_n})\phi'_{\alpha_n}(x_{\alpha_n} - y_{\alpha_n})}{\Lambda(s_{\alpha_n}, y_{\alpha_n})}.$$

Since this is being evaluated at a maximum of (14), we have

$$(20)\quad \left|\frac{\Lambda_x(t_{\alpha_n}, x_{\alpha_n})V(s_{\alpha_n}, y_{\alpha_n}) - C\eta_{\alpha_n}(t_{\alpha_n} - s_{\alpha_n})\phi'_{\alpha_n}(x_{\alpha_n} - y_{\alpha_n})}{\Lambda(s_{\alpha_n}, y_{\alpha_n})}\right| \le K_R(W).$$

Then, since Λ, Λ_x, and V are bounded on $Q_T^R \times Q_T^R$ and (16) implies that η_{α_n} is bounded away from zero, we have some $B_1 < \infty$ such that

$$|\phi'_{\alpha_n}(x_{\alpha_n} - y_{\alpha_n})| \leq B_1$$

for all n sufficiently large. Therefore, there exists $l_1 \in \Re^n$ such that $|l_1| \leq B_1$ and

$$\phi'_{\alpha_n}(x_{\alpha_n} - y_{\alpha_n}) \to l_1$$

along some subsequence (also denoted by α_n).

Thus, using assumption (A) (continuity of H), we see that all terms in (18) are bounded, with the possible exception of

$$\frac{C\eta'_{\alpha_n}\phi_{\alpha_n}}{\Lambda(s_{\alpha_n}, y_{\alpha_n})}.$$

This implies that there exists some $B_2 > -\infty$ such that

$$\frac{C\eta'_{\alpha_n}\phi_{\alpha_n}}{\Lambda(s_{\alpha_n}, y_{\alpha_n})} \geq B_2$$

for all n sufficiently large, which upon noting (16) implies

$$\eta'_{\alpha_n} \geq \hat{B}_2 > -\infty \tag{21a}$$

for all n sufficiently large. Proceeding similarly with (19), we have

$$\eta'_{\alpha_n} \leq \hat{B}_3 < \infty \tag{21b}$$

for all n sufficiently large. By inequalities (21), there exists $l_2 \in \Re$ such that $\hat{B}_2 \leq l_2 \leq \hat{B}_3$ and

$$\eta'_{\alpha_n}(t_{\alpha_n} - s_{\alpha_n}) \to l_2$$

along some further subsequence (also denoted by α_n).

We may now take the limit in (18) and (19) as $n \to \infty$ and employ (A1) again to obtain

$$\frac{\Lambda_t V - Cl_2}{\Lambda} + H\left(t_0, x_0, W, \frac{\Lambda_x V - Cl_1}{\Lambda}\right) \leq 0, \tag{22}$$

$$\frac{\Lambda_s W - Cl_2}{\Lambda} + H\left(t_0, x_0, V, \frac{\Lambda_y W - Cl_1}{\Lambda}\right) \geq 0, \tag{23}$$

where all terms are evaluated at (t_0, x_0).

Subtracting (23) from (22), we have

$$\begin{aligned} &\frac{\Lambda_t}{\Lambda}(V - W) + H\left(t_0, x_0, W, \frac{\Lambda_x V - Cl_1}{\Lambda}\right) \\ &\qquad - H\left(t_0, x_0, V, \frac{\Lambda_x W - Cl_1}{\Lambda}\right) \leq 0. \end{aligned} \tag{24}$$

By assumption (B), (12), and (20) (and its counterpart for V), this yields

$$\frac{\Lambda_t}{\Lambda}(V - W) \leq a(1 + R)\left|\frac{\Lambda_x}{\Lambda}(V - W)\right| + k_R(W - V)$$

at (t_0, x_0), or equivalently,

$$\frac{\Lambda_t}{\Lambda}(V - W) \leq a(1+R)\frac{|\Lambda_x|}{\Lambda}|V - W| + k_R(W - V),$$

which, by (12), implies

$$-\Lambda_t \leq a(1+R)|\Lambda_x| + k_R\Lambda$$

at (t_0, x_0). But this contradicts the assumption of the lemma. Therefore

$$\max_{Q_T^E \times Q_T^E} \Lambda(t,x)[W(t,x) - V(t,x)] \leq 0,$$

which is the desired result. □

LEMMA 2. *Let V and W be as above, and let assumptions* (A) *and* (B) *be satisfied. Let $R < \infty$. Then*

$$W \leq V \tag{25}$$

on

$$Q_T^E \cap \{(t,x) : |x| \leq R - a(1+R)t,\ t \in [0, \min(T, \tau_R)]\},$$

where $\tau_R = \frac{R}{a(1+R)}$.

Proof. Let $g_\gamma \in C^1(\Re)$ for γ sufficiently small. For $D > 0$, let

$$g_\gamma(r) = 0 \qquad \text{if } r \leq 0$$

and

$$\gamma g'_\gamma(r) \geq g_\gamma(r) > 0 \qquad \text{if } r \in (0, D].$$

To see that such g_γ exist, take for example

$$g_\gamma(x) = \begin{cases} \gamma\left(e^{x/\gamma} - 1\right) - x & \text{if } x \in (0, D], \\ 0 & \text{if } x \leq 0. \end{cases} \tag{26}$$

Choose $D \geq R_0 > 0$, and let

$$\Lambda(t,x) = g_\gamma\left(R_0 - a(1+R)t - \lambda|x|^{1+\beta}\right), \tag{27}$$

where $\beta > 0$ and λ is yet to be specified, so that

$$\begin{aligned} -\Lambda_t &= a(1+R)g'_\gamma\left(R_0 - a(1+R)t - \lambda|x|^{1+\beta}\right), \\ a(1+R)|\Lambda_x| &= a(1+R)\lambda(1+\beta)|x|^\beta g'_\gamma\left(R_0 - a(1+R)t - \lambda|x|^{1+\beta}\right), \\ k_R\Lambda &= k_R g_\gamma \leq \gamma k_R g'_\gamma\left(R_0 - a(1+R)t - \lambda|x|^{1+\beta}\right). \end{aligned} \tag{28}$$

We would like the conditions of Lemma 1 to be satisfied by this choice of Λ. In order that Λ be zero if $|x| \geq R$, we require $R_0 - \lambda R^{1+\beta} \leq 0$, or equivalently,

$$(\lambda^{-1}R_0)^{\frac{1}{1+\beta}} \leq R. \tag{29}$$

In order that $-\Lambda_t > a(1+R)|\Lambda_x| + k_R\Lambda$, by (27), (28), and the properties of g_γ, we require

$$1 > \lambda(1+\beta)\left(\lambda^{-1}R_0\right)^{\frac{1}{1+\beta}} + \gamma\frac{k_R}{a(1+R)}. \tag{30}$$

Let

$$(1+2\beta)R_0^{1+\beta} = R^{1+\beta}. \tag{31}$$

Then conditions (29) and (30) become

$$\lambda \geq \frac{1}{R^{\beta}(1+2\beta)^{\frac{1}{1+\beta}}} \tag{32}$$

and

$$1 > \lambda^{\frac{1}{1+\beta}}(1+\beta)R^{\frac{\beta}{1+\beta}}(1+2\beta)^{\frac{-\beta}{(1+\beta)^2}} + \gamma\frac{k}{a(1+R)}, \tag{33}$$

respectively. We take

$$\lambda = \frac{1}{R^{\beta}(1+2\beta)^{\frac{1}{1+\beta}}} \tag{34}$$

so that (32) is satisfied. Substituting (34) into (33), the last requirement becomes

$$1 > \frac{1+\beta}{(1+2\beta)^{\frac{1}{1+\beta}}} + \gamma\frac{k_R}{a(1+R)}. \tag{35}$$

The first term on the right-hand side of (35) is less than one and converges to one as $\beta \downarrow 0$. This implies that if we take

$$\gamma = \frac{1}{2}\frac{a(1+R)}{k_R}\left[1 - \frac{1+\beta}{(1+2\beta)^{\frac{1}{1+\beta}}}\right], \tag{36}$$

(35) will be satisfied.

Now note that as $\beta \downarrow 0$, we have $\lambda \to 1$, $R_0 \to R$, and $\gamma \to 0$. This implies that any $(t,x) \in \{(t,x) : |x| \leq R - a(1+R)t,\ t \in [0,\tau_R]\}$ is in $\operatorname{supp}(\Lambda)$ for sufficiently small β. Therefore, by Lemma 1, we have the desired result. □

THEOREM 3. *Let V and W be as above, and let assumptions* (A) *and* (B) *be satisfied. Then $W \leq V$ on Q_T^E.*

Proof. By Lemma 2, $W \leq V$ on $Q_T^E \cap \{(t,x) : |x| \leq R - a(1+R)t,\ t \in [0,\tau_R]\}$.

Letting $R \to \infty$, we see that

$$\tau_R \uparrow \tau_\infty = \frac{1}{a}.$$

This implies that as $R \to \infty$, the cones cover $[0,\tau_\infty) \times \Re^n$. Therefore, given any $\delta > 0$, $W \leq V$ on $[0, \min\{T, \tau_\infty - \delta\}] \times E$.

If $\tau_\infty - \delta \geq T$, we are finished. If not, we simply repeat the procedure, with the new initial time being $\tau_\infty - \delta$. This yields $W \leq V$ on $[0, \min\{T, 2(\tau_\infty - \delta)\}] \times E$. Iterating, we obtain the desired result. □

Under our assumptions, the uniqueness result is a trivial corollary to the comparison result.

COROLLARY 4. *Let $U_1, U_2 \in C_L$ be two viscosity solutions of* (1) *with some initial condition $U = g(t,x)$ on $(\{0\} \times E) \cup ([0,T] \times \partial E)$, where $g \in C_L$. Let assumptions* (A) *and* (B) *be satisfied. Then $U_1 = U_2$ on $[0,T] \times E$.*

Remark 5. By noting that $\tau_\infty = \frac{1}{a}$, the comparison and uniqueness results can clearly be extended to $[0,\infty) \times E$ if, for each $T < \infty$, there exists $a < \infty$ such that (B) is satisfied. This yields uniqueness for the Cauchy problem.

Remark 6. An extension in the spirit of the semicontinuity results is possible. In particular, one can weaken the assumption that W, V are in C_L to a one-sided Lipschitz assumption on each. That is, we may assume that $W \in USL$ and $V \in LSL$, which we define as follows. We say that $U \in USC$ is in USL if, given $R < \infty$, there exists $L_R < \infty$ such that, for any $(t_0, x_0) \in Q_T^R$, there exists $\delta > 0$ (depending on (t_0, x_0)) such that

$$U(t,x) \leq U(t_0, x_0) + L_R(|t-t_0| + \|x - x_0\|) \tag{37}$$

for all (t,x) in a ball of radius δ around (t_0, x_0). $LSL \subset LSC$ is defined analogously by changing the direction of the inequality and the sign preceding L_R in (37). The proof of Lemma 2 is unchanged by these assumptions. The proof of Lemma 1 requires some minor modifications. The only nontrivial modification in the proof is in obtaining the bound on $\phi'_{\alpha_n}(x_{\alpha_n} - y_{\alpha_n})$. If we take $\phi(x) = \exp(\frac{-|x|^2}{1-|x|^2})$ and use the one-sided Lipschitz condition, a bound is obtained. However, at present it is not clear that this extension leads to a measurable change in where or how the result may be applied.

4. Two applications. In this section, the two applications mentioned at the end of §2 are discussed.

4.1. Risk-sensitive control and the robust limit. Consider the following finite time-horizon, risk-sensitive stochastic control problem as discussed in Fleming and McEneaney [10], McEneaney [19], and James [18]. Let the dynamics be given by

$$\begin{aligned} dX_t^\epsilon &= f(t, X_t^\epsilon, u_t)\, dt + \sqrt{\frac{\epsilon}{2\gamma^2}}\, dB_t, \\ X_s^\epsilon &= x, \end{aligned}$$

where X_t^ϵ taking values in $\Re^n$ is the state, $u.$ is the control, and $B.$ is an n-dimensional Brownian motion with respect to some reference probability system $(\Omega, \{\mathcal{F}_t\}, P, B.)$. Let ϵ and γ be parameters.

The cost criterion is

$$J^\epsilon(s,x,u) = \epsilon \log \mathrm{E}_{s,x} \exp\left\{ \frac{1}{\epsilon} \left[\int_s^T L(t, X_t^\epsilon, u_t)\, dt + \Psi(X_T^\epsilon) \right] \right\}.$$

The value function is given by

$$V^\epsilon(s,x) = \inf_{u \in \mathcal{U}} J^\epsilon(s,x,u).$$

Here $\mathcal{U}$ is the set of F_t-progressively measurable controls (see [13]) taking values in some compact set U, L is the running cost, and Ψ is the terminal cost.

Make the following assumptions on f, L, and Ψ.

$$\text{(AF1)} \quad \begin{aligned} &f \in C([0,T] \times \Re^n \times U), \\ &|f(t,x,v) - f(t,y,v)| \leq K|x-y| \qquad \forall t \in [0,T],\ \forall x,y \in \Re^n,\ \forall v \in U; \end{aligned}$$

(AF2)
$L \in C([0,T] \times \Re^n \times U)$,
$L(t,x,v) \geq 0 \qquad \forall t \in [0,T],\ \forall x \in \Re^n,\ \forall v \in U$,
$|L(t,x,v) - L(t,y,v)| \leq C(1+|x|+|y|)|x-y| \quad \forall t \in [0,T],\ \forall x,y \in \Re^n,\ \forall v \in U$;

$$\begin{aligned} &\Psi \in C(\Re^n),\\ &\Psi(x) \geq 0 \qquad \forall x \in \Re^n,\\ &|\Psi(x) - \Psi(y)| \leq C(1+|x|+|y|)|x-y| \qquad \forall x,y \in \Re^n, \end{aligned} \tag{AF3}$$

where K and C are generic constants.

It can then be shown (see [19]) that V^ϵ is the solution of

$$\begin{aligned} 0 &= V_s^\epsilon + \frac{\epsilon}{4\gamma^2}\Delta V^\epsilon + \frac{1}{4\gamma^2}|\nabla V^\epsilon|^2 + \min_{v\in U}[f(s,x,v)\cdot\nabla V^\epsilon + L(s,x,v)],\\ V^\epsilon(T,x) &= \Psi(x). \end{aligned}$$

Furthermore,

$$\begin{aligned} &0 \leq V^\epsilon(s,x) \leq C_1 + C_2|x|^2,\\ &|V^\epsilon(s,x) - V^\epsilon(s,y)| \leq [C_3 + C_4 R]|x-y| \ \ \forall |x|,|y| \leq R,\ \forall R < \infty,\\ &|V^\epsilon(s,x) - V^\epsilon(t,x)| \leq g_R(|s-t|) \ \ \forall |x|,|y| \leq R,\ \forall R < \infty, \end{aligned} \tag{38}$$

where $g_R(\rho) \downarrow 0$ as $\rho \downarrow 0$. Here, the C_i represent generic constants.

By the Ascoli–Arzela theorem and the stability property of viscosity solutions [13], there exists a sequence $\{\epsilon_n\}$ such that $\epsilon \downarrow 0$ and $V^{\epsilon_n} \to V^0$ (uniformly on compact sets) where V^0 is a viscosity solution to the limit partial differential equation (PDE)

$$\begin{aligned} 0 &= V_s + H(s,x,\nabla V),\\ V(T,x) &= \Psi(x), \end{aligned} \tag{39}$$

where

$$H(s,x,p) = \frac{1}{4\gamma^2}|p|^2 + \min_{v\in U}[f(s,x,v)\cdot p + L(s,x,v)].$$

Clearly V^0 satisfies (38) as well.

Note that the PDE in (39) is equivalent to (8) with the exception of a time reversal due to the presence of a terminal condition rather than an initial condition.

Also consider the following finite time-horizon deterministic differential game. Let the dynamics be

$$\begin{aligned} \frac{dX}{dt} &= f(t,X_t,u_t) + w_t,\\ X_s &= x, \end{aligned}$$

and let the payoff be

$$P(s,x,u,w) = \int_s^T [L(t,X_t,u_t) - \gamma^2|w_t|^2]\, dt + \Psi(X_T).$$

The minimizing player's control is $u.$, and the maximizing player's control is $w.$. We require that $u.$ be measurable (with respect to t) and take values in U. Let this control set be denoted by $\mathcal{U}^0$. For the maximizing player, we let the control set be $\mathcal{W}^0 = L^2([s,T];\Re^n)$. The Elliott–Kalton definition of value [7] is used. In this context, we define the set of strategies for the minimizing player to be

$$\Theta(s) = \{\theta : \mathcal{W}^0 \to \mathcal{U}^0 \text{ such that given any } \tau \in [s,T],\ w_t = \hat{w}_t\ \forall t \in [s,\tau] \\ \text{implies } \theta[w]_t = \theta[\hat{w}]_t\ \forall t \in [s,\tau]\}$$

and the set of strategies for the maximizing player to be

$$\Lambda(s) = \{\lambda : \mathcal{U}^0 \to \mathcal{W}^0 \text{ such that given any } \tau \in [s,T],\ u_t = \hat{u}_t\ \forall t \in [s,\tau] \\ \text{implies } \lambda[u]_t = \lambda[\hat{u}]_t\ \forall t \in [s,\tau]\}.$$

The lower and upper values are defined to be

$$W(s,x) = \inf_{\theta \in \Theta(s)} \sup_{w \in \mathcal{W}^0} P(s,x,\theta[w],w)$$

and

$$U(s,x) = \sup_{\lambda \in \Lambda(s)} \inf_{u \in \mathcal{U}^0} P(s,x,u,\lambda[u]),$$

respectively. The game has value if $W = U$.

Generalizing a result in Evans and Souganidis [8], it is shown in McEneaney [19] that this game has value in the Elliott–Kalton sense and that this value is a viscosity solution of (39). Furthermore, $W = U$ satisfies conditions (38) (with proper choice of the constants).

Now, suppose that we have a unique solution, V^0, of (39). Then it is easy to show that $V^\epsilon \to V^0$ (not just along a sequence) and that $V^0 = W = U$. That is, the value of the risk-sensitive stochastic control problem converges to the value of the deterministic game. Note (see [19], James [17]) that this game is related to robust control.

All that remains is to verify that (39) satisfies the assumptions of §2 which were shown to be sufficient to yield uniqueness.

Assumptions (A) and (D1) obviously hold. All that remains is to show that (C2) holds. Let $r = T - s$ so that the problem becomes an initial value problem rather than a terminal value problem. Then

$$H(r,x,p) - H(r,x,q) = \frac{1}{4\gamma^2}\left[|q|^2 - |p|^2\right] + \min_{v \in U}[f(T-r,x,v)\cdot q + L(T-r,x,v)] \\ - \min_{v \in U}[f(T-r,x,v)\cdot p + L(T-r,x,v)].$$

Let $v_0 \in \operatorname{argmin}_{v \in U}[f(T-r,x,v)\cdot p + L(T-r,x,v)]$. Then

$$H(r,x,p) - H(r,x,q) \leq \frac{1}{4\gamma^2}\left[|q| + |p|\right]\left[|q| - |p|\right] + f(T-r,x,v_0)\cdot(q-p),$$

which by (AF1)

$$\leq \frac{1}{4\gamma^2}\left[|q| + |p|\right]\left[|q| - |p|\right] + C_1(1+R)|q-p| \tag{40}$$

for all $|x| \leq R$ where C_1 is a generic constant. By the second inequality of (38), there exists $a < \infty$ such that $K_R(V^0) \leq a(1+R)$ and $K_R(W) \leq a(1+R)$ so that (40) implies

$$H(r,x,p) - H(r,x,q) \leq \left(\frac{a}{2\gamma^2} + C_1\right)(1+R)|p-q|$$

for all $|p|, |q| \leq \max\{K_R(V^0), K_R(W)\}$. Switching the order of p and q, one obtains the analogous bound from below. Therefore, assumption (C2) is satisfied, and the uniqueness result can be applied.

The uniqueness result also has an application to the infinite time-horizon risk-sensitive limit analogous to the finite time-horizon limit discussed above. In particular, it can be shown that the value function of the infinite time-horizon risk-sensitive control problem, under sufficiently stringent conditions, converges to a viscosity solution $V^0(x)$ of

$$0 = H^0(x, \nabla V),$$

where

$$H^0(x,p) = \frac{1}{4\gamma^2}|p|^2 + \min_{v\in U}[f(x,v)\cdot p + L(x,v)].$$

See Fleming and McEneaney [11], [12]. (In [12], this is proved for globally Lipschitz L; however, it is anticipated that the result also holds for L exhibiting quadratic growth as in (AF2).)

Then V^0 is a (steady-state) solution to the finite time-horizon PDE

$$\begin{aligned} 0 &= V_s + H^0(x, \nabla V), \\ V(T,x) &= V^0(x) \end{aligned}$$

for any $T < \infty$. Again V^0 satisfies a growth condition analogous to the middle inequality of (38). Then, by the same uniqueness result as used above, one sees that V^0 is also the value of the corresponding finite time-horizon deterministic differential game, that is,

$$V^0(x) = \inf_{\theta\in\Theta(0)} \sup_{w\in\mathcal{W}^0} \left\{ \int_0^T [L(X_t,u_t) - \gamma^2|w_t|^2]\, dt + V^0(X_T) \right\}$$

for any $T < \infty$. This is the dissipation relation used to define nonlinear H^∞ control (see van der Schaft [20], James [17], Ball and Helton [1], and Isidori [16]).

4.2. Risk-sensitive limit in an optimal investment model. In this section, we consider an application leading to a Hamiltonian of the form (9). The following optimal investment problem is discussed in Fleming [9]. The dynamics are of the form

$$\begin{aligned} dX_t &= A(u_t)X_t\, dt + |\gamma|^{-1/2}B(u_t)X_t\, dB_t, \\ X_0 &= x. \end{aligned} \tag{41}$$

Here X_t is the wealth at time t, $u.$ is the control, and $B.$ is Brownian motion with respect to some reference probability system $(\Omega, \{\mathcal{F}_t\}, P, B.)$. Assume that $u. \in \mathcal{U}$, where $\mathcal{U}$ is the set of $\mathcal{F}_t$-progressively measurable controls taking values in some compact set U. Assume that A and B are continuous functions. γ will be a parameter measuring risk sensitivity. The risk-sensitive value at time T is given by

$$V^\gamma(T,x) = \sup_{u\in\mathcal{U}} F_\gamma^{-1}\left[E_{0,x}F_\gamma\left(h_\gamma(x)\right)\right], \tag{42}$$

where F_γ has hyperbolic absolute risk aversion, that is,

$$F_\gamma(r) = \frac{1}{\gamma}r^\gamma, \quad \gamma < 1, \quad \gamma \neq 0, \tag{43}$$

and h_γ will be a perturbation of a linear function (to be discussed below). This type of risk-sensitive utility function is also described in Barron and Jensen [3]. We are interested in the limit as $\gamma \to -\infty$. The dynamic programming equation corresponding to (41)–(43) is

(44)

$$0 = V_T^\gamma - \max_{v\in U}\left[\frac{B^2(v)}{2|\gamma|}x^2V_{xx}^\gamma + A(v)xV_x^\gamma - \left|\frac{\gamma-1}{\gamma}\right|\frac{B^2(v)}{2V^\gamma}x^2(V_x^\gamma)^2\right], \quad (T,x)\in G,$$

$$V^\gamma(T,x) = h_\gamma(x), \qquad (T,x)\in\partial G,$$

where $G = (0,\infty)\times(0,\infty)$. Taking the limit as $\gamma\to-\infty$, one obtains

$$\begin{aligned} 0 &= V_T^0 + H(x,V^0,V_x^0), && (T,x)\in G,\\ V^0(T,x) &= h_0(x), && (T,x)\in\partial G, \end{aligned} \tag{45}$$

where

$$H(x,r,p) = -\max_{v\in U}\left[A(v)xp - \frac{B^2(v)}{2}\frac{x^2p^2}{r}\right].$$

This is equation (4.2) in Fleming [9]; see [9] for further details.

Now suppose that $V,W\in C([0,\infty)\times[0,\infty))$ are both subsequential limits of solutions of (44) and thus viscosity solutions of (45). Some assumptions on A and h_γ will be made, and then the results of §3 will be used to show that $V=W$. (Alternatively, one can transform this problem into one of the form given in §4.1 and then apply the methods used there. This approach is discussed more fully in [9]. In either case, one utilizes the structure of the problem to show that the assumptions in §2 are satisfied.)

Assume that there exists $v_0\in U$ such that

$$A(v_0) > 0. \tag{46}$$

For simplicity, let the form of the h_γ be specified by assuming that $h_\gamma(x) = xk_\gamma(x)$, where

$$k_\gamma(1)\to c_1 > 0 \qquad \text{as } \gamma\to-\infty, \tag{47a}$$

and that there exist $\eta<0$ such that, for all $\alpha\ge 1$ and $x>0$,

$$\alpha^{-\eta/\gamma}k_\gamma(x) \le k_\gamma(\alpha x) \le \alpha^{\eta/\gamma}k_\gamma(x) \tag{47b}$$

and $\eta/\gamma\downarrow 0$ as $\gamma\to-\infty$. In this way, $h_\gamma(x)\to h_0(x) = kx$ for some $k>0$.

A simple computation using (46) and (47) yields

$$V(T,x), W(T,x) \ge C_1xe^{C_2T} \tag{48}$$

for appropriate $C_1, C_2 > 0$. Also there exists $C_3<\infty$ such that, for all $R<\infty$,

$$K_R(V), K_R(W) \le C_3, \tag{49}$$

where we are now considering (45) for some finite time-horizon $T\le T_0$. (Remark 5 can be used to extend the result for all $T<\infty$.)

One way to obtain (49) is as follows. Consider $Z_\cdot = \log X_\cdot$, where $X_\cdot$ is given by (41). In particular, let $Z_\cdot^1 = \log X_\cdot^1$, where $X_0^1 = x_1 = e^{z_1}$, and $Z_\cdot^2 = \log X_\cdot^2$, where $X_0^2 = x_2 = e^{z_2}$. Then one can see that

$$Z_t^1 - Z_t^2 = z_1 - z_2 \qquad \text{a.s.}$$

Also let

$$U^\gamma(T,z) = \log[V^\gamma(T,e^z)].$$

Then one may easily show that

$$U^\gamma(T,z_1) - U^\gamma(T,z_2) \le \frac{\gamma+\eta}{\gamma}(z_1 - z_2)$$

for all $z_1 \ge z_2$. (Note that V^γ and U^γ are monotonically increasing.) After a few calculations, this yields

$$V^\gamma(T,x_1+\delta) - V^\gamma(T,x_1) \le C_4 e^{C_5 T}\left[x_1^{-\eta/\gamma} + x_1^{\eta/\gamma}\right]\delta,$$

which by (47b) yields (49). (Note that the above argument can be used to obtain equicontinuity of the V^γ as well.)

To prove that $V = W$, (48) and (49) will be used to show that the Hamiltonian in (45) satisfies assumptions (A) and (B). Since assumption (A) is clearly satisfied, we need only show that (B) is satisfied.

Note that

$$H(x,r,p) = \min_{v\in U} F(x,r,p,v),$$

where

$$F(x,r,p,v) = -A(v)xp + \frac{B^2(v)}{2}\frac{x^2p^2}{r}.$$

Since U is compact and A and B are continuous, we need only prove that F satisfies (B).

Let $R < \infty$, $x \le R$, $T \le T_0$, and suppose $W(T,x) > V(T,x)$. Then

$$\begin{aligned} &F(x,V(T,x),p,v) - F(x,W(T,x),q,v)\\ &\quad= F(x,V,p,v) - F(x,W,p,v) + F(x,W,p,v) - F(x,W,q,v)\\ &\quad= \frac{B^2(v)}{2}x^2p^2\left(\frac{1}{V} - \frac{1}{W}\right) + A(v)x(q-p) + \frac{B^2(v)}{2}\frac{x^2}{W(T,x)}(p^2-q^2), \end{aligned} \tag{50}$$

where we drop the arguments of V and W where unnecessary.

Now, by the mean value theorem, there exists $\xi \in [V(T,x),W(T,x)]$ such that

$$\frac{B^2(v)}{2}x^2p^2\left(\frac{1}{V} - \frac{1}{W}\right) = \frac{B^2(v)}{2}x^2p^2\frac{1}{\xi^2}(W-V),$$

which by (48)

$$\le \frac{B^2(v)}{2}\frac{p^2}{C_1^2e^{2C_2T}}(W-V),$$

which for $|p| \le \max\{K_R(V),K_R(W)\}$ and using (49)

$$\le C_4(W-V) \tag{51}$$

for appropriate C_4.

Note also that

$$A(v)x(q-p) + \frac{B^2(v)}{2}\frac{x^2}{W(T,x)}(p^2-q^2) \le |A(v)|R|p-q| + \frac{B^2(v)}{2}\frac{x^2}{W(T,x)}|p+q||p-q|,$$

which by (48)

$$\le |A(v)|R|p-q| + \frac{B^2(v)}{2}\frac{R}{C_1 e^{C_2 T}}|p+q||p-q|,$$

which by (49)

$$\le C_5 R|p-q| \tag{52}$$

for appropriate C_5.

Combining (50), (51), and (52), we see

$$F(x,V(T,x),p,v) - F(x,W(T,x),q,v) \le C_4[W(T,x)-V(T,x)] + C_5R|p-q|.$$

Therefore, (B) is satisfied, and consequently, $V = W$. Of course, in this simple example with $h_0(x) = x$, it is trivial to show then that one must have

$$V = W = xe^{CT},$$

where

$$C = \max_{v\in U}\left[A(v) - \frac{B^2(v)}{2}\right].$$

REFERENCES

[1] J. A. Ball, J. W. Helton, and M. L. Walker, *H^∞ control for nonlinear systems with output feedback*, IEEE Trans. Automat. Control, 38 (1993), pp. 117–164.

[2] G. Barles and B. Perthame, *Exit time problems in optimal control and vanishing viscosity solutions of Hamilton–Jacobi equations*, SIAM J. Control Optim., 26 (1988), pp. 1133–1148.

[3] E. N. Barron and R. Jensen, *Total risk aversion, stochastic optimal control and differential games*, Appl. Math. Optim., 19 (1989), pp. 313–327.

[4] M. G. Crandall, H. Ishii, and P. L. Lions, *Uniqueness of viscosity solutions of Hamilton-Jacobi equations revisited*, J. Math. Soc. Japan, 39 (1987), pp. 581–595.

[5] M. G. Crandall and P. L. Lions, *Viscosity solutions of Hamilton–Jacobi Equations*, Trans. Amer. Math. Soc., 277 (1983), pp. 1–42.

[6] ———, *On existence and uniqueness of solutions of Hamilton–Jacobi equations*, Nonlinear Anal., 10 (1986), pp. 353–370.

[7] R. J. Elliott and N. J. Kalton, *The existence of value in differential games*, Mem. Amer. Soc., 126 (1972).

[8] L. C. Evans and P. E. Souganidis, *Differential games and representation formulas for solutions of Hamilton–Jacobi–Isaacs equations*, Indiana Univ. Math. J., 33 (1984), pp. 773–797.

[9] W. H. Fleming, *Optimal investment models and risk sensitive stochastic control*, in Mathematical Finance, IMS Vol. Math. Appl. 65, Springer-Verlag, New York, 1995, pp. 75–88.

[10] W. H. Fleming and W. M. McEneaney, *Risk sensitive optimal control and differential games*, in Springer Lecture Notes in Control and Information Science 184, Springer-Verlag, New York, 1992, pp. 185–197.

[11] W. H. Fleming and W. M. McEneaney, *Risk sensitive control with ergodic cost criteria*, in Proc. 31st IEEE Conference on Decision and Control, 1992.

[12] ———, *Risk sensitive control on an infinite time horizon*, SIAM J. Control Optim., 33 (1995), to appear.

[13] W. H. Fleming and H. M. Soner, *Controlled Markov Processes and Viscosity Solutions*, Springer-Verlag, New York, 1992.

[14] H. Ishii, *Uniqueness of unbounded viscosity solution of Hamilton-Jacobi equations*, Indiana Univ. Math. J., 33 (1984), pp. 721–748.

[15] ———, *Representation of solutions of Hamilton-Jacobi Equations*, Nonlinear Anal., 12 (1988), pp. 121–146.

[16] A. Isidori, *H^∞control via measurement feedback for affine nonlinear systems*, in Proc. 31st IEEE Conference on Decision and Control, Dec. 1992.

[17] M. R. James, *A partial differential inequality for dissipative nonlinear systems*, Systems Control Lett., 21 (1993), pp. 315–320.

[18] ———, *Asymptotic analysis of nonlinear stochastic risk-sensitive control and differential games*, Math. of Control Signals Systems, 5 (1992), pp. 401–417.

[19] W. M. McEneaney, *Connections between risk-sensitive stochastic control, differential games and H^∞control: The nonlinear case*, Ph.D. thesis, Brown University, 1993.

[20] A. J. van der Schaft, *Nonlinear state space H^∞control theory*, in Perspectives in Control, Progress in Systems and Control, Birkäuser, Boston, 1993.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 33, No. 5, pp. 1577 1586, September 1995

© 1995 Society for Industrial and Applied Mathematics
013

CONTROL TIME FOR GRAVITY-CAPILLARY WAVES ON WATER*

RUSSELL M. REID†

Abstract. This paper considers distributed open-loop control of small-amplitude linear waves on a fluid in which both surface tension and gravity are significant. It formulates the control system as a first-order evolution equation, reducing the null controllability problem to a moment problem involving frequency exponentials. For simple geometries in which eigenvalues can be calculated explicitly, controllability in arbitrary finite time is established, and a relationship between mode frequencies and controller norm is noted.

Key words. water waves, distributed control, moment problem

AMS subject classifications. 93B05, 93B60, 93C15

1. Introduction.

In 1967 D. L. Russell [11] showed that a vibrating string with distributed control could be steered to rest in a time given by

$$2\int_0^1 \sqrt{\frac{\rho(x)}{p(x)}}\, dx. \tag{1}$$

Open-loop control time for the string thus depends only on physical parameters (mass density, modulus of elasticity, and length) and not on the initial state or controller. Russell's result is sharp in that he shows that the string cannot be controlled in shorter time.

For a uniform string, (1) is the period of the fundamental mode, but mass density and elasticity can be chosen to alter the period of any mode but leave the asymptotic behavior of eigenvalues and control time unchanged, showing that control time is in general not the period of any mode. One might also view (1) as a signal time, the round-trip transit time for a wave on a string of length 1, because the integrand is the reciprocal of phase velocity at any point x. The string is nondispersive: all wavelengths propagate at the same speed.

Studies [9], [7] of gravity waves on fluids showed infinite control time, perhaps because wave velocity approaches zero as wavelength decreases, so that transit time for high-frequency modes is arbitrarily large. However, a general relationship between wave propagation speed and control time has not been established.

The primary aim of this paper is to investigate control time in a simple system with dispersive traveling waves, making note of the effect of wave velocity on control time and controller norm. Gravity-capillary waves are a physical system in which wave velocity varies with wavelength (as opposed to the vibrating string) and is bounded away from zero (as opposed to pure gravity waves). Mode shapes for simple geometries are identical to those for the vibrating string, but mode frequencies are different. If wave propagation velocity determines control time, one would expect finite control time for gravity-capillary waves, and this paper proves controllability in arbitrary finite time.

Using the same approach as [9] and [7] and some results found there, this paper writes the equations for gravity-capillary waves in first-order form in an energy space

* Received by the editors February 19, 1992; accepted for publication (in revised form) May 18, 1994.

† Department of Mathematical Sciences, Michigan Technological University, Houghton, Michigan 49931.

and argues that the evolution operator has a complete orthonormal set of eigenfunctions. It is then possible to study a distributed control system by reducing it to a moment problem along the classical lines of [1] and [11], considering state spaces of finite energy with L_2 controls. The wavemaker problem is complex, especially at the intersection of the controlling wall with the free surface (see, for example, recent work by Joo, Schultz, and Messiter [3]), where an infinite-dimensional controller has in effect a continuous range of Froude numbers, which can be arbitrarily small. This work adopts a small-amplitude linearization and simplifying assumptions on the geometry and the behavior of solutions at corners of the domain, in pursuit of a tractable model exhibiting the wave behavior of gravity-capillary waves. It is noted that boundary control applied to a section of the containing wall is equivalent to a distributed control at the surface, albeit one whose control distribution coefficients are not easy to estimate. For simple geometry, the eigenvalues of a gravity-capillary fluid system are calculated explicitly.

For the gravity-capillary system, the spectrum of the uncontrolled evolution operator has neither finite density nor asymptotic gap, and in studying the moment problem one cannot appeal to the celebrated results of Paley and Wiener [6], Levinson [4], and others.

This work shows controllability by applying the result [8] showing that moment problems are solvable on an arbitrarily short interval when eigenvalue spacing becomes uniformly large, as is the case for fluids in simple geometries.

The essential difference between distributed control of a fluid and a string lies in the frequencies of eigenmodes; this work shows how those frequencies influence the controller norm.

2. Governing equations.

Consider a two-dimensional region Ω, with fixed boundary Γ (assumed smooth) and no beaches, i.e., with nonzero vertical walls at each end (see Fig. 1).

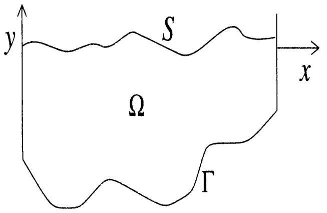


FIG. 1. *Domain and boundary.*

Let the free surface be denoted by S, in a coordinate system in which the undisturbed surface is at $y = 0$ and the vertical segments of the containing walls are at $x = 0$ and $x = \pi$.

Suppose that the fluid is irrotational and incompressible and has uniform density and no viscosity. Then the fluid velocity $u(x, y, t)$ can be expressed in terms of a harmonic potential by $u = -\nabla\Phi$, where Φ has zero normal derivative on the fixed boundary Γ. A small-amplitude linearization of the equations for surface waves, with the surface contour written as $y = \zeta(x, t)$ and $\Phi_y(x, y, t)$ and $\Phi_t(x, y, t)$ evaluated at $y = 0$, takes the form

$$\zeta_t = -\Phi_y, \tag{2}$$

$$\Phi_t = \zeta - T\zeta_{xx}, \tag{3}$$

where T is the surface tension and gravitational and density units have been chosen so that the gravitational constant is 1. Conservation of volume implies

$$\int_0^{\pi} \zeta(x)dx = 0. \tag{4}$$

Equations (2) and (3) can be put in first-order form by letting $\eta = \zeta_t$ and constructing the harmonic function Φ_t whose y-derivative at the surface gives ζ_{tt}. More specifically,

$$\frac{\partial}{\partial t}\begin{bmatrix} \zeta \\ \eta \end{bmatrix} = \begin{bmatrix} 0 & 1 \\ -(A+F) & 0 \end{bmatrix}\begin{bmatrix} \zeta \\ \eta \end{bmatrix}, \tag{5}$$

where A is the evolution operator for pure gravity waves and $F = AL$, with L the Sturm–Liouville operator $Lu = -Tu''$. It is desirable to embed (5) in an energy space where it is skew-symmetric with compact normal resolvent; to do so it is necessary to suppose energy-conserving boundary conditions on L, which we take to be $\zeta'(0) = \zeta'(\pi) = 0$, corresponding physically to an assumption of no friction with the containing walls. Let $L_0^2[0,\pi]$ denote those square integrable functions satisfying equation (4), with the usual L^2 inner product, and let Ψ_ζ be a function harmonic in Ω, with zero normal derivative on the fixed boundary Γ and value $\zeta(x)$ on the free surface. (Ψ_ζ would be Φ_t in (3) if T were 0.) It is shown in [9] and [7] that A defined by

$$A\zeta = \left.\frac{\partial \Psi_\zeta}{\partial y}\right|_{y=0} \tag{6}$$

is the evolution operator for pure gravity waves and is a positive self-adjoint operator with compact resolvent on $L_0^2[0,\pi]$, with domain $\mathcal{D}(A) = H^1 \cap L_0^2$.

The operator $F = AL$ can be seen to be positive and self-adjoint with compact resolvent on $L_0^2[0,\pi]$, with domain $\mathcal{D}(A) = H^3 \cap L_0^2 \cap BC$, where BC represents the Sturm–Liouville boundary conditions $\zeta'(0) = \zeta'(\pi) = 0$. When brackets are used to represent the L^2 inner product, note that for u and v in $\mathcal{D}(F)$,

$$\langle ALu, v\rangle = \langle Lu, Av\rangle = \langle u, LAv\rangle\,. \tag{7}$$

The first equality uses self-adjointness of A and the fact that $\mathcal{D}(A)$ is contained in $\mathcal{D}(F)$, whereas the second uses self-adjointness of L and requires that Av be in $\mathcal{D}(L)$, which comes from noticing that

$$Av = \left.\frac{\partial \Psi_v}{\partial y}\right|_{y=0} \tag{8}$$

so that

$$\frac{d}{dx}(Av) = \left.\frac{\partial^2 \Psi_v}{\partial y \partial x}\right|_{y=0}, \tag{9}$$

which is zero at $x = 0$ and $x = \pi$ because Ψ_v has zero normal (x) derivative there. (This is really a continuity assumption at the corners, the "no beaches" assumption.) One can conclude self-adjointness of F from the further observation that A^{-1} and

L^{-1} are compact operators on $L_0^2[0,\pi]$ (see, for example, [10, p. 353]). One can then embed (5) in a Hilbert space H, defining an inner product

$$\left\langle \begin{bmatrix} \zeta \\ \eta \end{bmatrix}, \begin{bmatrix} z \\ h \end{bmatrix} \right\rangle_H = \langle \zeta, (A+F)z \rangle + \langle \eta, h \rangle . \tag{10}$$

The space H of functions of finite norm is a Hilbert space, and the operator A_0 defined by

$$A_0 = \begin{bmatrix} 0 & 1 \\ -(A+F) & 0 \end{bmatrix}, \tag{11}$$

with domain

$$\mathcal{D}(A_0) = \left\{ \begin{bmatrix} \zeta \\ \eta \end{bmatrix} : \zeta \in H^3 \cap L_0^2 \cap BC \,, \eta \in H^{3/2} \cap L_0^2 \right\}, \tag{12}$$

is skew-symmetric with compact resolvent in this space.

Although the Hilbert space above is convenient for calculation, its norm does not give the actual physical energy; rather that is given by

$$\left\langle \begin{bmatrix} \zeta \\ \eta \end{bmatrix}, \begin{bmatrix} z \\ h \end{bmatrix} \right\rangle_E = \langle \zeta, z \rangle + \langle \eta, A^{-1}h \rangle + \langle \zeta, Lz \rangle , \tag{13}$$

where the first term can be shown to give gravitational potential energy, the second, kinetic energy, and the third, elastic potential energy. It turns out that eigenfunctions are orthogonal in the actual energy inner product also, but for now we use H.

One may summarize these observations in a theorem.

THEOREM 2.1. *The system* (2), (3) *describing gravity-capillary waves on a fluid surface can be written in the form*

$$X_t = A_0 X, \tag{14}$$

where X is in a Hilbert space H and A_0 is skew-symmetric with compact resolvent on H. There is an orthonormal basis for H consisting of eigenfunctions of A_0. The eigenvalues of A_0 are $\omega_k = i\sqrt{\lambda_k}$, $k = 1, 2, 3, \ldots$, with $\omega_{-k} = -\omega_k$, where λ_k denotes the kth eigenvalue of $A + F$. Solutions of (14) *form a group of bounded operators on H, denoted by $X(t) = e^{A_0 t} X_0$.*

3. Moment problem for a null control. Consider a distributed open-loop control applied at the surface, with spatial distribution of control fixed, which can be expressed in the form

$$X_t = A_0 X + Bu, \tag{15}$$

where B is in H and u is the control. An initial state X_0 can be steered to zero in time L (we used T for surface tension), provided that

$$\left\| e^{A_0 L} X_0 + \int_0^L e^{A_0(L-s)} Bu(s)\, ds \right\|_H = 0. \tag{16}$$

In terms of the basis of H consisting of eigenfunctions ϕ_n of A_0, denoting the corresponding eigenvalue of A_0 by ω_n, this becomes

$$\left\| \sum_{-\infty}^{\infty} \left(e^{\omega_n L} a_n \phi_n + \int_0^L e^{\omega_n (L-s)} b_n \phi_n u(s)\, ds \right) \right\|_H = 0, \tag{17}$$

where $X_0 = \sum a_n\phi_n$ and $B = \sum b_n\phi_n$. A necessary and sufficient condition for this to be true is, provided that no b_n is zero,

$$\int_0^L e^{-s\omega_n} u(s)ds = \frac{-a_n}{b_n} = c_n, \tag{18}$$

a standard moment problem for null control of an evolution equation.

One drawback to a moment approach to distributed control is the difficulty in characterizing the controllability space in terms of the coefficients b_n, which themselves may be difficult to calculate. Capillary waves differ little from other distributed linear systems in this regard, however; this work seeks primarily to consider the effect of mode frequencies on the controller norm for a given set of ratios c_n. Before doing so, we note that control applied to a boundary wall can take the form of a distributed control at the surface.

Consider control applied by small-amplitude vibration of the boundary wall at $x = 0$. The argument that such a control takes the form (14) is identical to that for gravity waves in [9], because the influence of the boundary control separates cleanly from the surface dynamics. To briefly review that calculation, let C denote some part of the wall at $x = 0$, controlled so that the harmonic function Φ_t in (3) has normal derivative on C given by

$$\frac{\partial \Phi_t}{\partial n} = h(y)u(t) \tag{19}$$

and normal derivative zero on the rest of Γ. It is required that $h(y) \in H^1$ and that its integral along C be zero to ensure conservation of volume. Let Ψ denote the harmonic function Φ_t that corresponds to a given surface configuration in an uncontrolled system, and let Θ be a harmonic function which is zero on the surface S and has normal derivative $h(y)u(t)$ on C and zero normal derivative elsewhere on Γ. Then it is easy to see that $\Psi + u(t)\Theta$ satisfies the boundary conditions required of Φ_t in the controlled system. It follows that

$$\zeta_{tt} = -\frac{\partial}{\partial y}(\Psi + u(t)\Theta)\bigg|_{y=0} = -(A+F)\zeta - u(t)\frac{\partial \Theta}{\partial y}\bigg|_{y=0} \tag{20}$$

so that (15) holds with

$$B = \begin{bmatrix} 0 \\ g(x) \end{bmatrix} = \begin{bmatrix} 0 \\ -\left.\frac{\partial\Theta}{\partial y}\right|_{y=0} \end{bmatrix}. \tag{21}$$

The control coefficients b_n that result from controlling part of a containing wall instead of applying control directly to the surface can be complicated for even the simplest controllers, however, as the following example, chosen because it can be computed exactly, shows. Suppose Ω is an infinite-depth domain, with controlled wall C the entire vertical wall at $x = 0$, and control function $h(y) = \sin(y)$ in equation (19). The y-derivative of the harmonic function Θ which defines the control element B in equation (21) can be constructed as an integral by using ideas from electrostatic potential theory. Specifying normal derivative on a surface is equivalent to specifying surface charge density on the wall at $x = 0$, and symmetry can be used to ensure zero normal derivative (electric field) at $x = \pi$ by placing the same charge density at $x = 2\pi$. Reflecting both charge distributions above the x-axis with the opposite sign ensures that the value of Θ (the potential) at $y = 0$ be zero. Then $g(x)$ in equation (21), temporarily thought of as the vertical component of an electric field at $y = 0$, can be constructed as an integral (recalling that in two dimensions electric field falls off as $1/r$). The y-component of electric field at location x due to charge $\sin(y)$ at depth y (doubled because of an antisymmetric charge above the x-axis), integrated over y to give the vertical component of electric field at the surface due to the charges at $x = 0$, gives

$$2\int_0^\infty \frac{\sin(y)}{\sqrt{x^2+y^2}}\frac{y}{\sqrt{x^2+y^2}}\,dy. \tag{22}$$

This integral can be calculated explicitly, giving πe^{-x}, and then added to the result for the charges at $x = 2\pi$, which is $\pi e^{-(2\pi - x)}$, to give $g(x) = \pi e^{-x} + \pi e^{-(2\pi-x)}$. The energy inner product of the control element B defined by this $g(x)$ with the nth eigenfunction of ϕ_n of A_0 (ϕ_n not normalized, because c_n is a ratio) can be calculated exactly, giving

$$b_n = \frac{\pi\sinh(\pi) + \pi\sqrt{Tn^3+n}\,\sin(\pi\sqrt{Tn^3+n})}{e^\pi(Tn^3+n+1)}. \tag{23}$$

It is not obvious whether any integer value of n gives exactly zero for b_n, but certainly one gets arbitrarily close to zero, and a plot shows a complicated dependence on n.

4. Eigenvalues for simple geometries. It is well known that existence of a solution $u(t)$ to a moment problem such as (18) depends on the interval L, the eigenvalues $\omega_n = \pm i\sqrt{\lambda_n}$, and the sequence $\{c_n\}$. This work now restricts its attention to a geometry in which the eigenvalues of $A + F$ can be computed explicitly.

Suppose that the region Ω has a flat bottom at $y = -H$. Then any $\zeta(x)$ in $L_0^2[0, \pi]$ satisfying $\zeta'(0) = \zeta'(\pi) = 0$ must be

$$\zeta = \sum_{n=1}^{\infty} a_n \cos(nx). \tag{24}$$

The harmonic function Ψ_ζ meeting boundary conditions and matching ζ at $y = 0$ is

$$\Psi(x,y) = \sum_{n=1}^{\infty} na_n \cos(nx)\frac{\cosh n(y+H)}{\cosh(nH)}, \tag{25}$$

giving

$$A\zeta = \sum_{n=1}^{\infty} na_n \cos(nx)\tanh(nH) \tag{26}$$

and

$$F\zeta = A(-T\zeta_{xx}) = \sum_{n=1}^{\infty} Tn^3 a_n \cos(nx)\tanh(nH). \tag{27}$$

The eigenfunctions of A and F are the same so that

$$(A+F)\cos(nx) = (Tn^3+n)\tanh(nH)\cos(nx). \tag{28}$$

Thus the eigenvalues of $A+F$ in this geometry are $\lambda_n = (Tn^3+n)\tanh(nH)$, so that the eigenvalues of A_0 are $\omega_n = i\sqrt{(Tn^3+n)\tanh(nH)}$, with $\omega_{-n} = -\omega_n$. A similar calculation for an infinite-depth domain leads to $\omega_n = i\sqrt{(Tn^3+n)}$. In particular, eigenvalues ω_n are $\mathcal{O}(n^{3/2})$, so that adjacent differences become arbitrarily large.

It is helpful to compare the infinite-depth example to the simplest vibrating string of length π, taking the string's density and elasticity to be 1. If both are put in skew-adjoint first-order form, the eigenvalues of the first-order evolution operator A_0 are $\omega_n = \pm in$ for the string and $\omega_n = \pm i\sqrt{Tn^3+n}$ for the fluid.

Allowing the ends of the string to slide freely rather than be fixed (to make the comparison to the fluid more obvious) and normalizing in the energy norm, eigenfunctions for the string are

$$\phi_n = \begin{bmatrix} n^{-1}\cos(nx) \\ i\cos(nx) \end{bmatrix}. \tag{29}$$

Similarly, eigenfunctions for the fluid, normalized in the norm of the Hilbert space H, are

$$\phi_n = \begin{bmatrix} (Tn^3+n)^{-1/2}\cos(nx) \\ i\cos(nx) \end{bmatrix}. \tag{30}$$

The same eigenfunctions are orthogonal in the actual energy inner product (13); normalized in that norm they are

$$\phi_n = \begin{bmatrix} (Tn^2+1)^{-1/2}\cos(nx) \\ i\sqrt{n}\cos(nx) \end{bmatrix}. \tag{31}$$

If control is applied as a distributed force, B has the first-order form

$$B = \begin{bmatrix} 0 \\ f(x) \end{bmatrix}. \tag{32}$$

Suppose one makes the assumption, used by Russell for the vibrating string [11, (2.17)], that the ordinary Fourier coefficients γ_n of $f(x)$ decay slowly enough that

$$\liminf_{n\to\infty} n|\gamma_n| > 0. \tag{33}$$

A simple $f(x)$ for which this is true is the piecewise linear function

$$f(x) = \begin{cases} x, & 0 \le x \le \pi/2, \\ (\pi/2 - x), & \pi/2 < x \le \pi. \end{cases}$$

Then for the string, and also for the fluid using the H-norm,

$$|b_n| = |\gamma_n|, \tag{34}$$

so that, under the assumption (33), $a_n/b_n = \mathcal{O}(na_n)$, which is square summable if the initial state is in $\mathcal{D}(A_0)$.

For the fluid, if we use the actual energy norm,

$$|b_n|\sqrt{n} = |\gamma_n|, \tag{35}$$

so that for large n, again under the assumption (33), $a_n/b_n = \mathcal{O}(n^{3/2}a_n)$, which is square summable for initial states in $\mathcal{D}(A_0)$.

However, in this work we are primarily interested in the effect of wave frequencies on control time and controller norm; henceforth we consider only the sequence of ratios c_n under the supposition that it is square summable.

5. Controllability and solutions to the moment problem.

A moment problem such as (18) is

$$(36) \qquad \langle f_n, u \rangle = c_n \, .$$

If the sequence $\{f_n\}$ has the property that the equations (36) have a solution $u \in H$ for every ℓ^2 sequence $\{c_n\}$ so that its moment space contains ℓ^2, it is a *Riesz-Fischer* sequence; if its moment space is contained in ℓ^2, it is a *Bessel* sequence. A sequence which is both a Bessel sequence and a Riesz–Fisher sequence is *strongly independent* and a Riesz basis for its closed linear span. The basic criterion is due to Boas: a sequence $\{f_n\}$ is a Riesz–Fischer sequence if there is an m and a Bessel sequence if there is an M such that, for every sequence of scalars $\{c_n\}$,

$$(37) \qquad m \sum |c_n|^2 \leq \left\| \sum c_n f_n \right\| 2 \leq M \sum |c_n|^2 \, .$$

The sequences $\{f_n\}$ which arise in the control of evolution equations are sequences of complex exponentials $\{e^{\pm i\lambda_n x}\}$; they are Bessel sequences in $L^2[-L, L]$ for every positive L provided only that the λ_n are real and separated [12]. Ingham showed in [2] that a sequence of exponentials $\{e^{\pm i\lambda_n x}\}$ is a Riesz–Fischer sequence in $L^2[-L, L]$, provided that every $|\lambda_{n+1} - \lambda_n|$ exceeds π/L. It can be shown that if a sequence of exponentials is incomplete and satisfies Ingham's condition for all but finitely many n, then it is a Riesz–Fischer sequence. (See, for example, [8] where it is shown that strong independence is unaffected by replacing finitely many λ_n, allowing rearrangement until Ingham's condition is met for all n.) For capillary waves the eigenvalues $|\lambda_n|$ are $\mathcal{O}(n^{3/2})$, so that the sequence $\{f_n\} = \{e^{\pm i\lambda_n x}\}$ is strongly independent in $L^2[-L, L]$ for every positive L.

If a solution u to (36) exists, there is a unique least-norm solution in the closed linear span of the $\{f_n\}$. If the sequence $\{f_n\}$ is strongly independent, every element in its closed linear span can be expressed as a series ([5, p. 317]). If we let $u = \sum a_n f_n$ and denote the sequences of coefficients (finite or infinite) by $a = \{a_1, a_2, a_3, \ldots\}$, $c = \{c_1, c_2, c_3, \ldots\}$, then the requirement (36) is

$$(38) \qquad \left\langle f_n, \sum a_j f_j \right\rangle = \sum a_j \langle f_n, f_j \rangle = c_n \, .$$

If G is the Gram matrix whose entries are $g_{nj} = \langle f_n, f_j \rangle$, this is simply

$$(39) \qquad Ga = c.$$

For the control moment problem (18), the entries of G are inner products of frequency exponentials independent of the initial state or controller; in $L_2[-L, L]$, a simple calculation gives

$$g_{ij} = \frac{\sin L|\omega_i - \omega_j|}{L\,|\omega_i - \omega_j|} \, .$$

The coefficients of the least-norm controller are determined by solving the system $Ga = c$, and the norm of the controller itself is

$$(40) \qquad \|u\|^2 = \left\langle \sum a_n f_n, \sum a_n f_n \right\rangle = a^T G a.$$

The moment problem has a solution for every $c \in \ell^2$ when G is invertible; the Gram matrix of a finite set of modes is always invertible. In either case, since G is symmetric,

$$\|u\|^2 = a^T G a = a^T c = \left(G^{-1}c\right)^T c = c^T G^{-1} c, \tag{41}$$

so that the relationship between controller norm and initial state depends fundamentally on the eigenstructure of the Gram matrix of inner products of mode frequencies.

If G_n denotes the Gram matrix of the first n modes (or rather, to preserve complex conjugates $\pm n$), then D. L. Russell's result can be viewed as showing that, for fixed $L \geq 2\pi$,

$$\lim_{n\to\infty} \left\|G_n^{-1}(L)\right\| < \infty,$$

whereas for any $L < 2\pi$,

$$\lim_{n\to\infty} \left\|G_n^{-1}(L)\right\| = \infty.$$

For capillary waves, for fixed $L > 0$,

$$\lim_{n\to\infty} \left\|G_n^{-1}(L)\right\| = K(L) < \infty,$$

because a Riesz–Fischer sequence is one for which eigenvalues of finite subsections of the Gram matrix are bounded away from zero; equivalently the ℓ^2 operator G is boundedly invertible with norm $\|G^{-1}\| \leq 1/m$ in equation (37).

The remarks above allow a final theorem.

THEOREM 5.1. *In terms of the orthonormal basis for H provided by the eigenfunctions ϕ_n of A_0, let $B = \sum b_n\phi_n$. Let $X_0 = \sum a_n\phi_n$ be an initial state in $\mathcal{D}(A_0)$ for which the sequence of ratios $c_n = a_n/b_n$ is square summable*; *in particular this is true of any initial state in $\mathcal{D}(A_0)$ under the assumption* $\inf|n^{3/2}b_n| > 0$, *or of any finite sequence $\{a_n\}$, provided only that no b_n is zero. Then for any time $L > 0$ there is a control u which steers X_0 to zero in time L and a constant $K(L)$ such that* $\|u\|_{L^2[0,L]} < K(L)\,\|\{c_n\}\|_{\ell^2}$.

It is worth remarking that for a capillary wave system, $K(L)$ is $\mathcal{O}(1)$ as long as L exceeds the natural control time $2\pi/(\omega_{n+1} - \omega_n)$ for every mode pair; it grows extremely rapidly with shorter control times.

Acknowledgments. The author wishes to thank the referees for significant contributions to this work.

REFERENCES

[1] A. G. BUTKOVSKII, *The method of moments in the theory of optimal control of systems with distributed parameters*, Automat. Remote Control, 24 (1963), pp. 1106–1113.

[2] A. E. INGHAM, *Some trigonometrical inequalites with applications to the theory of series*, Math. Z., 41 (1936), pp. 367–379.

[3] S. JOO, W. SCHULTZ, AND A. MESSITER, *An analysis of the inital-value wavemaker problem*, J. Fluid Mech., 214 (1990), pp. 161–183.

[4] N. LEVINSON, *Gap and Density Theorems*, AMS Colloquium Publications, Vol. 26, American Mathematical Society, Providence, RI, 1940.

[5] A. NAYLOR AND G. W. SELL, *Linear Operator Theory in Engineering and Science*, Applied Mathematical Sciences, Vol. 40, Springer-Verlag, Berlin, New York, 1982.

[6] R. E. PALEY AND N. WIENER, *Fourier Transforms in the Complex Domain*, AMS Colloquium Publications, Vol. 19, American Mathematical Society, Providence, RI, 1934.

[7] R. M. Reid, *Open loop control of water waves in an irregular domain*, SIAM J. Control Optim, 24 (1986), pp. 789–796.

[8] ———, *A pythagorean inequality*, Proc. Am. Math. Soc., 123 (1995), pp. 831–839.

[9] R. M. Reid and D. L. Russell, *Boundary control and stability of linear water waves*, SIAM J. Control Optim., 23 (1985), pp. 111–121.

[10] W. Rudin, *Functional Analysis*, McGraw-Hill, New York, 1991.

[11] D. L. Russell, *Nonharmonic Fourier series in the control theory of distributed parameter systems*, J. Math. Anal. Appl., 18 (1967), pp. 542–559.

[12] R. M. Young, *An Introduction to Nonharmonic Fourier Series*, Academic Press, New York, 1980.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 33, No. 5, pp. 1587 1602, September 1995

© 1995 Society for Industrial and Applied Mathematics
014

AN EMBEDDING OF DOMAINS APPROACH IN FREE BOUNDARY PROBLEMS AND OPTIMAL DESIGN*

P. NEITTAANMÄKI[†] AND D. TIBA[‡]

Abstract. Both free boundary problems and optimal design problems involve unknown or variable domains. We describe a direct approach extending the formulation of the problem to a larger fixed domain without modifying the differential operator. This argument is based on an approximate controllability-type property.

Key words. variable domains, controllability, optimal control, fictitious domain

AMS subject classifications. 49D37, 49D05, 49E05

1. Introduction.

For the solution of partial differential equations defined in exterior domains or in domains with a complicated geometry, one procedure consists of extending the given problem to the whole space or to a convenient simple region, where an associated control problem has to be solved; see the paper by C. Atamian and P. Joly [1].

This idea is very useful in problems involving unknown or variable domains, and we mention the works of J. Blum [3], [4] related to free boundary problems in the physics of plasma and the papers of K. H. Hoffmann, J. Haslinger, and M. Kocvara [8]; D. Tiba [12], [13]; and R. Mäkinen, P. Neittaanmäki, and D. Tiba [14] devoted to applications in optimal shape design.

It turns out that a main ingredient in the argument is given by certain geometric controllability properties of elliptic systems, with or without constraints. From this point of view, the literature is not very rich, and we quote a first result from the classic book by Lions [9, p. 85] and the recent works by V. Barbu and D. Tiba [2] and by D. Tiba [12].

In §2 we analyse from this point of view the rapid method of identification of the plasma boundary in a tokamak used by J. Blum [4, Chap. V], and in §3, we study a typical optimal shape design problem (see Pironneau [11, Chap. VI]). We use a boundary control approximation, which is efficient since it involves convex optimization problems even when the original problem is nonconvex. In §4 we describe a distributed control approach in the case of the optimal packaging problem [16]. This is again based on a controllability-type result, and it may be viewed as a general fixed domain method in optimal shape design. The last section is devoted to numerical experiments which investigate the accuracy of the proposed methods.

Let us underline that one of the main aims of this paper is to present a unitary treatment of the various problems mentioned above and studied previously by other methods. We emphasize the use of fixed domains (containing the original ones), which is advantageous from a numerical point of view. Moreover, the partial differential operator describing the system remains unchanged in this approach and we avoid the

* Received by the editors May 26, 1992; accepted for publication (in revised form) May 18, 1994.

[†] Department of Mathematics, University of Jyväskylä, P.O. Box 35, Jyväskylä, Finland.

[‡] Institute of Mathematics of the Romanian Academy, P.O. Box 1-764, RO-70700 Bucharest, Romania. The research of this author was partially performed while visiting l'Institut National de Recherche en Informatique et en Automatique, Paris, under the support of the French Minister of Research and Technology.

control in the coefficients techniques specific to optimal design problems (see Murat and Simon [10]).

2. Free boundary problems in plasma physics. In a tokamak, it is essential to know the shape and the position of the plasma boundary in the void chamber. In cross section, this gives a regular domain $\Omega \subset \mathbb{R}^2$ and the plasma covers a subdomain $D \subset \mathbb{R}^2$, confined by certain limitators F, which may have various shapes. (See Fig. 2.1, which is a schematic representation of the physics of the problem.)

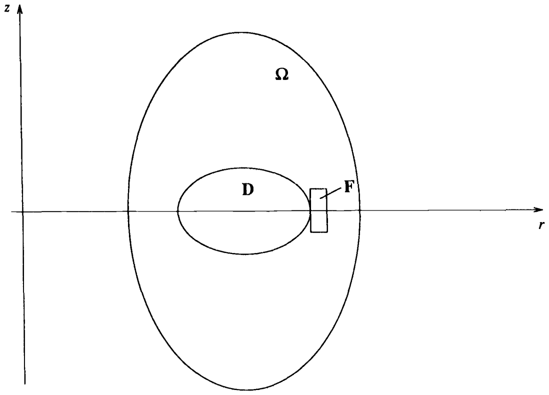


FIG. 2.1.

In void region $\Omega \setminus \overline{D}$ (unknown a priori), the poloidal flux ψ satisfies the linear elliptic equation

$$L\psi = -\frac{\partial}{\partial r}\left(\frac{1}{r}\frac{\partial \psi}{\partial r}\right) - \frac{\partial}{\partial z}\left(\frac{1}{r}\frac{\partial \psi}{\partial z}\right) = 0, \tag{2.1}$$

which is nonsingular since $r > c > 0$ in Ω under the natural choice of axes indicated in Fig. 2.1, which observes the symmetry of the torus representing the void chamber.

On the exterior boundary $\partial\Omega$, it is possible to measure

$$\psi = f \quad \text{on } \partial\Omega, \tag{2.2}$$

$$\frac{1}{r}\frac{\partial \psi}{\partial n} = g \quad \text{on } \partial\Omega, \tag{2.3}$$

thus obtaining a Cauchy problem for the elliptic operator (2.1) in $\Omega \setminus \overline{D}$. Assuming that ψ may be computed, the free boundary ∂D is characterized as a level set by the relation $M \in \partial D$ iff

$$\psi(M) = \sup_{x \in F} \psi(x). \tag{2.4}$$

The idea is to find ∂D using (2.4) and the overdetermined data on $\partial\Omega$. For arbitrary f and g, the obtained geometry of D may be very general, but for real data the method proposed by Blum [4, Chap. V] gives very accurate results, according to his numerical experiments. We follow the same procedure here, but our theoretical justification is different and avoids any extension assumptions, which are equivalent with exact boundary controllability properties and may be not true in general.

Let Γ be a closed regular curve in D, which is not in a neighbourhood of ∂D. (From the point of view of applications it is even preferable that Γ is "far" from ∂D

in order to allow more liberty for the searched position of the free boundary of the plasma.) We denote as Ω_0 the domain between $\partial\Omega$ and Γ, $\Omega_0 \supset \Omega \setminus D$.

We replace the ill-posed Cauchy problem (2.1)–(2.3) by the following noncoercive boundary control problem:

$$\underset{u\in L^2(\Gamma)}{\text{minimize}}\left\{ J(u) = \frac{1}{2}\left|\frac{1}{r}\frac{\partial\psi}{\partial n} - g\right|^2_{L^2(\partial\Omega)}\right\} \tag{2.5}$$

subject to

$$L\psi = 0 \quad \text{in } \Omega_0, \tag{2.6}$$

$$\psi = f \quad \text{on } \partial\Omega, \tag{2.7}$$

$$\psi = u \quad \text{on } \Gamma. \tag{2.8}$$

Here, we notice that if $f \in H^{3/2}(\partial\Omega)$, then $\psi \in H^2(\vartheta)$, where ϑ is a neighbourhood of $\partial\Omega$ in Ω_0 and the cost functional is well defined.

We use the approximate boundary controllability result of Lions [9, p. 85] and we see that

$$\inf_{u\in L^2(\Gamma)} J(u) = 0. \tag{2.9}$$

Therefore, a minimizing sequence of pairs $[u_n, \psi_n]$ will provide a good approximation of the Cauchy data (2.2), (2.3). Due to the absence of coercivity properties in (2.5)–(2.8), we haven't ensured the existence of optimal pairs.

We approximate the problem by the usual Tikhonov regularization technique, and we shall use an argument based on the uniqueness of the Cauchy problem for the adjoint system of the control problem

$$\text{minimize}\left\{ J_\varepsilon(u) = \frac{1}{2}\left|\frac{1}{r}\frac{\partial\psi}{\partial n} - g\right|^2_{L^2(\partial\Omega)} + \frac{\varepsilon}{2}\int_\Gamma u^2\, d\tau\right\} \tag{2.10}$$

subject to (2.6)–(2.8), $\varepsilon > 0$.

Let $[u_\varepsilon, \psi_\varepsilon]$ be the unique optimal pair for the problem (2.10).

THEOREM 2.1. *Assume that $f \in H^{3/2}(\partial\Omega)$, $g \in L^2(\partial\Omega)$. Then, for $\varepsilon \to 0$, we have*

$$\frac{1}{r}\frac{\partial\psi_\varepsilon}{\partial n} \to g \tag{2.11}$$

strongly in $L^2(\partial\Omega)$.

Proof. By taking $u = 0$ in (2.8), we infer that $\{\frac{1}{r}\frac{\partial\psi_\varepsilon}{\partial n}\}$ and $\{\varepsilon^{1/2}u_\varepsilon\}$ are bounded in $L^2(\partial\Omega)$, $L^2(\Gamma)$. We define the adjoint system by

$$Lp_\varepsilon = 0 \quad \text{in } \Omega_0, \tag{2.12}$$

$$p_\varepsilon = 0 \quad \text{on } \Gamma, \tag{2.13}$$

$$p_\varepsilon = \frac{1}{r}\frac{\partial\psi_\varepsilon}{\partial n} - g \quad \text{on } \partial\Omega, \tag{2.14}$$

which has a unique solution $p_\varepsilon \in H^{1/2}(\Omega_0)$, which is regular around Γ.

Standard calculus shows that the maximum principle (Pontryagin) takes the form

$$\frac{1}{r}\frac{\partial p_\varepsilon}{\partial n} = \varepsilon u_\varepsilon \quad \text{on } \Gamma. \tag{2.15}$$

Let $\ell = \lim \left(\frac{1}{r}\frac{\partial \psi_\varepsilon}{\partial n} - g\right)$ in the weak topology of $L^2(\partial\Omega)$ on a subsequence ε_n. The adjoint equation (2.12)–(2.14) defines a bounded linear operator from $L^2(\partial\Omega)$ to $H^{1/2}(\Omega_0)$, so we can pass to the limit $p_\varepsilon \to p$, which satisfies in a weak sense

$$Lp = 0 \quad \text{in } \Omega_0, \tag{2.16}$$

$$p = 0 \quad \text{on } \Gamma, \tag{2.17}$$

$$p = \ell \quad \text{on } \partial\Omega. \tag{2.18}$$

Passing to the limit in (2.15) too, we see that

$$\frac{1}{r}\frac{\partial p}{\partial n} = 0 \quad \text{on } \Gamma. \tag{2.19}$$

Then the uniqueness for the Cauchy problem (2.16), (2.17), (2.19) gives $p = 0$—that is, $\ell = 0$—by (2.18). Therefore, we get (2.11) with weak convergence instead of strong convergence.

Let $\{\tilde{\psi}_{\varepsilon_n}\}$ be a sequence of convex combinations of $\{\psi_{\varepsilon_n}\}$ corresponding to the sequence $\{\tilde{u}_{\varepsilon_n}\}$ of controls given by the Mazur theorem, i.e., $\frac{1}{r}\frac{\partial \tilde{\psi}_{\varepsilon_n}}{\partial n} \to g$ strongly in $L^2(\partial\Omega)$.

Let the fixed control $\tilde{u}_{\varepsilon_n}$ be a convex combination of controls $\{u_{\varepsilon_k}\}_{k=m,\dots,n}$. Then $\{\varepsilon_n^{1/2}\tilde{u}_{\varepsilon_n}\}$ is bounded in $L^2(\Gamma)$. We put the control $\tilde{u}_{\varepsilon_n}$ in the problem corresponding to ε_n^2 and we obtain

$$\begin{aligned}&\frac{1}{2}\left|\frac{1}{r}\frac{\partial \psi_{\varepsilon_n^2}}{\partial n} - g\right|^2_{L^2(\partial\Omega)} + \frac{\varepsilon_n^2}{2}\int_\Gamma u^2_{\varepsilon_n^2}\,d\tau \\ &\le \frac{1}{2}\left|\frac{1}{r}\frac{\partial \tilde{\psi}_{\varepsilon_n}}{\partial n} - g\right|^2_{L^2(\partial\Omega)} + \frac{\varepsilon_n^2}{2}\int_\Gamma \tilde{u}^2_{\varepsilon_n}\,d\tau.\end{aligned} \tag{2.20}$$

By the above argument, we notice that the right-hand side in (2.20) converges to 0; therefore $\frac{1}{r}\frac{\partial \psi_{\varepsilon_n^2}}{\partial n} - g \to 0$ strongly in $L^2(\partial\Omega)$. Finally, (2.11) is valid without taking subsequences. □

3. Boundary control and optimal shape design. We analyse the following standard design problem (see Pironneau [11, Chaps. III and VI]):

$$\underset{D}{\text{minimize}}\left\{V(D) = \int_E |\nabla y - y_d|^2\,dx\right\} \tag{P}$$

subject to

$$-\Delta y = f \quad \text{in } D, \tag{3.1}$$

$$y = 0 \quad \text{on } \partial D. \tag{3.2}$$

Here $y_d \in L^2(\Omega)^N$, $f \in L^2(\Omega)$, $E \subset D \subset \Omega \subset \mathbb{R}^2$ (compare with Fig. 2.1) are bounded domains, $\overline{E} \subset \Omega$ are fixed, D is variable (the minimization parameter), and y is a weak solution of the Dirichlet problem (3.1), (3.2) in D.

In the absence of regularity assumptions on ∂D, we haven't ensured the existence of an optimal subdomain for (P), and we shall study the minimizing sequences.

We associate with (P) the following constrained control problem:

$$\text{(Q)} \qquad \underset{u}{\text{minimize}} \left\{ W(u) = \int_E |\nabla y - y_d|^2 \, dx \right\},$$

$$\text{(3.3)} \qquad -\Delta y = f \quad \text{in } \Omega,$$

$$\text{(3.4)} \qquad y = u \quad \text{on } \partial\Omega,$$

$$\text{(3.5)} \qquad u \leq 0 \quad \text{on } \partial\Omega,$$

$$\text{(3.6)} \qquad y \geq 0 \quad \text{in } \overline{E}.$$

We assume that $\partial\Omega$ is regular, $u \in H^{3/2}(\partial\Omega)$, so $y \in H^2(\Omega)$ and it is continuous by the Sobolev embedding theorem. Then (3.5) and (3.6) have a clear meaning. We show that there is a strong relationship between the design problem (P) and the control problem (Q) defined in a fixed domain. Any admissible control for (Q) generates uniquely a subdomain of Ω with the same associated cost, and we may write the following theorem.

THEOREM 3.1. *The problem* (Q) *is a subproblem of* (P) *if* $f \geq 0$.

Proof. For every $u \in H^{3/2}(\partial\Omega)$ admissible for (Q), we define

$$\widetilde{D}_u = \text{int}\{x \in \Omega, \ \ y_u(x) \geq 0\},$$

which is an open set such that $E \subset \widetilde{D}_u \subset \Omega$, by (3.6). We have denoted by y_u the solution of (3.3), (3.4) corresponding to u. We define $D_u \subset \Omega$ to be the connected component of $\widetilde{D}_u$ which contains E, and we remark that $y_u|_{D_u}$ is the solution of the Dirichlet problem (3.1), (3.2) in D_u.

Therefore, by this correspondence $u \rightarrow D_u$, we associate, with each control $u \in H^{3/2}(\partial\Omega)$ and satisfying (3.5), (3.6), a domain $D_u \subset \Omega$ which gives the same cost $V(D_u) = W(u)$. No smoothness is valid in general for ∂D_u, but the existence of the weak solution of the Dirichlet problem in D_u is ensured. Since this is given by $y_u|_{D_u}$, it is in fact a strong solution. This correspondence is not void since we may take $u \equiv 0$ and $D_u = \Omega$ as an example by the weak maximum principle. It is also injective since otherwise, if u_1 and u_2 produce the same subdomain $\widetilde{D}$ by the above construction, then $y_1 - y_2 \equiv 0$ in $\widetilde{D}$ and, by analyticity, $y_1 - y_2 \equiv 0$ in Ω; that is, $u_1 \equiv u_2$. ▯

For the converse statement, we shall show that it is possible to construct a minimizing sequence for (P) with domains of the type $\{D_{u_k}\}$, where $\{u_k\}$ is an admissible sequence of controls for (Q), generally unbounded. Then, by Theorem 3.1, $\{u_k\}$ is a minimizing sequence for (Q) too, and we have

$$\text{(3.7)} \qquad \inf (\text{P}) = \inf (\text{Q}).$$

Finally, every minimizing sequence $\{w_n\}$ for (Q) will produce a minimizing sequence $\{D_{w_n}\}$ for (P) by (3.7), and this shows that the solution of the design problem (P) is reduced to the solution of the control problem (Q).

Remark. The advantage of this approach is that the control problem (Q) is defined in a fixed domain, which avoids the redefinition of new finite element mesh in each iteration of the algorithm, as in the boundary variation technique. Moreover, the problem (Q) is convex; therefore, the global minimum for (P) may be obtained, which is not ensured by the other methods which generally produce only a local minimizer.

We shall need the following approximate constrained controllability-type hypothesis:

(H) Let Ω be a regular domain and $f \geq 0$ in Ω. For any subdomain $D \subset \Omega$ there is a sequence $u_n \in H^{3/2}(\partial\Omega)$, $u_n \leq 0$ in $\partial\Omega$, such that the solution y_n of (3.3), (3.4) satisfies

$$y_n\Big|_D \to y_D \quad \text{strongly in } L^2(D), \tag{3.8}$$

where y_D is the weak solution of (3.1), (3.2).

Remark. If no control constraints are imposed, the statement is valid. In [12] and [14], it is proved that it remains valid if the control constraint acts up to an open part of $\partial\Omega$ of arbitrarily small measure. Moreover, Example 3.1 below shows that (H) is satisfied in dimension one and Example 3.2 gives a hint in the same sense in dimension two. So, we conjecture that (H) is true in any dimension.

Example 3.1. If $\Omega \subset \mathbb{R}$, we take $\Omega = (0,1)$ to fix the ideas. Let $a, b \in (0,1)$, $a < b$, be arbitrary. Let f be negative in $(0,1)$ and y be the solution of

$$\begin{aligned} y''(x) &= f(x) \quad \text{in } (a,b), \\ y(a) &= y(b) = 0; \end{aligned}$$

that is,

$$y(x) = \int_a^x (x-t)f(t)\,dt + \frac{x-a}{x-b}\int_a^b (b-t)f(t)\,dt.$$

Then y may be viewed as defined over the whole (0,1) and satisfying the same equation. Moreover, we have

$$\begin{aligned} y(0) &= \int_0^a tf(t)\,dt + \frac{a}{b-a}\int_a^b (b-t)f(t)\,dt \leq 0, \\ y(1) &= (1-b)\int_a^b f(t)\left(1 - \frac{b-t}{b-a}\right)dt + \int_b^1 (1-t)f(t)\,dt \leq 0, \end{aligned}$$

so (H) is satisfied as an exact controllability property with constraints. This is also related to the convexity of the mapping $-y$.

Example 3.2. In higher dimension, the situation is by far more complicated. Let $\Omega = D(1, \frac{3}{2}) \subset \mathbb{R}^2$ be the disc centered in the point $A(1,0)$ and with radius $r = \frac{3}{2}$ (see Fig. 5.1) and let

$$y(x_1,x_2) = \begin{cases} \frac{1}{4}x_1^2 - x_2^2, & x_1 \leq 1, \\ \frac{1}{4}x_1^2 - x_2^2 - (x_1-1)^4, & x_1 \geq 1, \end{cases}$$

$$f(x_1,x_2) = \begin{cases} \frac{3}{2}, & x_1 \leq 1, \\ \frac{3}{2} + 12(x_1-1)^2, & x_1 \geq 1, \end{cases}$$

for $(x_1, x_2) \in \Omega$.

We notice that $f \geq 0$ in Ω and y is the solution of (3.3). Moreover, if $D \subset \Omega$ is defined by $D = \{(x_1,x_2) \mid x_1 \geq 0,\ y(x_1,x_2) \geq 0\}$, we obtain, obviously, $y = 0$ on ∂D. As $u = y\big|_{\partial\Omega}$ satisfies $u > 0$ between lines $x_2 = \pm\frac{1}{2}x_1$ for $x_1 < 0$ one may conclude, apparently, that (H) is not valid. This is due to the fact that in higher dimensions the positivity of Δy does not ensure the convexity of y. In our case, y is a saddle function for $x_1 \leq 1$.

However, a standard numerical treatment of the constrained control problem

$$\text{minimize } \frac{1}{2}\int_{\partial D} y^2\, d\tau$$

subject to (3.3)–(3.5) (see Example 5.1) will produce a negative boundary control $\bar{u}$ and its associated state $\bar{y}$, whose level lines are shown in Fig. 5.3, as well as Ω and D (see also Fig. 5.1).

This shows that (H) also remains valid in this setting as an approximate controllability property, but the exact controllability-type property (with constraints) will no longer be true.

Remark. In the parabolic case, distributed controllability properties with constraints are proved in the thesis of J. Henry [7], and constrained approximate boundary controllability properties have been recently announced by J. I. Diaz [5] without complete proofs.

THEOREM 3.2. *Assume* (H) *is valid. Let $E \subset D \subset \Omega$ be any fixed subdomains and $\varepsilon > 0$ be an arbitrary positive parameter. There exists a control $u_{\varepsilon,D} \in H^{3/2}(\partial\Omega)$ admissible for* (Q) *such that*

$$|V(D) - V(D_{u_{\varepsilon,D}})| < \varepsilon. \tag{3.9}$$

Proof. By a variant of a result of Pironneau [11, p. 32] we may limit the analysis to the case $\overline{E} \subset D$.

We may also assume that f is regular enough such that the weak solution of the Poisson equation in various subdomains has C^2 interior regularity. This may be obtained by a regularization of f and doesn't affect (3.9). That is, if f_λ is a regularization of f, $\lambda > 0$, the $H^1(D)$ estimate for the weak solutions associated to f, f_λ shows that the difference between the corresponding costs in (P) is "small." Moreover, this estimate is independent of D.

Let $y \in C^2(D)$ be the weak solution of the Poisson equation

$$-\Delta y = f \quad \text{in } D, \tag{3.10}$$

$$y = 0 \quad \text{on } \partial D. \tag{3.11}$$

The strong maximum principle, since $f \geq 0$, gives $y(x) > 0$ in D; therefore, $y(x) \geq C > 0$ in $\overline{E}$ by the compactness of $\overline{E} \subset D$.

By the (H) condition, there is $u_n \in H^{3/2}(\partial\Omega)$, $u_n \leq 0$ in $\partial\Omega$ such that the corresponding solution of (3.3), (3.4) satisfies (3.8). We show that, for n sufficiently large, the pair $[y_n, u_n]$ will be admissible for (Q); that is, (3.6) is also satisfied.

Let $\bar{y}$ be given by

$$\begin{aligned} -\Delta\bar{y} &= f \quad \text{in } \Omega,\\ \bar{y} &= 0 \quad \text{on } \partial\Omega; \end{aligned}$$

then $y_n \leq \bar{y}$ in Ω by comparison. The harmonic mappings $y_n - y \leq \bar{y} - y$ in D and $y_n - y \to 0$ almost everywhere D by (H). Then, it is well known that

$$y_n - y \to 0$$

in $\mathcal{H}(D)$, the space of harmonic mappings on D; that is, $y_n - y \to 0$ uniformly in compact subsets of D.

In particular, this is valid in $\overline{E}$, and taking into account that $y(x) \geq C > 0$ in $\overline{E}$, we get (3.6) for n big enough. Moreover, we have $W(u_n) = V(D_{u_n}) \to V(D)$.

Choosing n big enough such that (3.9) is fulfilled, we obtain the desired $u_{\varepsilon,D}$. The subdomain $D_{u_{\varepsilon,D}} \subset \Omega$ is associated to it by the technique of Theorem 3.1. $\Box$

Remark. Under this approach, it is not necessary to study geometric convergence properties for minimizing sequences of subdomains as is usual in optimal shape design. We show directly that the control $u_{\varepsilon,n}$ will generate a subdomain which gives a cost close to the optimal value.

Remark. We also underline the simplicity of the method in applications, by comparison with other fixed domain methods. In the mapping method of Murat and Simon [10] the variable domain is mapped on a fixed one and the control appears in the coefficients of the differential operator, while in the above technique we preserve the differential operator unchanged.

4. A distributed control approach in the optimal packaging problem. It is our aim to give a fixed-domain approach to the optimal packaging problem introduced by Zolésio, Sokolowski, and Benedict [16] and discussed by other methods in books by Haslinger and Neittaanmäki [6, Chap. X] and Tiba [15, Chap. III, §5].

The distributed control approach we are proposing here has the advantage of a large range of applications with respect to the equation, the boundary conditions, and the assumptions on the data.

Let $E \subset \Omega \subset \mathbb{R}^2$ be fixed domains with regular boundary and $\varphi\colon \overline{\Omega} \mapsto \mathbb{R}$ be a $C^2(\overline{\Omega})$ mapping with $\varphi|_{\partial\Omega} < 0$. Let D be a variable domain such that $E_\varphi \subseteq D \subseteq \Omega$, where $E \subset E_\varphi = \{x \in \Omega\ ;\ \varphi(x) > 0\}$. In D, we consider the variational inequality

$$-\Delta y + \beta(y - \varphi) \ni f \quad \text{in } D, \tag{4.1}$$

$$y = 0 \quad \text{on } \partial D, \tag{4.2}$$

with $f \in L^2(\Omega)$ and $\beta \subset \mathbb{R} \times \mathbb{R}$, the maximal monotone graph defined by

$$\beta(r) = \begin{cases} 0, & r > 0, \\ ]-\infty, 0], & r = 0, \\ \emptyset, & r < 0. \end{cases} \tag{4.3}$$

The optimal packaging problem, denoted (PA), consists of finding the subdomain $D \supset E_\varphi$ with minimal area such that the coincidence set given by

$$Z_y = \{x \in D\ ;\ y(x) = \varphi(x)\} \tag{4.4}$$

contains (or equals) E, $Z_y \supseteq E$. We note that the whole ∂D may be variable, and no global analytic description is assumed for it.

This problem contains several difficulties: it is governed by variational inequalities, and it involves variable domains and nonstandard state constraints given by (4.4).

We suppose the existence of at least one admissible domain $\check{D}$ such that the associated solution $\check{y}$ satisfies (4.4). In order to get an existence result, one has to impose some uniform regularity assumptions on ∂D (see Pironneau [11, Chap. 3]).

We associate the control system in Ω

$$-\Delta y_u + \beta(y_u - \varphi) \ni u \quad \text{in } \Omega, \tag{4.5}$$

$$y_u = 0 \quad \text{on } \partial\Omega \tag{4.6}$$

and, for a fixed Lipschitzian subdomain $D \supset E_\varphi$, the variational inequalities

$$-\Delta \tilde{y} + \beta(\tilde{y} - \varphi) \ni f \quad \text{in } D, \tag{4.7}$$

$$\tilde{y} = 0 \quad \text{on } \partial D, \tag{4.8}$$

$$-\Delta y_v + \beta(y_v - \varphi) \ni v \quad \text{in } \Omega \setminus \overline{D}, \tag{4.9}$$

$$y_v = 0 \quad \text{on } \partial D \cup \partial\Omega \tag{4.10}$$

with $v \in L^2(\Omega \setminus D)$.

The following exact controllability-type result is true.

THEOREM 4.1. *There is $v \in L^2(\Omega \setminus D)$ such that y_v given by* (4.9), (4.10) *satisfies*

$$\frac{\partial y_v}{\partial n} = -\frac{\partial \tilde{y}}{\partial \nu} \quad \textit{on } \partial D. \tag{4.11}$$

Here n and ν are the normals to ∂D inside and outside D.

Proof. We give a direct argument.

By the trace theorem, there is $\hat{y} \in H^2(\Omega \setminus \overline{D}) \cap H_0^1(\Omega \setminus \overline{D})$ such that it satisfies (4.11). We define $\bar{y} \in H_0^1(\Omega \setminus \overline{D})$ by

$$\bar{y}(x) = \max(\hat{y}(x), \varphi(x)), \tag{4.12}$$

where we have used the fact that $\varphi(x) < 0$ in $\Omega \setminus \overline{D}$ by hypothesis and $\hat{y} \in H_0^1(\Omega \setminus \overline{D})$. We notice that for any $x \in \partial D$, $\hat{y}(x) > \varphi(x)$, which is preserved by continuity in a neighbourhood $\vartheta(x)$; therefore, $\bar{y}(x) = \hat{y}(x)$ in $\vartheta(x)$—that is, $\bar{y}(x)$ satisfies (4.11) too. Moreover, we have $\bar{y}(x) \geq \varphi(x)$ in $\overline{\Omega \setminus D}$. In order to get more smoothness, one may regularize locally $\bar{y}$, which is piecewise smooth. We denote y_ε the regularized function corresponding to the regularization parameter $\varepsilon > 0$. It turns out that y_ε does not necessarily satisfy $y_\varepsilon \geq \varphi$ in $\overline{\Omega \setminus D}$.

To overcome this difficulty, we remark that the set where jumps in the derivatives of $\bar{y}$ may occur is contained in the set $\{\hat{y}(x) = \varphi(x)\} \subset \Omega \setminus \overline{D}$, compact and at positive distance $2c > 0$ from the boundaries $\partial\Omega$ and ∂D where $\hat{y}(x) > \varphi(x)$. Some local regularization y_ε is different from $\bar{y}$ on a compact neighbourhood of this set at positive distance $c > 0$ from $\partial\Omega$ and ∂D. Then, adding to y_ε a regular mapping of the type

$$K \operatorname{dist}(x, \partial\Omega)^2 \cdot \operatorname{dist}(x, \partial D)^2 \tag{4.13}$$

with a sufficiently large $K > 0$, we obtain the function y_K, which satisfies

$$\begin{aligned}
&y_K > \varphi,\\
&\text{it is in } H^2(\Omega \setminus \overline{D}) \cap H_0^1(\Omega \setminus \overline{D}),\\
&\frac{\partial y_K}{\partial n} = \frac{\partial y_\varepsilon}{\partial n} = \frac{\partial \bar{y}}{\partial n} = \frac{\partial \hat{y}}{\partial n} = -\frac{\partial \tilde{y}}{\partial \nu} \quad \text{in } \partial D,
\end{aligned}$$

due to the properties of (4.13) and the definition of y_ε.

Finally, $v = -\Delta y_K \in L^2(\Omega \setminus D)$ is the desired control since $\beta(y_K - \varphi) \equiv 0$ by $y_K(x) > \varphi(x)$ in $\Omega \setminus \overline{D}$. ||

We define the distributed control problem

$$(\mathbb{P}_n) \qquad \min_{u \in L^2(\Omega)} \int_{E_{y_u}} (1 + n(u - f)^2)\, dx \tag{4.14}$$

subject to (4.5), (4.6), (4.4), and $n \in \mathbb{N}$. Here E_{y_u} is the smallest subdomain of Ω containing E_φ and such that $y_u = 0$ on ∂E_{y_u}. The existence of E_{y_u} is trivial due to (4.6). One may ask for ∂E_{y_u} to be Lipschitzian since we have this for $\partial\Omega$ and the intersection of two Lipschitzian domains remains Lipschitzian.

It is to be underlined that, generally, (4.14) has no optimal pair since there is no coercivity ensured for u.

We show that there is a strong approximation relationship between (4.14) and (PA) for $n \to \infty$, without any reference to convergence properties of sequences of subdomains of Ω, as it is a standard argument in optimal design.

THEOREM 4.2. *For any $n \in \mathbb{N}$ and any admissible domain for the problem* (PA) *there is $u \in L^2(\Omega)$ admissible for $(\mathbb{P}_n)$ with an associated lower cost. Then*

$$\inf(\text{PA}) \geq \inf(\mathbb{P}_n). \tag{4.15}$$

Conversely, if $\delta_n > 0$ is small and $[y_n, u_n]$ is a δ_n optimal pair for $(\mathbb{P}_n)$, then E_{y_n} is an ε_n optimal subdomain for (PA) *(in a sense to be specified) with a small $\varepsilon_n > 0$ depending on δ_n.*

Proof. For any Lipschitzian $D \supset E_\varphi$, Theorem 4.1 gives $v \in L^2(\Omega \setminus D)$ such that y_v defined by (4.9) and (4.10) satisfies (4.11), with $\tilde{y}$ given by (4.7), (4.8). Then, we have that

$$y_u(x) = \begin{cases} \tilde{y}(x) & \text{in } \overline{D}, \\ y_v(x) & \text{in } \Omega \setminus D \end{cases} \tag{4.16}$$

satisfies (4.5), (4.6), $y_u \in H^2(\Omega) \cap H_0^1(\Omega)$, and $u \in L^2(\Omega)$ is given by

$$u(x) = \begin{cases} f(x) & \text{in } D, \\ v(x) & \text{in } \Omega \setminus D. \end{cases} \tag{4.17}$$

Furthermore, $y_u|_{\partial D} = 0$, so $E_{y_u} \subset D$ by definition, and $u|_{E_{y_u}} = f|_{E_{y_u}}$. Consequently, the cost in (4.14) is less or equal to meas(D), the cost associated with (PA). It is in this sense that we may speak about embedding of (PA) in (4.14) for any $n \in \mathbb{N}$, which yields (4.15).

Conversely, by (4.15), we have

$$\int_{E_{y_n}} (1 + n(u_n - f)^2)\, dx \leq \inf(\text{PA}) + \delta_n = d + \delta_n. \tag{4.18}$$

That is, $|u_n - f|_{L^2(E_{y_n})} \leq \sqrt{\frac{d+\delta_n}{n}}$. Since $y_n|_{E_{y_n}}$ satisfies the variational inequality

$$\begin{aligned} -\Delta y_n + \beta(y_n - \varphi) \ni u_n \quad & \text{in } E_{y_n}, \\ y_n = 0 \quad & \text{on } \partial E_{y_n}, \end{aligned}$$

then it is close to $\tilde{y}_n$ such that

$$\begin{aligned} -\Delta \tilde{y}_n + \beta(\tilde{y}_n - \varphi) \ni f \quad & \text{in } E_{y_n}, \\ \tilde{y}_n = 0 \quad & \text{on } \partial E_{y_n}; \end{aligned}$$

that is, $|y_n - \tilde{y}_n|_{H_0^1(E_{y_n})} \leq \varepsilon_n$. Therefore E_{y_n} is an ε_n-suboptimal pair for (4.1) in the sense that

$$\begin{aligned} &\text{meas}(E_{y_n}) \leq \inf(4.1) + \delta_n, \\ &|\tilde{y}_n - \varphi|_{H^1(E)} \leq \varepsilon_n, \end{aligned}$$

since $y_n\big|_E \equiv \varphi$ by the definition of $(\mathbb{P}_n)$.

Remark. In practical solving of $(\mathbb{P}_n)$, one also penalizes the state constraint (4.4), which gives a weaker evaluation of the same type of its violation in the above argument.

Remark. It is the controllability property from Theorem 4.1 which, in fact, guarantees that the subdomains D may be automatically generated by the control system (4.5), (4.6), thus allowing us to renounce the geometric parameters. Moreover, another specific feature of the methods developed in §§3 and 4 is that the governing equations remain unchanged when we transfer the problem on the fixed domain.

5. Numerical examples. In this section we note three relevant numerical examples. We begin with the numerical treatment of Example 3.2. The other two examples concern optimal design problems (boundary control and distributed control approaches). A complete computational analysis of the plasma problem may be found in the book by Blum [4], whose "rapid method" is theoretically investigated in §2 in a general setting. Moreover, other approximation procedures and other plasma models are also discussed in [4].

Example 5.1. Let $\Omega = \{(x_1, x_2) \in \mathbb{R}^2 \ ; \ (x_1 - 1)^2 + x_2^2 < 1.5\}$ and $\Gamma = \{(x_1, x_2) \in \mathbb{R}^2; \ \frac{1}{2}x_1^2 - x_2^2 = 0$ for $0 < x_1 < 1$ and $\frac{1}{4}x_1^2 - x_2^2 - (x_1 - 1)^4 = 0$ for $x_1 > 1\}$; see Fig. 5.1.

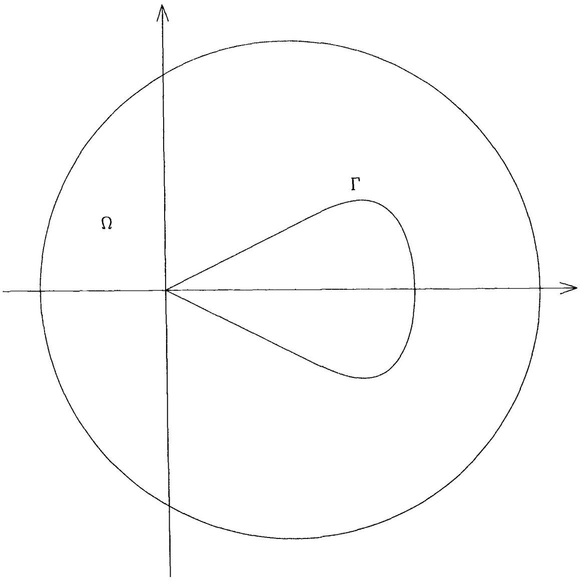


FIG. 5.1.

Consider the minimization problem

$$\text{(5.1)} \qquad \underset{u}{\text{minimize}} \left\{ J(u) = \frac{1}{2} \int_\Gamma y^2 \, d\Gamma \right\}$$

subject to

$$\text{(5.2)} \qquad \begin{cases} -\Delta y = f & \text{in } \Omega, \\ y = u & \text{on } \partial\Omega, \end{cases}$$

$$\text{(5.3)} \qquad u \le 0 \quad \text{on } \partial\Omega,$$

where f is indicated in Example 3.2.

We have discretized the state problem (5.2) using the finite element method with triangular elements consisting of 661 interior nodes and 120 boundary nodes (control variables) and piecewise linear approximation. We formulate the minimization problem (5.1)–(5.3) as a mathematical programming problem and use the sequential quadratic programming method (subroutine E04VDF of the Numerical Algorithms Group library). The state problem was solved by the Cholesky method. As usual, the adjoint state technique was applied to compute the gradient of the cost function $J(\mathbf{u})$ (see J. Haslinger and P. Neittaanmäki [6]).

We have chosen the initial guess as follows: $u_i^0 = -10.0$, $i = 1, \dots, 120$. This gives the value 187.77 for $J(\mathbf{u})$.

After 17 iterations the value of $J(\mathbf{u})$ was reduced to $4.23 \cdot 10^{-4}$. Figure 5.2 shows the value of $\mathbf{u}$ on $\partial\Omega$ (as a function of angle parameter). The expected symmetry of $\mathbf{u}$ with respect to 3.1416 is slightly perturbed due to the round-off errors which accumulate in the way that we solve the algebraic system and due to the chosen triangulation of the disc Ω.

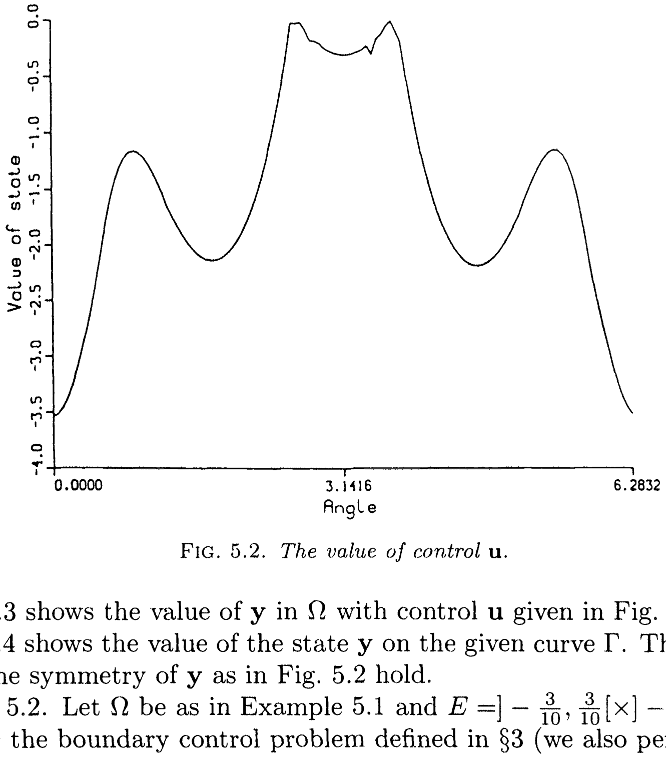


FIG. 5.2. *The value of control* $\mathbf{u}$.

Figure 5.3 shows the value of $\mathbf{y}$ in Ω with control $\mathbf{u}$ given in Fig. 5.2.

Figure 5.4 shows the value of the state $\mathbf{y}$ on the given curve Γ. The same considerations on the symmetry of $\mathbf{y}$ as in Fig. 5.2 hold.

Example 5.2. Let Ω be as in Example 5.1 and $E =]-\frac{3}{10}, \frac{3}{10}[\times]-\frac{7}{10}, \frac{7}{10}[$.

Consider the boundary control problem defined in §3 (we also penalize the state constraint):

$$\text{(5.4)} \qquad \underset{u}{\text{minimize}} \left\{ J(y(u)) = \frac{1}{2}\|y - y_d\|^2_{H^1(E)} + \frac{1}{2\lambda}\int_E y_-^2 \, dx \right\}$$

subject to

$$\text{(5.5)} \qquad \begin{cases} -\Delta y = f & \text{in } \Omega, \\ y = u & \text{on } \partial\Omega, \end{cases}$$

$$\text{(5.6)} \qquad u \leq 0 \quad \text{on } \partial\Omega.$$

Here $y_d = \frac{3}{2} - 4x_1^2 - x_2^2$ and $f = 10$.

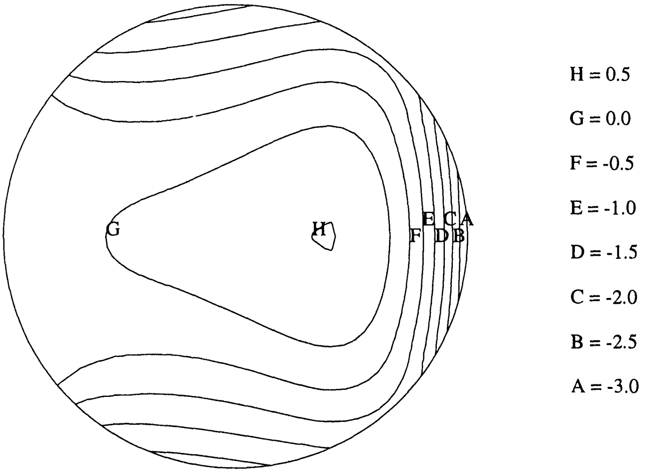


FIG. 5.3. *The value of* $\mathbf{y}$ *in* Ω *with control* $\mathbf{u}$.

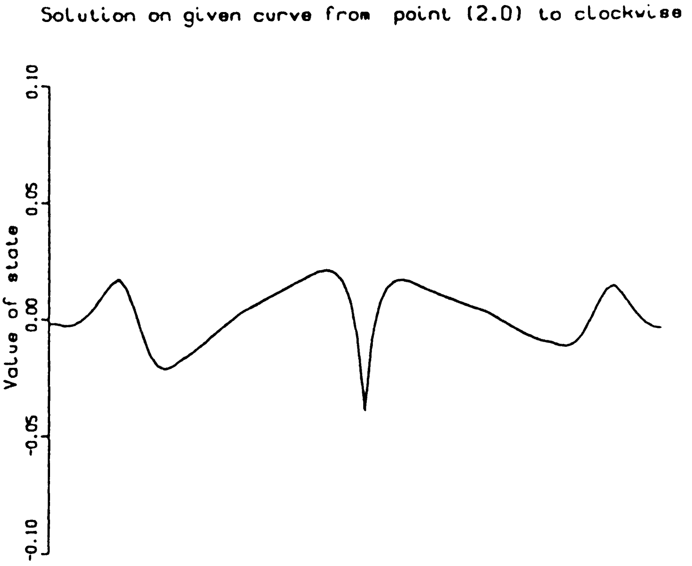


FIG. 5.4. *The value of the state* $\mathbf{y}$ *on the given curve* Γ.

We have applied a finite-element grid and elements similar to those in Example 5.1. The initial guess for the control was $u_i^0 = -4.5$, $i = 1, \dots, 120$, giving the value 3.35 for the cost functional J with the penalty parameter $\lambda = 10^{-3}$. After 12 iterations the value of the cost functional was reduced to $2 \cdot 10^{-4}$. The obtained control and the corresponding state can be seen in Figs. 5.5 and 5.6.

The same problem was solved in [14] with the conventional moving mesh technique, and the radial coordinates of the boundary nodes were chosen as the control variables. The initial guess chosen was a unit disc. The moving mesh approach gives essentially the same design with fewer sequential quadratic programming iterations than the proposed boundary control approach, but the total computational burden is much heavier, as the finite element mesh has to be updated at the beginning of each iteration. As the coefficient matrix in the moving grid method depends on control variables, one cannot utilize the same factorization of the coefficient matrix as in the case of the boundary control approach.

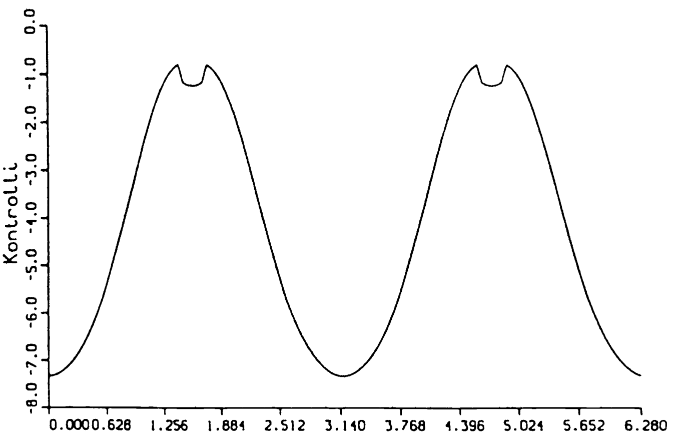


FIG. 5.5. *Obtained control in Ex. 5.2.*

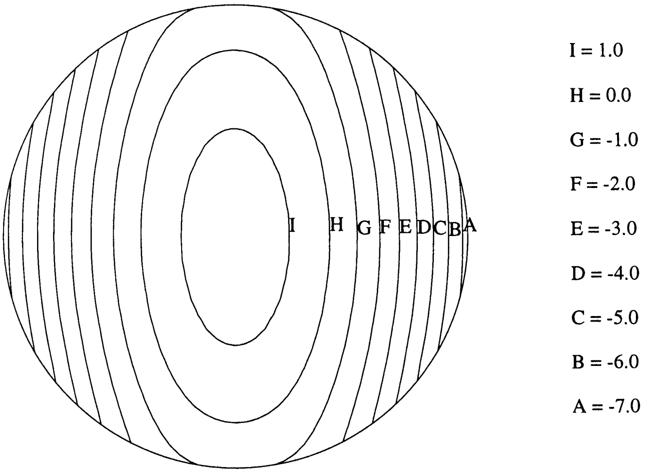


FIG. 5.6. *State corresponding to the control given in Ex. 5.2.*

Example 5.3. In connection with §4, we choose $\Omega = \{(x_1, x_2) \in \mathbb{R}^2\ ;\ x_1^2 + x_2^2 < 1\}$, the obstacle

$$\varphi(x_1, x_2) = \begin{cases} \frac{1}{7} - 2\left(x_1 + \frac{1}{2}\right) - \frac{1}{2}x_2^2, & x_1 < 0, \\ \frac{1}{7} - 2\left(x_1 - \frac{1}{2}\right) - \frac{1}{2}x_2^2, & x_1 \geq 0, \end{cases} \tag{5.7}$$

and $E = E_1 \cup E_2 = [-\frac{5}{8}, -\frac{3}{8}] \times [-\frac{3}{8}, \frac{3}{8}] \cup [\frac{3}{8}, \frac{5}{8}] \times [-\frac{3}{8}, \frac{3}{8}]$, $f \equiv -3$ in Ω. In (4.14), we also penalize the state constraint (4.4) and solve the control problem

$$\underset{u \in L^2(\Omega)}{\text{minimize}} \left\{ J = \int_{E_{y_u}} \left[1 + \frac{1}{2\varepsilon_1}(u - f)^2 \right] dx + \frac{1}{2\varepsilon_2} \int_E (y - \varphi)^2\, dx \right\} \tag{5.8}$$

subject to (4.5), (4.6). Here φ is given by (5.7).

If we take $\varepsilon_1 = 10^{-3}$, $\varepsilon_2 = 10^{-5}$, and initial iteration $u_0 = 2$, which gives $E_{y_{u_0}} = \Omega$, then the initial cost is about $3.569 \cdot 10^4$. After five iterations the cost was reduced to 0.9571 and the value of the penalization terms was about $3 \cdot 10^{-9}$, which shows that the error in the state constraints or in the solution is very small.

In Fig. 5.7, the final solution of the optimal shape problem is shown. One should

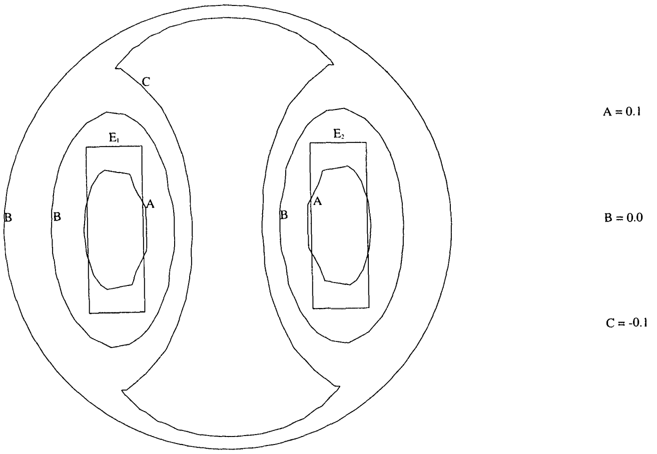


FIG. 5.7.

note that the solution consists of two disjoint open sets. Therefore the topology of the solution set may change during the computation. (Initially it was connected.) This is another advantage over the standard boundary variation technique.

Acknowledgments. The authors are indebted to R. Mäkinen and T. Räisänen for their help with the numerical tests.

REFERENCES

[1] C. ATAMIAN AND P. JOLY, *Une analyse de la méthode des domaines fictifs pour le problème de Helmholtz extérieur*, Rapport INRIA 1378, 1991.

[2] V. BARBU AND D. TIBA, *Boundary controllability of the coincidence set in the obstacle problem*, SIAM J. Control Optim., 29 (1991), pp. 1150–1159.

[3] J. BLUM, *Sur quelques problèmes d'analyse numérique et de contrôle optimal en physique des plasmas*, Thèse de l'Université Paris VI, 1985.

[4] ———, *Numerical Simulation and Optimal Control in Plasma Physics. With Applications to Tokamaks*, John Wiley, Gauthier-Villars, New York, 1989.

[5] J. I. DIAZ, *Sur la contrôlabilité approchée des inéquations variationnelles et d'autres problèmes paraboliques nonlinéaires*, C. R. Acad. Sci. Paris Sér. I Math., 312 (1991), pp. 519–522.

[6] J. HASLINGER AND P. NEITTAANMÄKI, *Finite Element Approximation of Optimal Shape Design: Theory and Applications*, John Wiley, New York, 1988.

[7] J. HENRY, *Quelques problèmes de contrôlabilité de systèmes paraboliques*, Thèse de l'Université de Paris VI, 1978.

[8] K. H. HOFFMANN, J. HASLINGER, AND M. KOCVARA, *Control/fictitious domain method in solving optimal shape design problems*, DFG preprint no. 278, University of Augsburg, 1991.

[9] J. L. LIONS, *Contrôle optimal de systèmes gouvernés par des équations aux dérivées partielles*, Dunod, Paris, 1968.

[10] F. MURAT AND J. SIMON, *Etude de problèmes d'optimal design*, in Optimization Techniques, Modelling and Optimization in the Service of Man, J. Cea, ed., Lecture Notes in Control and Information Sciences 41, Springer-Verlag, Berlin, 1976.

[11] O. PIRONNEAU, *Optimal Shape Design for Elliptic Systems*, Springer-Verlag, Berlin, 1984.

[12] D. TIBA, *Une approche par contrôlabilité frontière dans les problèmes de design optimal*, C. R. Acad. Sci. Paris Sér. I Math., 310 (1990), pp. 175–177.

[13] ———, *Controllability properties for elliptic systems, the fictitious domain method and optimal shape design problems*, Rapport no. 1500, INRIA, 1991.

[14] D. TIBA, P. NEITTAANMÄKI, AND R. MÄKINEN, *Controllability-type properties for elliptic systems and applications*, in Control and Estimation of Distributed Parameter Systems, F. Kappel and K. Kunisch, eds., Birkhäuser, Basel, 1991.

[15] D. TIBA, *Optimal Control of Nonsmooth Distributed Parameter Systems*, Lecture Notes in Mathematics 1459, Springer-Verlag, Berlin, 1990.

[16] J. P. ZOLÉSIO, J. SOKOLOWSKI, AND B. BENEDICT, *Shape optimization for contact problems*, in Lecture Notes in Control and Information Sciences 59, Springer-Verlag, Berlin, 1984, pp. 784–799.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 33, No. 5, pp. 1603 1635, September 1995

© 1995 Society for Industrial and Applied Mathematics
015

SECOND-ORDER OPTIMALITY CONDITIONS IN SETS OF L^∞ FUNCTIONS WITH RANGE IN A POLYHEDRON*

JOSEPH C. DUNN†

Abstract. Formal extensions of the general second-order necessary conditions and sufficient conditions for local optimality in a polyhedral convex set $U \subset \mathbb{R}^m$ are established for L^∞-local optimality and L^2-local optimality in the infinite-dimensional nonpolyhedral convex set Ω of L^∞ functions $u(\cdot) : [0,1] \to U$. A more refined analysis for nonconvex cost functions with specially structured differentials yields optimality conditions that apply to an important class of constrained input Bolza optimal control problems. The gap between the necessary conditions and sufficient conditions in this setting is uncharacteristically small for infinite-dimensional problems. In the control problem context, the L^∞-local optimality conditions and L^2-local optimality conditions entail a mild strengthening of a pointwise strict complementarity condition and variants of the Legendre–Clebsch condition and the Pontryagin minimum principle. In related recent studies, similar second-order sufficient conditions for the special case $U = [0,\infty)$ are the key hypotheses in corresponding local convergence theories for iterative constrained minimization algorithms.

Key words. constrained minimization, function spaces, necessary conditions, sufficient conditions, optimal control, Legendre–Clebsch condition, Pontryagin minimum principle

AMS subject classifications. 49K15, 46N10, 90C06, 49M07, 49M10, 65K10

1. Introduction.

Many problems in the calculus of variations and optimal control theory belong to the class of infinite-dimensional nonlinear programs

$$\min_{u \in \Omega \cap \mathcal{D}} J(u), \tag{1a}$$

$$\Omega = \{u \in \mathsf{L}_m^\infty\,[0,1] : u(t) \overset{\text{a.e.}}{\in} U\}, \tag{1b}$$

where $\mathsf{L}_m^\infty\,[0,1]$ is the vector space of Lebesgue-measurable essentially bounded functions $u : [0,1] \to \mathbb{R}^m$, J is a real-valued function that is differentiable in some sense on a domain $\mathcal{D}$ in $\mathsf{L}_m^\infty\,[0,1]$, and U is a nonempty polyhedral convex set in $\mathbb{R}^m$, i.e., an intersection of finitely many closed half-spaces in $\mathbb{R}^m$. Although Ω is *not* a polyhedral set, its special Cartesian product structure suggests that the well-known second-order necessary conditions and sufficient conditions for local optimality in U may have valid formal extensions in Ω. This question is explored here for two distinct species of local optimality corresponding to the nonequivalent norms $\|\cdot\|_\infty$ and $\|\cdot\|_2$ on $\mathsf{L}_m^\infty[0,1]$ and for nonconvex objective functions J with differentials satisfying structure and continuity hypotheses that are met by a large class of Bolza optimal control problems, and in other contexts as well.

The theorems in §§5 and 6 greatly extend the L^∞-local optimality and L^2-local optimality necessary conditions and sufficient conditions established in [1] for the special case $U = [0,\infty)$. These theorems are proved with geometric techniques that directly exploit the properties of polyhedra and the product structure in Ω, and thereby dispense with the constraint qualifications and multiplier functionals invoked in Lagrangian second-order necessary conditions for general infinite-dimensional nonlinear programs (cf. [3]–[11]). Smoothness hypotheses imposed in the L^∞-local optimality

* Received by the editors November 16, 1992; accepted for publication (in revised form) May 24, 1994. This research was supported by National Science Foundation research grant DMS-9205240.

† Mathematics Department, Box 8205, North Carolina State University, Raleigh, North Carolina 27695-8205.

sufficient conditions of Theorem 6.4 are also substantially weaker than the corresponding hypotheses in [1] and are now comparable to analogous assumptions in the L^∞ sufficiency analysis of [9]. Under stronger smoothness requirements like those in [1], Theorem 6.6 ensures that L^2-local optimality is implied by L^∞-local optimality sufficient conditions and a strengthened form of the L^2-local optimality necessary condition. The latter condition is closely related to the Pontryagin minimum principle in the context of ordinary differential equation (ODE) optimal control problems.

Reference [2] develops another far-reaching extension of the L^∞-local optimality sufficiency theorem in [1] for ODE optimal control problems with terminal state equality constraints and nonpolyhedral time-dependent admissible control input sets $U(t)$ prescribed by finitely many smooth nonlinear inequalities; however, the Lagrangian formulation in [2] imposes constraint qualifications on the inequalities that define U and differs in other important respects from the representation-free geometric approach pursued here. Lemma 6.5, Note 6.3, and the related discussion in §6 interpret the principal algebraic hypotheses of [2] in the geometric framework of Theorems 6.4 and 6.6.

The gaps between the main necessary conditions and sufficient conditions proved here are uncharacteristically small for infinite-dimensional nonlinear programs in general, and optimal control problems in particular [3], [9]. As noted in [10], the sufficiency gap has some computational significance since standard iterative constrained optimization algorithms tend to exhibit their best local convergence behavior near stationary points that satisfy known second-order sufficient conditions. The sufficient conditions in [1] and [2] have already produced new local convergence theorems for gradient projection methods [12], [13] and sequential quadratic programming algorithms [2] in infinite-dimensional settings. While the convergence rate estimates in these theorems are similar to their finite-dimensional counterparts, the developments in [12], [13], and [19] indicate potentially interesting differences in the computational implications of L^∞-local and L^2-local convergence theorems for approximate finite-dimensional implementations of the subject algorithms. The sufficient conditions proved in the present paper support similar local convergence theorems for gradient projection iterations and continuous-time optimal control problems with vector-valued control functions satisfying affine inequality constraints pointwise in bounded time intervals; these theorems will be stated and proved in a later paper.

2. Preliminary considerations.

The following brief review is meant to serve as a bridge between familiar terminology and theorems in $\mathbb{R}^n$ and analogous formulations for Ω in the infinite-dimensional space $\mathsf{L}^\infty_m[0,1]$.

Let J be a real valued function on a domain $\mathcal{D}$ in a vector space V. Then:

(i) J is *twice directionally differentiable* iff each point of $\mathcal{D}$ is an *internal point* of $\mathcal{D}$ [14] and the following limits exist for all u in $\mathcal{D}$ and v, w in V:

$$d^1J(u\ ;\ v) = \lim_{s\to 0} \frac{J(u+sv)-J(u)}{s},$$

$$d^2J(u\ ;\ v,w) = \lim_{s\to 0} \frac{d^1J(u+sv\ ;\ w)-d^1J(u\ ;\ w)}{s}.$$

(ii) J is *twice Gâteaux differentiable relative to a norm* $\|\cdot\|$ *on* V iff J is twice directionally differentiable and, for each u in $\mathcal{D}$, the differentials $d^1J(u\ ;\ \cdot)$ and

$d^2\ J(u\ ;\ \cdot,\cdot)$ are linear and bilinear, respectively, and continuous relative to $\|\cdot\|$.[1]

(iii) J is twice *Fréchet differentiable relative to a norm* $\|\cdot\|$ *on* V iff J is twice Gâteaux differentiable relative to $\|\cdot\|$, $\mathcal{D}$ is open relative to $\|\cdot\|$, and, for all u in $\mathcal{D}$,

$$|J(u+v) - J(u) - d^1J(u\ ;\ v)| = o(\|v\|),$$

$$\sup_{\|w\|=1} |d^1J(u+v\ ;\ w) - d^1J(u\ ;\ w) - d^2J(u\ ;\ v,w)| = o(\|v\|).$$

(iv) J is twice *continuously* Fréchet differentiable relative to a norm $\|\cdot\|$ on V iff J is twice Fréchet differentiable relative to $\|\cdot\|$ and, for all u in $\mathcal{D}$,

$$\lim_{\|\Delta u\|\to 0}\ \sup_{\|v\|=1} |d^1J(u+\Delta u\ ;\ v) - d^1J(u\ ;\ v)| = 0,$$

$$\lim_{\|\Delta u\|\to 0}\ \sup_{\|w\|=1}\ \sup_{\|v\|=1} |d^2J(u+\Delta u\ ;\ v,w) - d^2J(u\ ;\ v,w)| = 0.$$

The uniform approximation property imposed in the definition of Fréchet differentiabilty is closely related to the continuity properties of the directional derivative maps $d^1J(\cdot\ ;\ v)$ and $d^2J(\cdot\ ;\ v,w)$. More precisely, if J is twice Gâteaux differentiable relative to a norm $\|\cdot\|$ on V, if $\mathcal{D}$ is open relative to $\|\cdot\|$, and if the real functions $d^1J(\cdot\ ;\ v)$ and $d^2J(\cdot\ ;\ v,w)$ are continuous at each u in $\mathcal{D}$ relative to $\|\cdot\|$, *uniformly in v and w on the unit sphere* $S(0,1) = \{v \in \mathsf{V} : \|v\| = 1\}$, then J is twice continuously Fréchet differentiable relative to $\|\cdot\|$. When J is twice continuously Fréchet differentiable, the second differential is *symmetric*, i.e.,

$$d^2J(u\ ;\ v,w) = d^2J(u\ ;\ w,v),$$

and Taylor's formula yields the important estimate

$$J(u+v) = J(u) + d^1J(u\ ;\ v) + \tfrac{1}{2}\ d^2J(u\ ;\ v,v) + o(\|v\|^2).$$

Note that in $\mathbb{R}^n$, all norms are equivalent, linearity implies continuity in any norm, continuity and Gâteaux differentiability and Fréchet differentiability are norm-invariant properties, and the directional derivative uniform continuity requirements are met iff all first and second order partial derivatives of J are continuous. On the other hand, in infinite-dimensional vector spaces, two norms are not necessarily equivalent; linearity does not imply continuity; and continuity, Gâteaux differentiability and Fréchet differentiability are norm-dependent properties.

Let J be a twice Gâteaux differentiable real-valued function on a domain $\mathcal{D}$ in a vector space V with inner product $\langle\cdot,\cdot\rangle$ and induced norm $\|u\| = \sqrt{\langle u,u\rangle}$. At each u in $\mathcal{D}$, there is at most one vector $\nabla J(u)$ in V and one continuous linear map $\nabla^2J(u) : \mathsf{V} \to \mathsf{V}$ such that

$$d^1J(u\ ;\ v) = \langle \nabla J(u), v\rangle \quad \text{and} \quad d^2J(u\ ;\ v,w) = \langle v, \nabla^2J(u)w\rangle.$$

The vector $\nabla J(u)$ and operator $\nabla^2J(u)$ are called the *gradient* and *Hessian* of J at u, *corresponding to the inner product* $\langle\cdot,\cdot\rangle$. When $\nabla J(u)$ and $\nabla^2J(u)$ exist at every

[1] This definition of Gâteaux differentiability conforms to [18]. In some references, the terms directional differentiability and Gâteaux differentiability are used interchangeably [23].

u in $\mathcal{D}$ for the given inner product, the function J is then twice continuously Fréchet differentiable (relative to the induced norm) iff $\mathcal{D}$ is open and the maps $\nabla J(\cdot)$ and $\nabla^2 J(\cdot)$ are continuous (relative to the induced norm and the corresponding operator norm). If V is a *Hilbert space*, then the Riesz representation theorem ensures that the gradient and Hessian do exist at each point in the domain of J. In particular, in $\mathbb{R}^n$, every twice Gâteaux differentiable J has gradients and Hessians for the standard inner product, with the familiar matrix representations in the standard basis, i.e.,

$$(\nabla J(u))_i = \frac{\partial J}{\partial u_i}(u) \quad \text{and} \quad (\nabla^2 J(u)v)_i = \sum_{j=1}^{m} \frac{\partial^2 J}{\partial u_i \partial u_j}(u) v_j.$$

Suppose that C is a *convex* set in a vector space V with inner product $\langle \cdot, \cdot \rangle$ and induced norm $\|u\| = \sqrt{\langle u, u \rangle}$. At each u in C, let $\mathcal{N}_C(u)$, $\mathcal{T}_C(u)$, $\mathsf{N}_C(u)$, $\mathsf{T}_C(u)$, and $\operatorname{ri} \mathcal{N}_C(u)$ denote *the cone of outer normals, the polar cone of tangents, and the closed linear hull, orthogonal complement, and relative interior of* $\mathcal{N}_C(u)$ *in* V, respectively; i.e.,

$$\mathcal{N}_C(u) = \{w \in \mathsf{V} : \forall v \in C \quad \langle w, v - u \rangle \leq 0\},$$

$$\mathcal{T}_C(u) = \{v \in \mathsf{V} : \forall w \in \mathcal{N}_C(u) \quad \langle v, w \rangle \leq 0\},$$

$$\mathsf{N}_C(u) = \text{cl span } [\mathcal{N}_C(u)],$$

$$\mathsf{T}_C(u) = \mathcal{N}_C(u)^{\perp} = \mathsf{N}_C(u)^{\perp},$$

and

$$\operatorname{ri} \mathcal{N}_C(u) = \{w \in \mathcal{N}_C(u) : \exists \delta > 0 \; \forall x \in \mathsf{N}_C(u) \quad \|x\| \leq \delta \Rightarrow w + x \in \mathcal{N}_C(u)\}.$$

The closed subspace $\mathsf{T}_C(u)$ is the largest subspace contained in $\mathcal{T}_C(u)$ and is sometimes called the *lineality* in the tangent cone. In a *Hilbert space* V, $\mathcal{N}_C(u)$ is the polar cone for $\mathcal{T}_C(u)$, $\mathsf{N}_C(u)$ is the orthogonal complement of $\mathsf{T}_C(u)$, and for all v in V, the usual orthogonal projection decompositions hold with respect to the pairs $(\mathcal{N}_C(u), \mathcal{T}_C(u))$ and $(\mathsf{N}_C(u), \mathsf{T}_C(u))$; i.e., for all v in V,

$$v = P_{\mathcal{N}_C(u)}\, v + P_{\mathcal{T}_C(u)}\, v, \qquad v = P_{\mathsf{N}_C(u)}\, v + P_{\mathsf{T}_C(u)}\, v,$$

where P_X denotes orthogonal projection into the set $X \subset \mathsf{V}$.

3. Structure and smoothness assumptions. Much of the analysis in §§5 and 6 applies to objective functions J that satisfy the following conditions:

(i) J is twice directionally differentiable on its domain $\mathcal{D}$ in $\mathsf{L}_m^\infty[0,1]$.

(ii) For each u in $\mathcal{D}$ there is a vector $\nabla J(u) \in \mathsf{L}_m^\infty[0,1]$, an essentially bounded $m \times m$ matrix-valued function $S(\cdot) \in \mathsf{L}_{m \times m}^\infty[0,1]$, and a square-integrable $m \times m$ matrix-valued function $K(\cdot) \in \mathsf{L}_{m \times m}^2([0,1] \times [0,1])$ such that

$$d^1 J(u\,;\, v) = \langle \nabla J(u), v \rangle_2 \tag{2a}$$

and

$$d^2 J(u\,;\, v, w) = \langle v, \nabla^2 J(u) w \rangle_2 \tag{2b}$$

with

$$(2c)\qquad \left(\nabla^2 J(u)w\right)(t) = S(u)(t)w(t) + \int_o^1 K(u)(t,s)w(s)ds, \quad t \in [0,1],$$

where $\langle v, w\rangle_2 = \int_0^1 \langle v(t), w(t)\rangle dt$, $\|u\|_2 = \sqrt{\langle u, u\rangle_2}$, and $\langle \xi, \eta\rangle = \sum_{i=1}^m \xi_i\eta_i$ for ξ, η in $\mathbb{R}^m$.

(iii) For $\nu = 2$ or ∞, the domain $\mathcal{D}$ is open relative to $\|\cdot\|_\nu$, and the associated mappings $S(\cdot)$ and $K(\cdot)$ are continuous in the sense that for all u in $\mathcal{D}$

$$(3a)\qquad \lim_{\|v-u\|_\nu \to 0} \|S(v) - S(u)\|_\infty = 0$$

and

$$(3b)\qquad \lim_{\|v-u\|_\nu \to 0} \|K(v) - K(u)\|_2 = 0,$$

where $\|v\|_\infty = \operatorname{ess\,sup}_{t\in[0,1]} \|v(t)\|$, $\|\xi\| = \sqrt{\langle \xi, \xi\rangle}$ and

$$\|S(u)\|_\infty = \operatorname*{ess\ sup}_{t\in[0,1]} \|S(u)(t)\|,$$

$$\|K\|_2 = \left(\int_0^1 \int_0^1 \|K(u)(t,s)\|^2 ds dt\right)^{\frac{1}{2}},$$

$$\|S(u)(t)\| = \sup_{\|\xi\|=1} \|S(u)(t)\xi\|,$$

$$\|K(u)(t,s)\| = \sup_{\|\xi\|=1} \|K(u)(t,s)\xi\|.$$

Assumptions (i) and (ii) imply that J is Gâteaux differentiable relative to *either* of the norms $\|\cdot\|_\infty$ or $\|\cdot\|_2$, with gradients $\nabla J(u)$ and Hessians $\nabla^2 J(u)$ in the incomplete inner product space $\{\mathsf{L}_m^\infty[0,1], \langle\cdot,\cdot\rangle_2, \|\cdot\|_2\}$. In $\mathbb{R}^n$, the analogue of the structure condition (2c) in (ii) can always be met by writing $\nabla^2 J(u)$ as the sum of its diagonal and off-diagonal parts; however, (2c) is a *nontrivial* hypothesis in the vector space $\mathsf{L}_m^\infty[0,1]$. In this infinite-dimensional setting, condition (2c) ensures that the Hessian operator $\nabla^2 J(u)$ acts essentially like the simple multiplication operator $S(u)$ on vectors v that vanish outside a set with small Lebesgue measure in $[0,1]$. This "local diagonal dominance" property is essential for the proofs of the L^∞-local optimality and L^2-local optimality sufficient conditions in Theorems 6.4 and 6.6 and the matching necessary conditions in Theorems 5.3 and 5.4.

If assumption (iii) holds along with (i) and (ii), then J is *twice continuously Fréchet differentiable* relative to $\|\cdot\|_\nu$ on $\mathsf{L}_m^\infty[0,1]$ and, in addition,

$$(4a)\qquad J(v) - J(u) = \langle \nabla J(u), v-u\rangle_2 + \tfrac{1}{2}\langle v-u, \nabla^2 J(u)\,(v-u)\rangle_2 + r(u;v)$$

with

$$(4b)\qquad \lim_{\|v-u\|_\nu \to 0} \frac{|r(u;v)|}{\|v-u\|_2^2} = 0.$$

For $\nu = 2$, (4b) asserts that $r(u;v) = o(\|v-u\|_2^2)$, as in formula 4 of [1]; this standard L^2 version of Taylor's remainder estimate is invoked in the L^2-local optimality *and* L^∞-local optimality analyses in [1]. For $\nu = \infty$, (4) reduces to the weaker condition in the L^∞ analyses of [9] and [2]; this form of (4) is imposed here in §6 (and will also support the L^∞ sufficient conditions in [1]). Note that the continuity hypotheses in (iii) are roughly analogous to partial derivative continuity conditions in $\mathbb{R}^n$; however, the norms $\|\cdot\|_\infty$ and $\|\cdot\|_2$ are not equivalent on the infinite-dimensional space $\mathsf{L}_m^\infty[0,1]$, and conditions (3) with $\nu = 2$ are *stronger* than (3) with $\nu = \infty$. Note also that for $\nu = \infty$, (4b) is stronger than the standard Taylor remainder estimate for objective functions J that are twice continuously Fréchet differentiable relative to $\|\cdot\|_\infty$; thus, conditions (3) with $\nu = \infty$ imply *more* than twice continuous Fréchet differentiability in the L^∞ norm but *less* than twice continuous Fréchet differentiability in the L^2 norm.

Note 3.1. In an optimal control setting, the structure/continuity conditions (2) and (3) are satisfied with $\nu = \infty$ by a large class of Bolza objective functions,

$$J(u) = P(x(1)) + \int_0^1 f^0(t, x(t), u(t))dt, \tag{5a}$$

where $x(\cdot) : [0,1] \to \mathbb{R}^n$ is the solution of an initial value problem

$$x(0) = x_0, \tag{5b}$$

$$\frac{dx}{dt} = f(t, x(t), u(t)), \quad t \in [0,1], \tag{5c}$$

and P, f^0, and f satisfy relatively weak smoothness and growth restrictions. Conditions (2) and (3) are also satisfied with $\nu = 2$ by a smaller but still important class of *quasi-quadratic* Bolza objective functions with associated system Hamiltonians $H(t,\psi,x,u) = \langle \psi, f(t,x,u)\rangle + f^0(t,x,u)$ that are quadratic in $u \in \mathbb{R}^m$. (See [13] for a discussion of the case $m = 1$.)

4. Formal extensions of the optimality conditions in polyhedra.

If C is a *polyhedral* convex set in $\mathbb{R}^n$, then the tangent cone $\mathcal{T}_C(u)$ coincides with the *cone of feasible direction vectors at* u, every vector in the subspace $\mathsf{T}_C(u)$ is a feasible direction vector, and $u + \mathsf{T}_C(u)$ is the affine hull of the unique polyhedral face of C containing u in its relative interior. These observations and an application of elementary calculus immediately produce the following geometric expression of the basic necessary conditions for optimality in polyhedral convex subsets of $\mathbb{R}^n$.

THEOREM 4.1. *Suppose that C is a polyhedral convex set in $\mathbb{R}^n$, u is a local minimizer for the restriction of $J : \mathcal{D} \to \mathbb{R}^1$ to any line in C, and J is twice directionally differentiable at u. Then*

$$\forall v \in C \quad d^1J(u\,;\, v-u) \geq 0$$

and

$$\forall v \in \mathsf{T}_C(u) \quad d^1J(u\,;\, v) = 0 \quad \textit{and} \quad d^2J(u\,;\, v,v) \geq 0.$$

Note that if J is Gâteaux differentiable, then the first-order necessary condition in Theorem 4.1 says that $-\nabla J(u) \in \mathcal{N}_C(u)$.

The corresponding second-order *sufficient* conditions for local optimality in polyhedra also have a representation-free geometric expression. In this theorem, the directional differentiability hypothesis is superceded by a stronger continuous Fréchet differentiability assumption, and the first- and second-order necessary conditions are replaced by a *strict* complementarity condition on $\nabla J(u)$ and a *coercivity* condition on the quadratic form $\langle v, \nabla^2 J(u) v\rangle$ in the subspace $\mathsf{T}_C(u)$. Under these circumstances, Taylor's formula can be used to establish explicit quadratic growth rate estimates for J near u in C.

THEOREM 4.2. *Suppose that C is a polyhedral convex set in $\mathbb{R}^n$ and that $J : \mathcal{D} \to \mathbb{R}^1$ is twice continuously Fréchet differentiable. In addition, suppose that the following conditions hold at a point $u \in C \cap \mathcal{D}$:*

$$-\nabla J(u) \in \operatorname{ri} \mathcal{N}_C(u)$$

and

$$\exists \hat{\mu} > 0 \ \forall v \in \mathsf{T}_C(u) \quad \langle v, \nabla^2 J(u) v\rangle \geq \hat{\mu}\, \|v\|^2.$$

Then u is a strict local minimizer for J in C, and for each $\mu \in (0, \hat{\mu})$, there is a corresponding $\delta > 0$ such that

$$\forall v \in C \quad \|v-u\| \leq \delta \Rightarrow J(v) - J(u) \geq \tfrac{1}{2}\mu\, \|v-u\|^2.$$

Theorem 4.2 can be proved by using the coercivity condition to estimate $J(v) - J(u)$ for increments $v - u$ in a cone of vectors *nearly orthogonal* to the subspace $\mathsf{N}_C(u)$, and the strict complementarity condition to make a similar estimate for $v - u$ in the complement of this cone and v in C. In contrast to the standard indirect proof of the second-order sufficient conditions for nonlinear programs in $\mathbb{R}^n$ [22], this proof technique is not tied to compactness of the unit sphere and can therefore be applied in infinite-dimensional settings as well as $\mathbb{R}^n$; moreover, it establishes explicit quadratic growth estimates for J that are needed in convergence theories for iterative constrained minimization algorithms [17], [20], [21]. Variants of this proof strategy are implicit in the treatment of infinite-dimensional nonlinear programs with finitely many scalar-valued nonlinear inequality constraints in [18], and similar techniques are used in [1] and [2] and here in the proof of Theorem 6.1.

When U is a polyhedral convex set in $\mathbb{R}^m$, the k-fold Cartesian product $\Omega_k = U \times \cdots \times U$ is a polyhedral convex set in $\mathbb{R}^{km} = \mathbb{R}^m \times \cdots \times \mathbb{R}^m$, and for each $u = (u_1 \cdots u_k)$ in $\mathbb{R}^{km}$,

$$\mathcal{N}_{\Omega_k}(u) = \{w \in \mathbb{R}^{km} : w_i \in \mathcal{N}_U(u_i) \quad i = 1, \cdots, k\}, \tag{6a}$$

$$\mathcal{T}_{\Omega_k}(u) = \{w \in \mathbb{R}^{km} : w_i \in \mathcal{T}_U(u_i) \quad i = 1, \cdots, k\}, \tag{6b}$$

$$\mathsf{N}_{\Omega_k}(u) = \{w \in \mathbb{R}^{km} : w_i \in \mathsf{N}_U(u_i) \quad i = 1, \cdots, k\}, \tag{6c}$$

$$\mathsf{T}_{\Omega_k}(u) = \{w \in \mathbb{R}^{km} : w_i \in \mathsf{T}_U(u_i) \quad i = 1, \cdots, k\}. \tag{6d}$$

These cones and subspaces have the following formal counterparts in the nonpolyhedral set Ω:

$$\mathcal{N}_\Omega(u) = \{w \in \mathsf{L}^\infty_m[0,1] : w(t) \overset{\text{a.e.}}{\in} \mathcal{N}_U(u(t))\}, \tag{7a}$$

$$\text{(7b)} \qquad \mathcal{T}_\Omega(u) = \{w \in \mathsf{L}^\infty_m[0,1] : w(t) \overset{\text{a.e.}}{\in} \mathcal{T}_U(u(t))\},$$

$$\text{(7c)} \qquad \mathsf{N}_\Omega(u) = \{w \in \mathsf{L}^\infty_m[0,1] : w(t) \overset{\text{a.e.}}{\in} \mathsf{N}_U(u(t))\},$$

$$\text{(7d)} \qquad \mathsf{T}_\Omega(u) = \{w \in \mathsf{L}^\infty_m[0,1] : w(t) \overset{\text{a.e.}}{\in} \mathsf{T}_U(u(t))\}.$$

It can be seen that $\mathcal{N}_\Omega(u)$ actually *is* the cone of exterior normals at u in Ω in the incomplete inner product space $\mathsf{V} = \{\mathsf{L}^\infty_m[0,1], \langle\cdot,\cdot\rangle_2, \|\cdot\|_2\}$. Similarly, $\mathcal{T}_\Omega(u)$ is the polar of $\mathcal{N}_\Omega(u)$ in this space, and $\mathsf{N}_\Omega(u)$ and $\mathsf{T}_\Omega(u)$ are, respectively, the closed linear hull and orthogonal complement of $\mathcal{N}_\Omega(u)$. Moreover, the cones $\mathcal{N}_\Omega(u)$ and $\mathcal{T}_\Omega(u)$ and subspaces $\mathsf{N}_\Omega(u)$ and $\mathsf{T}_\Omega(u)$ are *pointwise* polar and orthogonal, respectively; i.e., for all v and w,

$$v \in \mathcal{N}_\Omega(u) \quad \text{and} \quad w \in \mathcal{T}_\Omega(u) \Rightarrow \langle v(t), w(t)\rangle \overset{\text{a.e.}}{\leq} 0$$

and

$$v \in \mathsf{N}_\Omega(u) \quad \text{and} \quad w \in \mathsf{T}_\Omega(u) \Rightarrow \langle v(t), w(t)\rangle \overset{\text{a.e.}}{=} 0.$$

Although $\{\mathsf{L}^\infty_m[0,1], \langle\cdot,\cdot\rangle_2, \|\cdot\|_2\}$ is incomplete, it can also be seen that $\mathcal{N}_\Omega(u)$ is the polar cone for $\mathcal{T}_\Omega(u)$, $\mathsf{N}_\Omega(u)$ is the orthogonal complement of $\mathsf{T}_\Omega(u)$, and every vector v in $\mathsf{L}^\infty_m[0,1]$ has a unique orthogonal projection into each of these objects, with

$$v = P_{\mathcal{N}_\Omega(u)}\, v + P_{\mathcal{T}_\Omega(u)}\, v, \qquad v = P_{\mathsf{N}_\Omega(u)}\, v + P_{\mathsf{T}_\Omega(u)}\, v,$$

$$\left(P_{\mathcal{N}_\Omega(u)}\, v\right)(t) \overset{\text{a.e.}}{=} P_{\mathcal{N}_U(u(t))}\, v(t), \qquad \left(P_{\mathcal{T}_\Omega(u)}\, v\right)(t) \overset{\text{a.e.}}{=} P_{\mathcal{T}_U(u(t))}\, v(t),$$

and

$$\left(P_{\mathsf{N}_\Omega(u)}\, v\right)(t) \overset{\text{a.e.}}{=} P_{\mathsf{N}_U(u(t))}\, v(t), \qquad \left(P_{\mathsf{T}_\Omega(u)}\, v\right)(t) \overset{\text{a.e.}}{=} P_{\mathsf{T}_U(u(t))}\, v(t).$$

The necessary conditions in Theorem 4.1 make no reference to a norm on $\mathbb{R}^n$, and formal extensions of these conditions for the nonpolyhedral set Ω are now obtained by simply replacing $\mathsf{T}_C(u)$ with $\mathsf{T}_\Omega(u)$. The formal first-order necessary condition is well known and follows at once from the convexity of Ω and the definition of the directional derivative; however, the formal second-order necessary condition is a nontrivial assertion, since the subspace $\mathsf{T}_\Omega(u)$ is not contained in the cone of feasible directions at u (Example 5.1). When the degree-2 homogeneous function $w \to d^2J(u\ ;\ w,w)$ is L^2 continuous, the second-order necessary condition can be proved by demonstrating that $\mathsf{T}_\Omega(u)$ is the L^2 closure of the union of an increasing sequence of subspaces that *are* contained in the cone of feasible directions (Lemma 5.1 and Theorem 5.2). On the other hand, when $w \to d^2J(u\ ;\ w,w)$ is merely L^∞ continuous, the analogous proof construction fails since $\mathsf{T}_\Omega(u)$ contains vectors w that *cannot* be approximated by feasible direction vectors with arbitrarily small error in the norm $\|\cdot\|_\infty$ (Note 5.2). Thus, the norm-free second-order necessary condition for polyhedra in Theorem 4.1 does have a valid formal extension for Ω, but the extension is evidently norm-dependent and tied to $\|\cdot\|_2$ or other similar integral norms. Note that $w \to d^2J(u\ ;\ w,w)$ is automatically L^2 continuous when J satisfies the structure conditions (2).

Note 4.1. The existence of nontrivial subspaces in the cone of feasible directions at u in the nonpolyhedral set Ω is a special consequence of Ω's product structure. For nonpolyhedral closed convex sets prescribed by finitely many smooth nonlinear inequalities or an infinitely indexed family of affine inequalities, the cone of feasible directions at a boundary point u typically contains *no* subspace other than $\{0\}$; moreover, for nonconvex J, the directional derivative $d^2J(u\ ;\ v,v)$ is typically *negative* for some v in the orthogonal complement of the normal cone at a local minimizer u. For example, in any Hilbert space V, the closed unit ball has the affine inequality constraint representation

$$\overline{B(0,1)} = \{u \in \mathsf{V} : \forall i \in S(0,1) \quad \langle i, u\rangle \leq 1\},$$

where $S(0,1)$ is the unit *sphere* in V. If $\|u\| = 1$, then the normal cone at u consists of all nonnegative multiples of u and there are no nonzero feasible direction vectors v orthogonal to u. Furthermore, if J is twice directionally differentiable and *strictly concave*, then $d^2J(u\ ;\ v,v)$ is negative for all $v \neq 0$ orthogonal to u.

The sufficient conditions in Theorem 4.2 are directly tied to the norm on $\mathbb{R}^n$ in three ways, through the differentiability, strict complementarity, and coercivity hypotheses. Since norm equivalence is lost in the infinite-dimensional vector space $\mathsf{L}^\infty_m[0,1]$, it can be seen that nonequivalent formal extensions of the sufficient conditions in Ω may arise from the assignment of different combinations of norms in these hypotheses. Some of these extensions are essentially vacuous and can be ruled out at once. In particular, any formal extension that invokes an L^2 strict complementarity condition or an L^∞ coercivity condition is uninteresting for present purposes, since the L^2 relative interior of the normal cone $\mathcal{N}_\Omega(u)$ is typically *empty*, and the L^∞ coercivity condition can't hold on nontrivial subspaces $\mathsf{T}_\Omega(u)$ if $w \longrightarrow d^2J(u\ ;\ w,w)$ is continuous in the L^2 norm. Nonvacuous sufficient conditions for L^∞-local optimality in Ω *are* obtained by combining an L^∞ strict complementarity condition with an L^2 coercivity condition and any smoothness conditions that imply the two-norm Taylor formula (4) with $\nu = \infty$ (Theorem 6.1); however, the gap between these sufficient conditions and the corresponding necessary conditions in Theorem 5.2 is far wider than the sufficiency gap in polyhedral subsets of $\mathbb{R}^n$. This is true partly because coercivity is stronger than positive-definiteness in infinite-dimensional spaces, but mainly because L^∞ strict complementarity requires that the distance from $-\nabla J(u)(t)$ to the relative boundary of $\mathcal{N}_U(u(t))$ in the subspace $\mathsf{N}_U(u(t))$ is *bounded away from* 0 *almost everywhere on the interval* $[0,1]$, and this condition typically can't be met if $u(\cdot)$ and $\nabla J(u)(\cdot)$ are continuous at some point τ in $[0,1]$ where $u(t)$ passes from the relative interior of one face of the polyhedron U to the relative interior of a contiguous face. Thus, L^∞ strict complementarity is a nonvacuous but very stringent hypothesis, and the aim of the sufficiency analysis in §6 is to weaken this condition *with no compensatory reinforcement of the* L^2 *coercivity condition on* $\mathsf{T}_\Omega(u)$. Theorems 6.4 and 6.6 achieve this goal with a mild strengthening of the *pointwise strict complementarity condition*

$$-\nabla J(u)(t) \overset{\text{a.e.}}{\in} \operatorname{ri} \mathcal{N}_U(u(t)), \tag{8}$$

which may be viewed as a formal extension of the component-wise expression of strict complementarity in k-fold Cartesian products $\Omega_k = U \times \cdots \times U$ in $\mathbb{R}^{km}$, i.e.,

$$-\left(\nabla J(u)\right)_i \in \operatorname{ri} \mathcal{N}_U(u_i), \quad i = 1, \ldots, k$$

(cf. equations (6)). Reference [1] supplies the prototype for this analysis in the special case $U = [0, \infty)$.

5. Necessary conditions. The special product structure in Ω ensures that subspaces $\mathsf{T}_\Omega(u)$ (analogues of the lineality) can be approximated from below by an increasing sequence of subspaces $T_{(1/n)}$, each of which is contained in the cone of feasible directions, even though $\mathsf{T}_\Omega(u)$ itself is *not* contained in this cone. It follows that each vector w in $\mathsf{T}_\Omega(u)$ is the L^2 limit of a sequence of feasible direction vectors w^n for which $d^1 f(u; w^n) = 0$ and $d^2 f(u; w^n, w^n) \geq 0$. Hence, *for this specially structured nonpolyhedral set* Ω, it *is* possible to extend the inequality $d^2 f(u; w, w) \geq 0$ from the subspaces $T_{(1/n)}$ to $\mathsf{T}_\Omega(u)$ when the directional derivative map $w \to d^2 J(u\ ;\ w, w)$ is L^2 continuous. These points are developed in the following example, lemma, and theorem.

Example 5.1. For $m = 1$ and $U = [0, \infty)$, Ω is the set of real-valued essentially bounded and nonnegative measurable functions on $[0, 1]$. Consider the function u in Ω defined by the rule

$$u(t) = \begin{cases} 0, & t \in [0, \frac{1}{2}], \\ t - \frac{1}{2}, & t \in (\frac{1}{2}, 1]. \end{cases}$$

Note that

$$\mathcal{N}_U(\xi) = \begin{cases} (-\infty, 0], & \xi = 0, \\ \{0\}, & \xi > 0, \end{cases}$$

$$[\text{span}\,\mathcal{N}_U(\xi)]^\perp = \begin{cases} \{0\}, & \xi = 0, \\ \mathbb{R}^1, & \xi > 0, \end{cases}$$

and therefore

$$\mathcal{N}_\Omega(u) = \left\{w \in \mathsf{L}^\infty(0,1) : w(t) \leq 0 \text{ a.e. in } [0, \tfrac{1}{2}] \text{ and } w(t) = 0 \text{ a.e. in } (\tfrac{1}{2}, 1]\right\},$$

$$\mathsf{T}_\Omega(u) = \{w \in \mathsf{L}^\infty(0,1) : w(t) = 0 \text{ a.e. in } [0, \tfrac{1}{2}]\}.$$

Fix $h > 0$ and put

$$w_h(t) = \begin{cases} 0, & t \in [0, \frac{1}{2}], \\ -h, & t \in (\frac{1}{2}, 1]. \end{cases}$$

The function w_h lies in $\mathsf{T}_\Omega(u)$ but is not a feasible direction vector, since $u + s w_h$ has negative values on the interval $[\frac{1}{2}, \frac{1}{2} + sh)$. Thus $\mathsf{T}_\Omega(u)$ is not contained in the cone of feasible directions at u. On the other hand, the subspaces

$$\mathsf{T}_{(1/n)}(u) = \{w \in \mathsf{L}^\infty(0,1) : w(t) = 0 \text{ a.e. in } [0, \tfrac{1}{2} + \tfrac{1}{n}]\}$$

are contained in $\mathsf{T}_\Omega(u)$ and the cone of feasible directions at u, and converge to $\mathsf{T}_\Omega(u)$ in the sense that their union is dense in $\mathsf{T}_\Omega(u)$, relative to the L^2 norm.

The construction in Example 5.1 will now be generalized to show that the subspaces $\mathsf{T}_\Omega(u)$ are approximated from below in the L^2 norm by subspaces of $\mathsf{T}_\Omega(u)$ lying in the cone of feasible directions at $u \in \Omega$. With reference to [15], the polyhedral convex set U has d distinct (polyhedral) *faces* $\mathcal{F}_i$, the *relative interiors* of these faces do not meet, and

$$U = \bigcup_{i=1}^{d} \text{ri}\mathcal{F}_i. \tag{9}$$

Moreover, if $\mathcal{F}(\xi)$ denotes the unique face in $\{\mathcal{F}_1, \ldots, \mathcal{F}_d\}$ containing ξ in its relative interior, then the cones $\mathcal{N}_U(\eta)$ and the subspaces $\mathsf{N}_U(\eta)$ and $\mathsf{T}_U(\eta)$ are invariant for $\eta \in \mathrm{ri}\mathcal{F}(\xi)$, and $\xi + \mathsf{T}_U(\xi)$ is the *affine hull* of $\mathcal{F}(\xi)$; i.e., for some $\{\mathcal{N}_i\}_{i=1}^d$, $\{\mathsf{N}_i\}_{i=1}^d$, and $\{\mathsf{T}_i\}_{i=1}^d$ and for all $\xi \in U$ and $i = 1, \ldots, d$,

(10)

$$(\mathcal{F}(\xi) = \mathcal{F}_i) \Rightarrow (\mathcal{N}_U(\xi) = \mathcal{N}_i,\ \mathsf{N}_U(\xi) = \mathsf{N}_i,\ \mathsf{T}_U(\xi) = \mathsf{T}_i, \text{ and } \mathrm{aff}\, \mathcal{F}(\xi) = \xi + \mathsf{T}_i\ ).$$

(These assertions are deduced in [15] from affine inequality representations for the polyhedral set U and derived representations for the polyhedral cones $\mathcal{N}_U(\xi)$; for further details, see the proof of Lemma 6.5.) For ξ in U, define

$$\rho(\xi) = \sup\{\rho \in (0,1] : B(\xi, \rho) \cap \mathrm{aff}\, \mathcal{F}(\xi) \subset \mathcal{F}(\xi)\}, \tag{11a}$$

where

$$B(\xi, \rho) = \{\eta \in \mathbb{R}^m : \|\xi - \eta\| < \rho\}. \tag{11b}$$

Note that $\rho(\xi)$ is the distance from ξ to the relative boundary of the face $\mathcal{F}(\xi)$ when $\rho(\xi) < 1$. For ξ in U, u in Ω, and $\sigma > 0$, define the corresponding subspaces $\mathsf{T}_{(\sigma)}(u) \subset \mathsf{T}_\Omega(u)$ by the rule

$$\mathsf{T}_{(\sigma)}(u) = \{w \in \mathsf{L}_m^\infty(0,1) : w(t) \overset{\text{a.e.}}{\in} \mathsf{T}_{(\sigma)}(u)(t)\}, \tag{11c}$$

$$\mathsf{T}_{(\sigma)}(u)(t) = \begin{cases} \mathsf{T}_U(u(t)), & \rho\,(u(t)) > \sigma, \\ \{0\}, & \rho\,(u(t)) \le \sigma. \end{cases} \tag{11d}$$

It can be seen that $\rho(\xi)$ is *positive* and the associated restrictions $\rho(\cdot)|_{\mathrm{ri}\mathcal{F}_i}$ are *continuous* for $i = 1, \ldots, d$. Moreover, $\mathrm{ri}\mathcal{F}_i$ is a Borel set for $i = 1, \ldots, d$, and thus for each u in Ω, the partition (9) induces a corresponding family of d pairwise disjoint Lebesgue measurable sets

$$\alpha_i(u) = \{t \in [0,1] : u(t) \in \mathrm{ri}\mathcal{F}_i\} \tag{12a}$$

with

$$\mu\left[[0,1] \setminus \bigcup_{i=1}^d \alpha_i(u)\right] = 0. \tag{12b}$$

The continuity properties of $\rho(\cdot)$ therefore imply that for each u in Ω, the composite function $\rho\,(u(\cdot))$ is *Lebesgue measurable and positive almost everywhere in* $[0,1]$.

LEMMA 5.1. *For all u in Ω and $\sigma > 0$, the subspace $\mathsf{T}_{(\sigma)}(u)$ lies in the cone of feasible directions at u, and the subspace $\bigcup_{n=1}^\infty \mathsf{T}_{(1/n)}(u)$ is dense in $\mathsf{T}_\Omega(u)$ relative to the L^2 norm.*

Proof. Fix u in Ω and $\sigma > 0$. Suppose that $w \in \mathsf{T}_{(\sigma)}(u)$. If $w = 0$, then w is trivially a feasible direction vector at u. Suppose that $w \neq 0$ and $s \in [0, \frac{\sigma}{\|w\|_\infty}]$. By construction,

$$u(t) + sw(t) \overset{\text{a.e.}}{\in} \mathcal{F}\,(u(t)) \subset U.$$

Thus, for all s,

$$s \in [0, \frac{\sigma}{\|w\|_\infty}] \Rightarrow u + sw \in \Omega,$$

and therefore w is a feasible direction vector at u.

Fix u in Ω and w in $\mathsf{T}_\Omega(u)$. For $n = 1, 2, \ldots$, let

$$\theta_n = \{t \in [0,1] : \rho(u(t)) \leq \tfrac{1}{n}\},$$

and construct

$$w_n(t) = \begin{cases} w(t), & t \in [0,1] \setminus \theta_n, \\ 0, & t \in \theta_n. \end{cases}$$

Since $\rho(u(\cdot))$ is measurable and positive almost everywhere, it follows that θ_n is measurable with $\theta_n \supset \theta_{n+1}$ for all n, and

$$\lim_{n\to\infty} \mu[\theta_n] = \mu\left[\bigcap_{n=1}^{\infty} \theta_n\right] = 0.$$

Hence, $w_n \in \mathsf{T}_{(1/n)}(u)$ and

$$\lim_{n\to\infty} \|w_n - w\|_2 = \lim_{n\to\infty} \left(\int_{\theta_n} \|w\|^2 dt\right)^{\frac{1}{2}} = 0. \qquad \square$$

Note 5.2. If u and w_h are defined as in Example 5.1, then w_h *cannot* be approximated by feasible direction vectors with arbitrarily small error in the norm $\|\cdot\|_\infty$. Hence, there is no L^∞ counterpart of Lemma 5.1.

Lemma 5.1 and elementary calculus now produce a basic representation-free second-order necessary condition in the inner product space $\{\mathsf{L}^\infty_m[0,1], \langle\cdot,\cdot\rangle_2, \|\cdot\|_2\}$. Note that the structure/continuity conditions (2)–(3) are not invoked in this theorem.

THEOREM 5.2. *Let u be a local minimizer for the restriction of $J : \mathcal{D} \to \mathbb{R}^1$ to any line in Ω. In addition, suppose that J is twice directionally differentiable and that the associated maps $w \to d^1J(u;w)$ and $w \to d^2J(u;w,w)$ are continuous with respect to the L^2 norm on $\mathsf{L}^\infty_m(0,1)$. Then*

$$\forall v \in \Omega \quad d^1J(u; v-u) \geq 0, \tag{13a}$$

$$\forall w \in \mathsf{T}_\Omega(u) \quad d^1J(u;w) = 0 \quad \text{and} \quad d^2J(u;w,w) \geq 0. \tag{13b}$$

Proof. Fix v in Ω. Since Ω is convex and u is an internal point of $\mathcal{D}$ and a local minimizer on lines, it follows that for all sufficiently small ϵ, the vectors $u + \epsilon(v-u)$ fall in $\Omega \cap \mathcal{D}$ and therefore

$$0 \leq J(u + \epsilon(v-u)) - J(u) = \epsilon d^1J(u; v-u) + o(\epsilon).$$

In the limit as $\epsilon \to 0$, this proves the well-known first-order necessary condition (13a). Suppose that w lies in one of the subspaces $\mathsf{T}_{(1/n)}(u)$ of Lemma 5.1. Then w and $-w$ are feasible direction vectors at u and hence $d^1J(u;w) = 0$ and $d^2J(u;w,w) \geq 0$, since u is a local minimizer on lines. Conditions (13b) now follow from the second part of Lemma 5.1 and the L^2 continuity of $w \to d^1J(u;w)$ and $w \to d^2J(u;w,w)$. $\square$

The next two theorems establish refinements of the second-order necessary conditions for twice Gâteaux differentiable functions that satisfy some or all of the structure/continuity hypotheses (2)–(3) for $\nu = 1$ or 2. The proofs for these results are closely modelled on the proofs for Theorems 2 and 3 in [1] and convert integral conditions to pointwise conditions by arguments that are well established in optimal control theory [18].

THEOREM 5.3. *Let u be a local minimizer for the restriction of $J : \mathcal{D} \to \mathbb{R}^1$ to any line in Ω, and suppose that J has first and second differentials satisfying* (2). *Then conditions* (13) *hold at u, and in addition,*

$$-\nabla J(u)(t) \in \mathcal{N}_U(u(t)) \quad a.e.\ in\ [0,1] \tag{14a}$$

and

$$(\forall \xi \in \mathsf{T}_U(u(t)), \quad \langle \xi, S(u)(t)\xi\rangle \geq 0)) \quad a.e.\ in\ [0,1]. \tag{14b}$$

Proof. Conditions (2) imply that $d^1J(u;w)$ and $d^2J(u;w,w)$ are L^2-continuous in w on $\mathsf{L}^\infty_m(0,1)$. By Theorem 5.2, conditions (13) then hold at u, and (13a) asserts that

$$\forall u \in \Omega \quad \int_0^1 \langle \nabla J(u)(t), v(t) - u(t)\rangle dt \geq 0$$

and hence

$$(\forall \xi \in U \quad \langle \nabla J(u)(t), \xi - u(t)\rangle \geq 0) \quad \text{a.e. in } [0,1].$$

This proves (14a). Now let $\{\alpha_i(u)\}_{i=1}^d$ be the subsets of $[0,1]$ in (12), and for t in $[0,1]$ and $\epsilon > 0$ define a corresponding family of sets $\phi(t,\epsilon) \subset [0,1]$ by the rule

$$t \in \alpha_i(u) \Rightarrow \phi(t,\epsilon) = \alpha_i(u) \cap (t-\epsilon, t+\epsilon). \tag{15}$$

Evidently, $\mu[\phi(t,\epsilon)] \leq 2\epsilon$, and therefore

$$\mu[\phi(t,\epsilon)] \cdot \int\int_{\phi(t,\epsilon)\times\phi(t,\epsilon)} \|K(u)(\tau,s)\|^2 d\tau ds = o(\epsilon). \tag{16a}$$

Moreover, since almost every t in $\alpha_i(u)$ is a point of density for $\alpha_i(u)$ [24], it follows that

$$\int_{\phi(t,\epsilon)} S(u)(\tau)d\tau = 2\epsilon S(u)(t) + o(\epsilon) \quad \text{a.e. in } [0,1]. \tag{16b}$$

Choose any t where the estimates in (16) hold. Suppose that $t \in \alpha_i(u)$. Fix ξ in $\mathsf{T}_i = \mathsf{T}_U(u(t))$, and for $\epsilon > 0$ construct $w_\epsilon \in \mathsf{T}_\Omega(u)$ by the rule

$$w_\epsilon(\tau) = \begin{cases} \xi, & \tau \in \phi(t,\epsilon), \\ 0, & \tau \in [0,1] \setminus \phi(t,\epsilon). \end{cases}$$

Theorem 5.2, conditions (2), and the estimates (16) then yield

$$0 \leq d^2J(u; w_\epsilon, w_\epsilon) = 2\epsilon\langle \xi, S(u)(t)\xi\rangle + o(\epsilon)$$

and hence

$$\langle \xi, S(u)(t)\xi\rangle \geq 0$$

in the limit as $\epsilon \to 0$. Since ξ can be any vector in $\mathsf{T}_i = \mathsf{T}_U(u(t))$, this proves (14b).
□

THEOREM 5.4. *Let u be an L^2-local minimizer of $J : \mathcal{D} \to \mathbb{R}^1$ in $\Omega \cap \mathcal{D}$, and suppose that J has first and second Gâteaux differentials satisfying* (2)–(3) *with $\nu = 2$. Then conditions* (13) *and* (14) *hold at u, and in addition,*

$$(17)\quad \inf_{\xi \in U} \left(\langle \nabla J(u)(t), \xi - u(t) \rangle + \tfrac{1}{2} \langle \xi - u(t), S(u)(t)\,(\xi - u(t)) \rangle \right) = 0 \quad \textit{a.e. in } [0,1].$$

Proof. Conditions (13) and (14) follow at once from Theorems 5.2 and 5.3, since every L^2-local minimizer is also a local minimizer on lines. To prove (17), fix t in $[0,1]$ and ξ in U, and for $\epsilon > 0$ construct $v_{t,\xi,\epsilon} \in \Omega$ by the rule

$$(18)\qquad v_{t,\xi,\epsilon}(\tau) = \begin{cases} \xi, & \tau \in [0,1] \cap (t-\epsilon, t+\epsilon), \\ u(\tau), & \tau \in [0,1] \setminus (t-\epsilon, t+\epsilon). \end{cases}$$

As in the proof of Theorem 5.3, it can be seen that for almost all t in $[0,1]$,

$$\forall \xi \in U \quad \|v_{t,\xi,\epsilon} - u\|_2^2 = 2\epsilon \|\xi - u(t)\|^2 + o(\epsilon),$$

and therefore

$$\forall \xi \in U \quad \|v_{t,\xi,\epsilon} - u\|_2 = O(\epsilon^{\frac{1}{2}}).$$

In the present instance, conditions (2)–(3) imply (4) with $\nu = 2$; hence for almost all t in $[0,1]$, all ξ in U, and sufficiently small $\epsilon > 0$,

$$\begin{aligned} 0 &\le J(v_{t,\xi,\epsilon}) - J(u) \\ &= \langle \nabla J(u), v_{t,\xi,\epsilon} - u \rangle_2 + \tfrac{1}{2} \langle v_{t,\xi,\epsilon} - u, \nabla^2 J(u)\,(v_{t,\xi,\epsilon} - u) \rangle_2 + o(\epsilon) \end{aligned}$$

since u is an L^2-local minimizer. Moreover, for almost all t in $[0,1]$,

$$\forall \xi \in U \quad \langle \nabla J(u), v_{t,\xi,\epsilon} - u \rangle_2 = 2\epsilon \langle \nabla J(u)(t), \xi - u(t) \rangle + o(\epsilon)$$

and

$$\forall \xi \in U \quad \langle v_{t,\xi,\epsilon} - u, \nabla^2 J(u)\,(v_{t,\xi,\epsilon} - u) \rangle_2 = 2\epsilon \langle \xi - u(t), S(u)(t)\,(\xi - u(t)) \rangle + o(\epsilon).$$

Condition (17) now follows in the limit as $\epsilon \to 0$. □

Note 5.3. Assertion (17) in Theorem 5.4 remains true if the range set U in (1b) is an *arbitrary* subset of $\mathbb{R}^m$. Condition (14a) in Theorem 5.3 is implied by (17) if U is a convex set in $\mathbb{R}^m$. Condition (14b) is implied by (17) if U is a polyhedral convex set in $\mathbb{R}^m$.

Note 5.4. For Bolza objective functions (5), the values of the quantities $\nabla J(u)$ and $S(u)$ are given almost everywhere in $[0,1]$ by

$$(19a)\qquad \nabla J(u)(t) = \nabla_u H\,(t, \psi(t), x(t), u(t))$$

and

$$(19b)\qquad S(u)(t) = \nabla^2_{uu} H\,(t, \psi(t), x(t), u(t)),$$

where H is the associated Hamiltonian

$$(20a)\qquad H(t, \psi, x, u) = \langle \psi, f(t,x,u) \rangle + f^0(t,x,u)$$

and $\psi(\cdot)$ solves the adjoint final value problem

$$\frac{d\psi}{dt} = -\nabla_x H\left(t, \psi, x(t), u(t)\right), \tag{20b}$$

$$\psi(1) = \nabla P(x(1)). \tag{20c}$$

(See [13] for a discussion of the case m=1.) If u is an L^2-local minimizer of (5) in the set (1b), then admissible perturbations of the form (18) cannot decrease J when ϵ is sufficiently small, and the Pontryagin minimum principle asserts that

$$H\left(t, \psi(t), x(t), u(t)\right) = \inf_{\xi \in U} H\left(t, \psi(t), x(t), \xi\right). \tag{21}$$

In this setting, conditions (14) in Theorem 5.3 are now seen as first- and second-order necessary conditions for the finite-dimensional constrained minimization problems (21); moreover, if (2) and (3) are to hold with $\nu = 2$ for control problems, then H must be quadratic in u for each fixed (t, ψ, x) (Note 3.1), and condition (17) in Theorem 5.4 is *equivalent* to (21). On the other hand, the Pontryagin minimum principle need not hold at a local minimizer on lines, where (14a) and (14b) may be viewed as extensions of the Euler and Legendre necessary conditions for weak local minimizers in the calculus of variations.

Note 5.5. Theorems 5.2 and 5.3 contain their counterparts in [1]; in particular, for $m = 1$ and $U = [0, \infty)$, condition (17) implies that $S(u)(t) \geq 0$ almost everywhere in $[0, 1]$.

6. Sufficient conditions. A nonvacuous but stringent formal extension of the sufficient conditions in polyhedra is obtained by replacing $\mathcal{N}_C(u)$ and $\mathsf{T}_C(u)$ in Theorem 4.2 with $\mathcal{N}_\Omega(u)$ and $\mathsf{T}_\Omega(u)$ and requiring that Taylor's formula (4) hold with $\nu = \infty$ at the point u in $\Omega \cap \mathcal{D}$, that $-\nabla J(u)$ lie in the L^∞ relative interior of $\mathcal{N}_\Omega(u)$, and that $\nabla^2 J(u)$ be L^2-*coercive* on $\mathsf{T}_\Omega(u)$. As noted in §4, the scope of this theorem is severely limited by its L^∞ strict complementarity condition; however, the proof technique employed here also establishes the more flexible analytical tool in Lemma 6.2. Lemma 6.2 is the key to the proof of the L^∞-local optimality sufficient conditions in Theorem 6.4 for objective functions J that satisfy the structure/continuity conditions (2)–(3) with $\nu = \infty$. Theorem 6.4 postulates the natural L^2 coercivity condition on $\mathsf{T}_\Omega(u)$, but its strict complementarity condition derives from the pointwise condition (8) and is much weaker than L^∞ strict complementarity.

THEOREM 6.1. *Let J be a real-valued function on a domain $\mathcal{D} \subset \mathsf{L}^\infty_m[0,1]$ that is open relative to the norm $\|\cdot\|_\infty$. Suppose that at some point u in $\Omega \cap \mathcal{D}$ there is a vector $\nabla J(u)$ in $\mathsf{L}^\infty_m(0,1)$ and an L^2 continuous linear operator $\nabla^2 J(u)$ on $\mathsf{L}^\infty_m(0,1)$ such that*

$$d^1 J(u; v) = \langle \nabla J(u), v \rangle_2, \tag{22a}$$

$$d^2 J(u; v, w) = \langle v, \nabla^2 J(u) w \rangle_2 \tag{22b}$$

for all v and w in $\mathsf{L}^\infty_m[0,1]$, and

$$J(v) - J(u) = \langle \nabla J(u), v - u \rangle_2 + \tfrac{1}{2}\langle v - u, \nabla^2 J(u)\,(v - u) \rangle_2 + r(u; v), \tag{23a}$$

$$\lim_{\|v-u\|_\infty \to 0} \frac{|r(u;v)|}{\|v-u\|_2^2} = 0 \tag{23b}$$

for all v in $\mathcal{D}$ near u. Suppose that the following L^∞ strict complementarity and L^2 coercivity conditions also hold at u:

$$\exists c_N > 0 \,\forall w \quad (w \in \mathsf{N}_\Omega(u) \textit{ and } \|w\|_\infty \le c_N \Rightarrow -\nabla J(u) + w \in \mathcal{N}_\Omega(u)) \tag{24a}$$

and

$$\exists c_T > 0 \,\forall w \quad \left(w \in \mathsf{T}_\Omega(u) \Rightarrow \langle w, \nabla^2 J(u) w\rangle_2 \ge c_T \|w\|_2^2\right). \tag{24b}$$

Then for each number $c_\infty \in (0, c_T)$, there is a corresponding $\delta_\infty > 0$ such that for all v,

$$v \in \Omega \textit{ and } \|v-u\|_\infty < \delta_\infty \Rightarrow J(v) - J(u) \ge \tfrac{1}{2} c_\infty \|v-u\|_2^2. \tag{25}$$

Proof. Recall first that the subspaces $\mathsf{N}_\Omega(u)$ and $\mathsf{T}_\Omega(u)$ are pointwise orthogonal and that $\mathsf{L}_m^\infty(0,1)$ is the *direct sum* of $\mathsf{N}_\Omega(u)$ and $\mathsf{T}_\Omega(u)$. To prove the latter assertion, fix w in $\mathsf{L}_m^\infty(0,1)$, and for t in $[0,1]$ put

$$w_N(t) = P_{\mathsf{N}_U(u(t))} w(t)$$

and

$$w_T(t) = P_{\mathsf{T}_U(u(t))} w(t).$$

Since the projection operators are nonexpansive and $\mathsf{N}_U(u(t))$ and $\mathsf{T}_U(u)(t))$ are constant almost everywhere on each of the measurable sets $\alpha_i(u)$ in (12), it follows that w_N and w_T are measurable essentially bounded functions in $\mathsf{N}_\Omega(u)$ and $\mathsf{T}_\Omega(u)$, with

$$w = w_N + w_T \tag{26a}$$

and

$$\|w(t)\|^2 = \|w_N(t)\|^2 + \|w_T(t)\|^2 \quad \text{a.e. in } [0,1]. \tag{26b}$$

Moreover, if $w = v_N + v_T$ for some $v_N \in \mathsf{N}_\Omega(u)$ and $v_T \in \mathsf{T}_\Omega(u)$, then the orthogonality of $\mathsf{N}_\Omega(u)$ and $\mathsf{T}_\Omega(u)$ ensures that $v_N = w_N$ and $v_T = w_T$.

Now fix c_∞ in $(0, c_T)$ and v in $\Omega \setminus \{u\}$, and let

$$w = c_N \frac{(v-u)_N}{\|v-u\|_\infty}.$$

With reference to (26), it can be seen that

$$\|w\|_\infty = \frac{c_N}{\|v-u\|_\infty} \|(v-u)_N\|_\infty \le c_N.$$

Hence (24a) implies that

$$\begin{aligned} 0 &\ge \langle -\nabla J(u) + w, v-u\rangle_2 \\ &= -\langle \nabla J(u), v-u\rangle_2 + \frac{c_N}{\|v-u\|_\infty} \langle (v-u)_N, v-u\rangle_2, \end{aligned}$$

and therefore

$$\forall v \in \Omega \quad \langle \nabla J(u), v-u\rangle_2 \geq \frac{c_N}{\|v-u\|_\infty}\|(v-u)_N\|_2^2. \tag{27}$$

Furthermore, conditions (23) and (24b) yield

(28a)

$$\langle v-u, \nabla^2 J(u)\,(v-u)\rangle_2 \geq c_T\|(v-u)_T\|_2^2 - M\|(v-u)_N\|_2\,(\|(v-u)_N\|_2 + 2\|(v-u)_T\|_2)$$

where

$$M = \|\nabla^2 J(u)\| \overset{\text{def}}{=} \sup_{\|v\|_2=1} \|\nabla^2 J(u)v\|_2 < \infty. \tag{28b}$$

The estimates (23), (27), and (28) and a completion of the square in (28) show that for all v in $\Omega \cap \mathcal{D}$,

(29)

$$J(v) - J(u)$$
$$\geq \left(\frac{c_N}{\|v-u\|_\infty} - \frac{M}{2} - \frac{M^2}{(c_T - c_\infty)}\right)\|(v-u)_N\|_2^2 + \frac{(c_T + c_\infty)}{4}\|(v-u)_T\|_2^2 + r(u;v)$$

(cf. the proof of Lemma 1 in [1]). Then assertion (25) holds if δ_∞ is chosen so small that

$$c_N\delta_\infty^{-1} - \frac{M}{2} - \frac{M^2}{(c_T - c_\infty)} \geq \tfrac{1}{4}(c_T + c_\infty)$$

and

$$0 < \|v-u\|_\infty < \delta_\infty \Rightarrow v \in \mathcal{D} \text{ and } \frac{|r(u;v)|}{\|v-u\|_2^2} \leq \tfrac{1}{4}(c_T - c_\infty). \qquad \square$$

The foregoing proof works equally well if the subspaces $\mathsf{N}_\Omega(u)$ and $\mathsf{T}_\Omega(u)$ in Theorem 6.1 are replaced by *any* complementary pointwise orthogonal subspaces $\hat{\mathsf{N}}(u)$ and $\hat{\mathsf{T}}(u)$ in $\mathsf{L}_m^\infty[0,1]$, with $\hat{\mathsf{N}}(u) \subset \mathsf{N}_\Omega(u)$ and $\hat{\mathsf{T}}(u) \supset \mathsf{T}_\Omega(u)$. This observation immediately produces the following extension of Theorem 6.1.

LEMMA 6.2. *Let J be a real-valued function on a domain $\mathcal{D} \subset \mathsf{L}_m^\infty[0,1]$ that is open relative to the norm $\|\cdot\|_\infty$, and suppose that hypotheses* (22) *and* (23) *in Theorem* 6.1 *are satisfied. Let $\hat{\mathsf{N}}(u)$ and $\hat{\mathsf{T}}(u)$ be pointwise orthogonal subspaces in $\mathsf{L}_m^\infty(0,1)$ such that $\hat{\mathsf{N}}(u) \subset \mathsf{N}_\Omega(u)$, $\hat{\mathsf{T}}(u) \supset \mathsf{T}_\Omega(u)$, and $\mathsf{L}_m^\infty(0,1)$ is the direct sum of $\hat{\mathsf{N}}(u)$ and $\hat{\mathsf{T}}(u)$. Assume that*

$$\exists c_{\hat{N}} > 0\; \forall w \quad \left(w \in \hat{\mathsf{N}}(u) \text{ and } \|w\|_\infty \leq c_{\hat{N}} \Rightarrow -\nabla J(u) + w \in \mathcal{N}_\Omega(u)\right) \tag{30a}$$

and

$$\exists c_{\hat{T}} > 0\; \forall w \quad \left(w \in \hat{\mathsf{T}}(u) \Rightarrow \langle w, \nabla^2 J(u)w\rangle_2 \geq c_{\hat{T}}\|w\|_2^2\right). \tag{30b}$$

Then for each number $c_\infty \in (0, c_{\hat{T}})$, there is a corresponding $\delta_\infty > 0$ such that for all v,

$$v \in \Omega \text{ and } \|v-u\|_\infty < \delta_\infty \Rightarrow J(v) - J(u) \geq \frac{1}{2}c_\infty\|v-u\|_2^2. \tag{31}$$

In general, if $\hat{\mathsf{N}}(u)$ is a proper subspace of $\mathsf{N}_\Omega(u)$, then the strict complementarity condition (30a) is *weaker* than (24a), and the L^2 coercivity condition (30b) is correspondingly *stronger* than (24b). However, when J satisfies the structure/continuity conditions (2)–(3) and u meets certain additional regularity conditions, it is possible to construct $\hat{\mathsf{N}}(u)$ and $\hat{\mathsf{T}}(u)$ so that (30a) is implied by a variant of the pointwise strict complementarity condition (8), while (30b) is obtained at no cost by a continuous extension of (24b); in other words, the L^∞ strict complementarity condition can be weakened substantially *with no corresponding reinforcement of the* L^2 *coercivity condition*. The relevant formal strict complementarity and coercivity conditions are collected in the following equations:

$$-\nabla J(u)(t) \in \mathcal{N}_U\,(u(t)) \quad \text{a.e. in } [0,1], \tag{32a}$$

$$\forall \beta \subset \cup_{i=1}^{d} \operatorname{int} \alpha_i(u) \quad (\beta \text{ compact} \Rightarrow \exists c_\beta > 0\ \Delta(u)(t) \geq c_\beta \text{ a.e. in } \beta)\,, \tag{32b}$$

$$\exists c_T > 0\ \forall w \quad \left(w \in \mathsf{T}_\Omega(u) \Rightarrow \langle w, \nabla^2 J(u) w\rangle_2 \geq c_T \|w\|_2^2\right), \tag{32c}$$

where $\Delta(u)(t)$ is the distance from $-\nabla J(u)(t)$ to the relative boundary of $\mathcal{N}_U\,(u(t))$ in the subspace $\mathsf{N}_U(u(t))$ or, equivalently,

$$\Delta(u)(t) = \sup\{\Delta \geq 0 : \forall \xi \in \mathsf{N}_U(u(t)),\ \|\xi\| \leq \Delta \Rightarrow -\nabla J(u)(t) + \xi \in \mathcal{N}_U\,(u(t))\}. \tag{32d}$$

While the strict complementarity conditions (30a)–(30b) are stronger than the pointwise strict complementarity condition (8), they still permit

$$\lim_{t \to \tau} \operatorname{ess\,inf}\ \Delta(u)(t) = 0 \tag{33}$$

at τ in the frontier of $\alpha_i(u)$ and are therefore substantially weaker than (24a). When (33) happens at frontier points of $\alpha_i(u)$, example 1 of [1] shows that (32) alone is in fact not sufficient for L^∞-local optimality in Ω, even for functionals J with differentials satisfying (2)–(3). On the other hand, theorem 4 in [1] shows that (32) may *become* sufficient for L^∞-local optimality when certain mild restrictions are imposed on the sets $\alpha_i(u)$ and the behavior of the function $S(u)(\cdot)$ near the frontiers of $\alpha_i(u)$. The proof construction in [1] applies a restricted version of Lemma 6.2 for $U = [0,1]$ to pointwise orthogonal subspaces

$$\hat{\mathsf{N}}(u) = \{w \in \mathsf{L}_m^\infty[0,1] : w(t) \overset{\text{a.e.}}{\in} \hat{\mathsf{N}}(u)(t)\}, \tag{34a}$$

$$\hat{\mathsf{T}}(u) = \{w \in \mathsf{L}_m^\infty[0,1] : w(t) \overset{\text{a.e.}}{\in} \hat{\mathsf{T}}(u)(t)\}, \tag{34b}$$

where $\hat{\mathsf{N}}(u)(t) \subset \mathsf{N}_U(u(t))$, $\hat{\mathsf{T}}(u)(t) \supset \mathsf{T}_U(u(t))$, and $\hat{\mathsf{N}}(u)(t)$ and $\hat{\mathsf{T}}(u)(t)$ are essentially constant on small neighborhoods of the frontier points of $\alpha_i(u)$, with range in the collection of subspaces $\{\mathsf{N}_i\}_{i=1}^d$ and $\{\mathsf{T}_i\}_{i=1}^d$ in (10). The same basic proof strategy used in [1] also works in the present wider context; however, certain new complications arise from the vector-valuedness of u. More specifically, let

$$\pi(u)(t) = \{i : t \in \partial\alpha_i(u)\} \tag{35a}$$

and

$$\gamma(u) = \{t \in [0,1] : \pi(u)(t) \neq \emptyset\}. \tag{35b}$$

If condition (34) is to hold with $\hat{\mathsf{N}}(u) \subset \mathsf{N}_\Omega(u)$ and $\hat{\mathsf{N}}_U(u)(t)$ essentially constant near $\tau \in \gamma(u)$, the collection of normal spaces N_i corresponding to sets $\alpha_i(u)$ for which τ is a frontier point must have a *minimal element*, i.e., for all τ in $\gamma(u)$, there is an integer ν_τ in $\pi(u)(\tau)$ such that

$$\forall i \in \pi(u)(\tau) \quad \mathsf{N}_{\nu_\tau} \subset \mathsf{N}_i \text{ (and therefore } \mathsf{T}_{\nu_\tau} \supset \mathsf{T}_i\text{)}. \tag{36}$$

This requirement is trivially met for $U \subset \mathbb{R}^1$, since N_i is then always either $\{0\}$ or $\mathbb{R}^1$; however, (36) may not hold for vector-valued u, even in the typical case where $\pi(u)(\tau)$ contains just two elements at each τ in $\gamma(u)$. The following development of an example of Tian [16] demonstrates what can happen when (36) is not satisfied.

Example 6.1. Let $U = \{\xi \in \mathbb{R}^2 : \xi_1 \geq 0 \text{ and } \xi_2 \geq 0\}$. For v in $\mathsf{L}_2^\infty(0,1)$, put

$$J(v) = \int_0^1 (\langle r(t), v(t)\rangle + \tfrac{1}{2}\langle v(t), S(u)(t)v(t)\rangle)dt$$

with

$$r(t) = \begin{cases} \left(3(\tfrac{1}{2} - t), -(\tfrac{1}{2} - t)\right), & t \in [0, \tfrac{1}{2}], \\ \left((\tfrac{1}{2} - t), -3(\tfrac{1}{2} - t)\right), & t \in (\tfrac{1}{2}, 1], \end{cases}$$

and

$$[S(u)(t)\eta]_i = \sum_{j=1}^{2} S_{ij}n_j \quad i = 1, 2,$$

where

$$S = \begin{bmatrix} 1 & -2 \\ -2 & 1 \end{bmatrix}.$$

Note that

$$U = \operatorname{ri} \mathcal{F}_1 \cup \operatorname{ri} \mathcal{F}_2 \cup \operatorname{ri} \mathcal{F}_3 \cup \operatorname{ri} \mathcal{F}_4$$

with

$$\begin{aligned} &\operatorname{ri} \mathcal{F}_1 = \{(0,\xi_2) : \xi_2 > 0\}, && \mathcal{N}_1 = \{(\xi_1, 0) : \xi_1 \leq 0\}, \\ &\operatorname{ri} \mathcal{F}_2 = \{(0,0)\}, && \mathcal{N}_2 = \{(\xi_1,\xi_2) : \xi_1 \leq 0, \xi_2 \leq 0\}, \\ &\operatorname{ri} \mathcal{F}_3 = \{(\xi_1,0) : \xi_1 > 0\}, && \mathcal{N}_3 = \{(0,\xi_2) : \xi_2 \leq 0\}, \\ &\operatorname{ri} \mathcal{F}_4 = \{(\xi_1,\xi_2) : \xi_1 > 0, \xi_2 > 0\}, && \mathcal{N}_4 = \{(0,0)\}, \end{aligned}$$

and therefore

$$\begin{aligned} &\mathsf{N}_1 = \operatorname{span}\{(1,0)\}, && \mathsf{T}_1 = \operatorname{span}\{(0,1)\}, \\ &\mathsf{N}_2 = \mathbb{R}^2, && \mathsf{T}_2 = \{(0,0)\}, \\ &\mathsf{N}_3 = \operatorname{span}\{(0,1)\} && \mathsf{T}_3 = \operatorname{span}\{(1,0)\}, \\ &\mathsf{N}_4 = \{(0,0)\}, && \mathsf{T}_1 = \mathbb{R}^2. \end{aligned}$$

Now let

$$u(t) = \begin{cases} (0, -r_2(t)), & t \in [0, \tfrac{1}{2}], \\ (-r_1(t), 0), & t \in (\tfrac{1}{2}, 1]. \end{cases}$$

By construction, u lies in Ω, with

$$\alpha_1(u) = [0, \tfrac{1}{2}), \quad \alpha_2(u) = \{\tfrac{1}{2}\}, \quad \alpha_3(u) = (\tfrac{1}{2}, 1], \quad \alpha_4(u) = \emptyset,$$

$$\gamma(u) = \{\tfrac{1}{2}\},$$

$$\nabla J(u)(t) = \begin{cases} \left((\tfrac{1}{2} - t), 0\right), & t \in [0, \tfrac{1}{2}], \\ \left(0, -(\tfrac{1}{2} - t)\right), & t \in (\tfrac{1}{2}, 1], \end{cases}$$

$$\langle w, \nabla^2 J(u) w\rangle_2 = \int_0^1 \langle w(t), S(u)(t) w(t)\rangle dt,$$

$$\mathcal{N}_U(u(t)) = \begin{cases} \mathcal{N}_1, & t \in [0, \tfrac{1}{2}), \\ \mathcal{N}_2, & t = \tfrac{1}{2}, \\ \mathcal{N}_3, & t \in (\tfrac{1}{2}, 1], \end{cases}$$

$$\mathsf{T}_U(u(t)) = \begin{cases} \mathsf{T}_1, & t \in [0, \tfrac{1}{2}), \\ \mathsf{T}_2, & t = \tfrac{1}{2}, \\ \mathsf{T}_3, & t \in (\tfrac{1}{2}, 1], \end{cases}$$

and

$$\Delta(u)(t) = \begin{cases} (\tfrac{1}{2} - t), & t \in [0, \tfrac{1}{2}), \\ 0, & t = \tfrac{1}{2}, \\ (t - \tfrac{1}{2}), & t \in (\tfrac{1}{2}, 1]. \end{cases}$$

Conditions (32) are seen to hold at u, and $S(u)(\cdot)$ is constant and therefore continuous everywhere in $[0,1]$; nevertheless, u is not an L^∞-local minimizer of J in Ω. To prove this, construct $v_\epsilon \in \Omega$ by the rule

$$v_\epsilon(t) = \begin{cases} (\epsilon, \epsilon) + u(t), & t \in (\tfrac{1}{2} - \epsilon, \tfrac{1}{2} + \epsilon), \\ u(t), & t \in [0,1] \setminus (\tfrac{1}{2} - \epsilon, \tfrac{1}{2} + \epsilon), \end{cases}$$

for $\epsilon \in (0, \tfrac{1}{2})$, and observe that $\|v_\epsilon - u\|_\infty = \sqrt{2}\epsilon$ and $J(v_\epsilon) - J(u) = -\epsilon^3$.

In Example 6.1, the sole frontier point for the sets α_1, α_2, and α_3 is $\tau = \tfrac{1}{2}$, and every other $t \in [0,1]$ lies in the interior of one of these sets in $[0,1]$. As t approaches $\tfrac{1}{2}$ from the right or the left, $\Delta(u)(t)$ converges to 0 and hence the first-order term in Taylor's formula at u does not satisfy (27). Moreover, the second-order term in Taylor's formula cannot carry the resulting additional burden, since it is not possible to extend coercivity of $S(u)(t)$ on T_1 forward beyond $[0, \tfrac{1}{2})$ (where $\mathsf{N}_U(u(t)) = \mathsf{N}_2 = \mathsf{T}_1$) or to extend coercivity of $S(u)(t)$ on T_2 backward beyond $(\tfrac{1}{2}, 1]$ (where $\mathsf{N}_U(u(t)) = \mathsf{N}_1 = \mathsf{T}_2$), even though $S(u)(\cdot)$ is *constant* and therefore continuous. Circumstances of this kind are ruled out in the present analysis when we enforce the "minimal normal subspace" hypothesis (36) and require that

$$\text{(37)} \qquad \forall \tau \in \gamma(u) \quad \text{cl int } \alpha_{\nu_\tau}(u) \supset \alpha_{\nu_\tau}(u) \supset \text{int } \alpha_{\nu_\tau}(u) \neq \emptyset.$$

The following lemma establishes key features of the set $\gamma(u)$ under the aforementioned conditions.

LEMMA 6.3. *For all u in Ω the set $\gamma(u)$ is closed. In addition, if conditions* (36) *and* (37) *hold at u, then*

$$\mu[\gamma(u)] = 0 \tag{38}$$

and

$$\forall \tau \in \gamma(u)\ \forall \epsilon > 0 \quad \mu[\alpha_{\nu_\tau} \cap (\tau - \epsilon, \tau + \epsilon)] > 0. \tag{39}$$

Proof. The set $\pi(u)(\tau)$ is empty iff for some i, τ lies in the interior of $\alpha_i(u)$ in $[0,1]$. Consequently, $\gamma(u)^c$ is open in $[0,1]$, and hence $\gamma(u)$ is closed. If conditions (36) and (37) hold, then

$$\forall \tau \in \gamma(u) \quad \tau \in (\text{cl int } \alpha_{\nu_\tau}(u)) \setminus \text{int } \alpha_{\nu_\tau}(u).$$

Therefore,

$$\begin{aligned} 0 &\leq \mu[\gamma(u)] \\ &\leq \mu[\cup_{i=1}^{d} \ (\text{cl int } \alpha_i(u)) \setminus \text{int } \alpha_i(u)] \\ &\leq \sum_{i=1}^{d} \mu[(\text{cl int } \alpha_i(u)) \setminus \text{int } \alpha_i(u)] \\ &= 0. \end{aligned}$$

Moreover, for all τ in $\gamma(u)$ and $\epsilon > 0$, the interval $(\tau - \epsilon, \tau + \epsilon)$ contains a nonempty open interval in the interior of $\alpha_{\nu_\tau}(u)$. □

Since the set $\gamma(u)$ is closed and therefore compact, it follows that if

$$\lim_{t \to \tau} \text{ess inf } \Delta(u)(t) > 0 \tag{40}$$

at each τ in $\gamma(u)$, then $\Delta(u)(t)$ is essentially bounded away from 0 on some open neighborhood of $\gamma(u)$ in $[0,1]$, and thus on *all* of $[0,1]$ when (32b) holds. Under these circumstances, u satisfies the L^∞ strict complementarity condition (24a) in Theorem 6.1. Our concern now is with the less tractible but frequently encountered cases in which (32), (33), (36), and (37) hold at $\tau \in \gamma(u)$, but

$$\lim_{t \to \tau} \text{ess inf } \Delta_\tau(u)(t) > 0, \tag{41a}$$

where

$$\Delta_\tau(u)(t) = \sup\{\Delta \geq 0 : \forall \xi \in \mathsf{N}_{\nu_\tau},\ \|\xi\| \leq \Delta \Rightarrow -\nabla J(u)(t) + \xi \in \mathcal{N}_U\,(u(t))\}. \tag{41b}$$

Near τ in $[0,1]$, condition (36) implies that $\mathsf{N}_{\nu_\tau} \subset \mathsf{N}_U(u(t))$ and therefore $\Delta_\tau(u)(t) \geq \Delta(u)(t)$; thus (41) is always consistent with (33). Moreover, the next example demonstrates why (41) is often satisfied when τ is a point of continuity for $u(\cdot)$ and $\nabla J(u)(\cdot)$, whereas (40) cannot hold at such points. (Also see the discussion following Theorem 6.4.)

Example 6.2. Let U, $\mathcal{F}_i$, $\mathcal{N}_i$, N_i, and T_i be defined as in Example 6.1, and let

$$J(v) = \int_0^1 \left(v_1(t) + (2t-1)v_2(t) + v_2(t)^2\right) dt$$

for v in L_2^∞. The functional J has a global minimizer in Ω at

$$u(t) = \begin{cases} (0, \frac{1}{2} - t), & t \in [0, \frac{1}{2}), \\ (0,0), & t \in [\frac{1}{2}, 1], \end{cases}$$

with

$$\alpha_1(u) = [0, \tfrac{1}{2}), \quad \alpha_2(u) = [\tfrac{1}{2}, 1], \quad \alpha_3(u) = \emptyset, \quad \alpha_4(u) = \emptyset,$$

$$\gamma(u) = \{\tfrac{1}{2}\},$$

$$\nu_{1/2} = 1, \quad \mathsf{N}_{\nu_{1/2}} = \mathsf{N}_1 = \text{span}\,\{(1,0)\},$$

$$\mathcal{N}_U\,(u(t)) = \begin{cases} \mathcal{N}_1, & t \in [0, \frac{1}{2}), \\ \mathcal{N}_2, & t \in [\frac{1}{2}, 1], \end{cases}$$

$$\nabla J(u)(t) = \begin{cases} (1,0), & t \in [0, \frac{1}{2}), \\ (1, 2t-1), & t \in [\frac{1}{2}, 1], \end{cases}$$

and therefore

$$\Delta(u)(t) = \begin{cases} 1, & t \in [0, \frac{1}{2}), \\ 2t-1, & t \in [\frac{1}{2}, 1], \end{cases}$$

and

$$\lim_{t \to \frac{1}{2}^+} \Delta(u)(t) = 0,$$

whereas

$$\Delta_{1/2}(u)(t) = 1, \quad t \in [0,1].$$

It is now possible to state and prove the following sufficient conditions for L^∞-local optimality in Ω.

THEOREM 6.4. *Let J be a twice directionally differentiable real-valued function on a domain $\mathcal{D} \subset \mathsf{L}_m^\infty[0,1]$ that is open relative to the norm $\|\cdot\|_\infty$, and assume that the structure/continuity conditions* (2)–(3) *hold with $\nu = \infty$. Suppose that the formal sufficient conditions* (32) *are satisfied at some point u in $\Omega \cap \mathcal{D}$, along with* (36), (37), *and* (41) *at each τ in $\gamma(u)$. In addition, assume that $S(u)(\cdot)$ is continuous on the set $\gamma(u)$ in* (35). *Then for each number $c_\infty \in (0, c_T)$ there is a corresponding δ_∞ such that for all v,*

$$(42) \qquad v \in \Omega \text{ and } \|v - u\|_\infty < \delta_\infty \Rightarrow J(v) - J(u) \geq \tfrac{1}{2} c_\infty \|v - u\|_2^2.$$

Proof. Conditions (22) and (23) follow at once from the hypotheses. Suppose $\gamma(u) = \emptyset$. Then for some i, $[0,1] = \alpha_i(u) = \text{int}\,\alpha_i(u)$, in which case $[0,1]$ is a compact set in int $\alpha_i(u)$, conditions (32a)–(32b) imply the L^∞ strict complementarity condition (24a), and assertion (42) is established by Theorem 6.1. Suppose $\gamma(u) \neq \emptyset$. Then (42) will follow from Lemma 6.2 if it can be shown that for each $c_{\hat{T}} \in (0, c_T)$ the coercivity condition (32c) lifts to (30b) on a larger subspace $\hat{\mathsf{T}}(u) \supset \mathsf{T}_\Omega(u)$, while (32a)–(32b) and (41) imply the strict complementarity condition (30a) on the corresponding pointwise orthogonal complement $\hat{\mathsf{N}}(u) \subset \mathsf{N}_\Omega(u)$.

Fix $c_{\hat{T}}$ in $(0, c_T)$, and note first that by (38) in Lemma 6.3 there is an open set $\mathcal{O}$ in $[0,1]$, with $\mathcal{O} \supset \gamma(u)$ and $\mu[\mathcal{O}]$ so small that

$$\left(\int\int_{(\mathcal{O}^c\times\mathcal{O}^c)^c} \|K(u)(t,s)\|^2 dtds\right)^{\frac{1}{2}} \leq \frac{1}{2}(c_T - c_{\hat{T}}), \tag{43}$$

where $\mathcal{O}^c = [0,1]\setminus\mathcal{O}$, $(\mathcal{O}^c\times\mathcal{O}^c)^c = [0,1]\times[0,1]\setminus(\mathcal{O}^c\times\mathcal{O}^c)$. Moreover, (32c) implies that for all $i = 1,\ldots,d$ and all ξ,

$$\xi \in \mathsf{T}_i \Rightarrow \langle\xi, S(u)(t)\xi\rangle \geq c_T\|\xi\|^2 \quad \text{a.e. in } \alpha_i(u).$$

This can be proved in the same way that (14b) is established in the proof of Theorem 5.3. Since $S(u)(\cdot)$ is continuous on $\gamma(u)$, it now follows from (39) in Lemma 6.3 that for all τ and ξ,

$$\tau \in \gamma(u) \text{ and } \xi \in \mathsf{T}_{\nu_\tau} \Rightarrow \langle\xi, S(u)(\tau)\xi\rangle \geq c_T\|\xi\|^2. \tag{44}$$

By continuous extension, this implies that for each $\tau \in \gamma(u)$ there is a δ_τ such that

$$(\tau - \delta_\tau, \tau + \delta_\tau) \cap [0,1] \subset \mathcal{O}, \tag{45a}$$

and for all t in $[0,1]$ and all ξ,

$$t \in (\tau - \delta_\tau, \tau + \delta_\tau) \cap [0,1] \text{ and } \xi \in \mathsf{T}_{\nu_\tau} \Rightarrow \langle\xi, S(u)(t)\xi\rangle \geq \tfrac{1}{2}(c_T + c_{\hat{T}})\|\xi\|^2. \tag{45b}$$

On the other hand, condition (41) implies that for each τ in $\gamma(u)$, there are positive numbers c_τ and $\delta'_\tau \in (0, \delta_\tau]$ such that

$$\Delta_\tau(u)(t) \geq c_\tau$$

and

$$\mathsf{N}_{\nu_\tau} \subset \mathsf{N}_U(u(t)) \text{ and } \mathsf{T}_{\nu_\tau} \supset \mathsf{T}_U(u(t))$$

for almost all t in $(\tau - \delta'_\tau, \tau + \delta'_\tau) \cap [0,1] \subset \mathcal{O}$. Since $\gamma(u)$ is compact, this ensures the existence of a finite set $\{\tau_1, \ldots, \tau_k\} \subset \gamma(u)$, a corresponding system of intervals $\{\mathcal{I}_1, \cdots, \mathcal{I}_k\}$ in $[0,1]$, and a positive number c_γ such that the set

$$\mathcal{I} = \bigcup_{l=1}^{k} \mathcal{I}_l \tag{46a}$$

is open in $[0,1]$, and for all i and j and almost all t in $\mathcal{I}_i$,

$$i \neq j \Rightarrow \mathcal{I}_i \cap \mathcal{I}_j = \emptyset, \tag{46b}$$

$$\mathcal{I}_i \subset (\tau - \delta'_\tau, \tau + \delta'_\tau) \cap [0,1] \subset \mathcal{O}, \tag{46c}$$

$$\Delta_{\tau_i}(u)(t) \geq c_\gamma, \tag{46d}$$

$$\mathsf{N}_{\nu_{\tau_i}} \subset \mathsf{N}_U(u(t)) \text{ and } \mathsf{T}_{\nu_{\tau_i}} \supset \mathsf{T}_U(u(t)), \tag{46e}$$

and finally

$$\gamma(u) \subset \mathcal{I} \subset \mathcal{O}. \tag{46f}$$

The requisite subspaces $\hat{\mathsf{N}}(u)$ and $\hat{\mathsf{T}}(u)$ are now defined by (34) and the following rule: for all i and t

$$t \in \mathcal{I}_i \Rightarrow \hat{\mathsf{N}}(u)(t) = \mathsf{N}_{\nu_{\tau_i}} \text{ and } \hat{\mathsf{T}}(u)(t) = \mathsf{T}_{\nu_{\tau_i}}, \tag{47a}$$

$$t \in \mathcal{I}^c \Rightarrow \hat{\mathsf{N}}(u)(t) = \mathsf{N}_U(u(t)) \text{ and } \hat{\mathsf{T}}(u)(t) = \mathsf{T}_U(u(t)). \tag{47b}$$

By construction, $\hat{\mathsf{N}}(u) \subset \mathsf{N}_\Omega(u)$, $\hat{\mathsf{T}}(u) \supset \mathsf{T}_\Omega(u)$, $\hat{\mathsf{N}}(u)$ and $\hat{\mathsf{T}}(u)$ are pointwise orthogonal, and the same argument used in the proof of Theorem 6.1 shows that $\mathsf{L}_m^\infty(0,1)$ is the direct sum of $\hat{\mathsf{N}}(u)$ and $\hat{\mathsf{T}}(u)$. Furthermore, since $\mathcal{I}^c$ is compact and $\mathcal{I}^c \subset \gamma(u)^c = \cup_{i=1}^d \operatorname{int} \alpha_i(u)$, conditions (32a)–(32b) and (46d) imply (30a) with $c_{\hat{N}} = \min\{c_\beta, c_\gamma\}$ and $\beta = \mathcal{I}^c$. To prove (30b), note that for all w in $\mathsf{L}_m^\infty(0,1)$,

$$\begin{aligned}\langle w, \nabla^2 J(u) w\rangle_2 &= \int_{[0,1]} \langle w(t), S(u)(t)w(t)\rangle dt + \int\int_{[0,1]\times[0,1]} \langle w(t), K(u)(t,s)w(s)\rangle dtds \\ &= \int_{\mathcal{I}^c} \langle w(t), S(u)(t)w(t)\rangle dt + \int\int_{\mathcal{I}^c\times\mathcal{I}^c} \langle w(t), K(u)(t,s)w(s)\rangle dtds \\ &\quad + \int_{\mathcal{I}} \langle w(t), S(u)(t)w(t)\rangle dt + \int\int_{(\mathcal{I}^c\times\mathcal{I}^c)^c} \langle w(t), K(u)(t,s)w(s)\rangle dtds.\end{aligned}$$

Hence, for all w in $\hat{\mathsf{T}}(u)$, conditions (32c) and (43)–(46) imply that

$$\begin{aligned}\langle w, \nabla^2 J(u) w\rangle_2 &\geq c_T \int_{\mathcal{I}^c} \|w(t)\|^2 dt + \tfrac{1}{2}(c_T + c_{\hat{T}}) \textstyle\int_{\mathcal{I}} \|w(t)\|^2 dt \\ &\quad - \left(\int\int_{(\mathcal{O}^c\times\mathcal{O}^c)^c} \|K(u)(t,s)\|^2 dtds\right)^{\frac{1}{2}} \|w\|_2^2 \\ &\geq \tfrac{1}{2}(c_T + c_{\hat{T}})\|w\|_2^2 - \tfrac{1}{2}(c_T - c_{\hat{T}})\|w\|_2^2 \\ &= c_{\hat{T}}\|w\|_2^2. \qquad \square\end{aligned}$$

Theorem 6.4 has an immediate corollary for optimal control problems with Bolza objective functions (5) and admissible control input sets

$$U = \{\xi \in \mathbb{R}^m; g_j(\xi) \leq 0, \quad j = 1, \ldots, l\}, \tag{48a}$$

where each g_j is an affine real function with values

$$g_j(\xi) = \langle a^j, \xi\rangle - b_j. \tag{48b}$$

Theorem 3 and lemmas 3–5 in [2] produce similar L^∞ sufficient conditions for a large class of Bolza optimal control problems with $\mathcal{C}^2$ equality constraints on the terminal state $x(1)$ and admissible control input sets $U \subset \mathbb{R}^m$ prescribed by l $\mathcal{C}^2$ inequality constraints. Although the analysis in §4 of [2] is also rooted in the proof strategy of [1], it proceeds from different assumptions and does not subsume Theorem 6.4 for the class of fixed-time free-endpoint Bolza problems with polyhedral control input sets U. When U is given by (48), the Farkas lemma implies that at each ξ in U, the normal cone $\mathcal{N}_U(\xi)$ consists of all nonnegative linear combinations of the active constraint gradients

$$\nabla g_j(\xi) = a^j, \quad j \in A_0(\xi), \tag{49a}$$

where

$$A_0(\xi) = \{j : g_j(\xi) = 0\}. \tag{49b}$$

Under these circumstances, the pointwise stationarity condition (14a) holds everywhere in $[0,1]$ iff there are Lagrange multipliers $\lambda(t)$ in $\mathbb{R}^l$ such that

$$\forall t \in [0,1] \quad -\nabla J(u)(t) = \sum_{j=1}^{l} \lambda_j(t) \nabla g_j\left(u(t)\right) \tag{50a}$$

with

$$\lambda_j(t) \geq 0, \quad j \in A_0\left(u(t)\right), \tag{50b}$$

$$\lambda_j(t) = 0, \quad j \notin A_0\left(u(t)\right). \tag{50c}$$

In §4 of [2], the active constraint gradients in (49) are assumed to be linearly independent at $\xi = u(t)$ for $t \in [0,1]$, and the complementarity condition (50b) is strengthened by requiring that

$$\forall i \, \forall \beta_i \subset \text{int } A_0^i \quad \left(\beta_i \text{ compact } \Rightarrow \inf_{t \in \beta_i} \lambda_i(t) > 0\right), \tag{51a}$$

where

$$A_0^i = \{t \in [0,1] : g_i\left(u(t)\right) = 0\}. \tag{51b}$$

It is further assumed in [2] that for all τ,

$$\tau \in \bigcup_{j=1}^{l} \partial A_0^j \Rightarrow \exists c_\tau > 0 \, \forall \xi \in \overline{T}_0(\tau), \quad \langle \xi, S(\tau)\xi \rangle \geq c_\tau \|\xi\|^2, \tag{52a}$$

where

$$\overline{T}_0(\tau) = \left[\{\nabla g_j\left(u(\tau)\right)\}_{j \in \overline{A}_0(\tau)}\right]^\perp, \tag{52b}$$

$$\overline{A}_0(\tau) = \bigcup_{\sigma > 0} \bigcap_{s \in (\tau - \sigma, \tau + \sigma) \cap [0,1]} A_0\left(u(s)\right), \tag{52c}$$

and

$$S(\tau) = \nabla^2_{uu} H\left(\tau, \psi(\tau), x(\tau), u(\tau)\right). \tag{52d}$$

(See Note 5.4.) The following lemma and note explain how the strengthened complementarity condition (51) and the pointwise coercivity condition (52) are related to the hypotheses in Theorem 6.4.

LEMMA 6.5. *Let U be a polyhedral convex set with representation* (48) *in* $\mathbb{R}^m$. *Then for each face $\mathcal{F}_i$ in U, there is a unique subset $\mathcal{A}_i$ of $\{1,\ldots,l\}$ such that*

$$\text{ri } \mathcal{F}_i = \{\xi \in U : A_0(\xi) = \mathcal{A}_i\}. \tag{53}$$

Suppose that

$$u \in \mathsf{L}_m^\infty(0,1) \textit{ and } (\forall t \in [0,1] \quad u(t) \in U). \tag{54}$$

Then

$$\forall i \quad \text{int } \alpha_i(u) \subset \bigcap_{j \in \mathcal{A}_i} \text{int } A_0^j, \tag{55}$$

$$\gamma(u) = \bigcup_{j=1}^{l} \partial A_0^j, \tag{56}$$

and

$$\forall \tau \in \gamma(u) \quad \overline{A}_0(\tau) = \bigcap_{j \in \pi(u)(\tau)} \mathcal{A}_j. \tag{57}$$

Furthermore, assume that (54) *holds at u and, for each τ in $\gamma(u)$, the collection of index sets $\{\mathcal{A}_j\}_{j\in\pi(u)(\tau)}$ has a minimal element, i.e.,*

$$\exists \nu_\tau \in \pi(u)(\tau)\ \forall j \in \pi(u)(\tau) \quad \mathcal{A}_{\nu_\tau} \subset \mathcal{A}_j. \tag{58}$$

Then for all τ in $\gamma(u)$,

$$\overline{A}_0(\tau) = \mathcal{A}_{\nu_\tau}, \tag{59a}$$

$$\forall j \in \pi(u)(\tau) \quad \mathsf{N}_{\nu_\tau} \subset \mathsf{N}_j, \tag{59b}$$

and

$$\mathsf{T}_{\nu_\tau} = \overline{\mathsf{T}}_0(\tau). \tag{59c}$$

Finally, assume that (50), (54), *and* (58) *hold at u and that the active constraint gradients in* (49) *are also linearly independent at $\xi = u(t)$ for all t in $[0,1]$. Then the algebraic strict complementarity condition* (51) *is equivalent to the geometric strict complementarity conditions*

$$\forall \beta \subset \bigcup_{i=1}^{d} \text{int } \alpha_i(u) \quad \left(\beta \text{ compact } \Rightarrow \inf_{t\in\beta} \Delta(u)(t) > 0\right), \tag{60a}$$

$$\forall \tau \in \gamma(u) \quad \lim_{t\to\tau} \inf \Delta_\tau(u)(t) > 0. \tag{60b}$$

Proof. For each $\xi \in U$,

$$\mathcal{N}_U(\xi) = \{\zeta \in \mathbb{R}^m : \exists \lambda \in \mathbb{R}^l_+ \quad \zeta = \sum_{j \in A_0(\xi)} \lambda_j a^j\}, \tag{61a}$$

where

$$\mathbb{R}^l_+ = \{\lambda \in \mathbb{R}^l : \forall j \quad \lambda_j \geq 0\}. \tag{61b}$$

It follows from (48) and (61) that for all $\xi \in U$ the set

$$\mathcal{F}(\xi) = \left[\xi + \mathcal{N}_U(\xi)^\perp\right] \cap U$$

is a polyhedral face in U, with

$$\operatorname{aff} \mathcal{F}(\xi) = \xi + \mathcal{N}_U(\xi)^\perp,$$

$$\operatorname{ri} \mathcal{F}(\xi) = \{\eta \in U : A_0(\eta) = A_0(\xi)\},$$

and

$$\eta \notin \operatorname{ri} \mathcal{F}(\xi) \Rightarrow \operatorname{ri} \mathcal{F}(\eta) \cap \operatorname{ri} \mathcal{F}(\xi) = \emptyset$$

for all $\eta \in U$ [15]. Since $A_0(\cdot)$ has range in the subsets of $\{1, \ldots, l\}$, there are d distinct index sets $\mathcal{A}_i \subset \{1, \ldots, l\}$ and d corresponding polyhedral faces $\mathcal{F}_i \subset U$ such that (53) holds, with

$$\eta \in \operatorname{ri} \mathcal{F}_i \Rightarrow \mathcal{N}_U(\eta) = \mathcal{N}_i = \left\{\zeta \in \mathbb{R}^m : \exists \lambda \in \mathbb{R}^l_+ \quad \zeta = \sum_{j \in \mathcal{A}_i} \lambda_j a^j\right\}, \tag{62a}$$

$$\eta \in \operatorname{ri} \mathcal{F}_i \Rightarrow \operatorname{span} \mathcal{N}_U(\eta) = \mathsf{N}_i = \operatorname{span} \{a^j\}_{j \in \mathcal{A}_i}, \tag{62b}$$

$$\operatorname{ri} \mathcal{F}_i \cap \operatorname{ri} \mathcal{F}_j = \emptyset, \tag{62c}$$

and

$$\cup_{i=1}^d \operatorname{ri} \mathcal{F}_i = U \tag{62d}$$

for all $\eta \in U$ and all $i, j \in \{1, \ldots, l\}$.

Suppose that (54) holds. By (12a) and (53),

$$\alpha_i(u) = \{t \in [0,1] : A_0\left(u(t)\right) = \mathcal{A}_i\} \tag{63a}$$

and

$$\bigcup_{i=1}^d \alpha_i(u) = [0,1]. \tag{63b}$$

If int $\alpha_i(u) = \emptyset$, then (55) holds trivially. Suppose that $\tau \in$ int $\alpha_i(u)$. Then for some $\delta > 0$ and all t,

$$\begin{aligned} t \in (\tau - \delta, \tau + \delta) \cap [0,1] &\Rightarrow A_0\left(u(t)\right) = \mathcal{A}_i \\ &\Rightarrow \forall j \in \mathcal{A}_i \quad g_j\left(u(t)\right) = 0 \\ &\Rightarrow \forall j \in \mathcal{A}_i \quad t \in A_0^j, \end{aligned}$$

and therefore $\tau \in \text{ int } A_0^j$ for all $j \in \mathcal{A}_i$. This proves (55).

Suppose that $\tau \in \partial A_0^i$ for some i. Then every neighborhood of τ in $[0,1]$ contains points t and s where $g_i(u(t)) = 0$ and $g_i(u(s)) \neq 0$, and hence $A_0(u(t)) \neq A_0(u(s))$. By (35), it follows that $\tau \in \left[\cup_{i=1}^d \text{int } \alpha_i(u)\right]^c = \gamma(u)$. Conversely, suppose that $\tau \notin \cup_{i=1}^l \partial A_0^i$ or, equivalently,

$$\forall i \quad \tau \in \text{ int } A_0^i \cup \text{int } (A_0^i)^c.$$

Then

$$\exists \delta > 0 \, \forall t \in (\tau - \delta, \tau + \delta) \cap [0,1] \quad A_0(u(t)) = A_0(u(\tau)),$$

in which case $\tau \in \text{ int } \alpha_i(u)$ for some i and therefore $\tau \in \gamma(u)^c$. This proves (56).

Since $A_0(\cdot)$ has finitely many values $\mathcal{A}_i$, it follows that for all $t \in [0,1]$ there is a $\sigma_t > 0$ such that

$$\forall \sigma \in (0, \sigma_t] \quad \overline{A}_0(t) = \bigcap_{s \in (t-\sigma, t+\sigma) \cap [0,1]} A_0(u(s)).$$

Moreover, if $\tau \in \gamma(u)$, then by (35) it is also true that

$$\forall \sigma \in (0, \sigma_\tau] \quad A_0[(\tau - \sigma, \tau + \sigma) \cap [0,1]] = \{\mathcal{A}_j\}_{j \in \pi(u)(\tau)}.$$

Assertion (57) is now proved, and (59) is an immediate consequence of (35), (52), (57), (62b), and the hypothesis (58).

Suppose that (50), (58), and (60) hold at u and that the active constraint gradients in (49) are linearly independent at $\xi = u(t)$ for $t \in [0,1]$. Let β_q be a compact set in $\text{int} A_0^q$ for some q. Suppose that $\beta_q \cap \gamma(u) \neq \emptyset$. By Lemma 6.3, the set $\beta_q \cap \gamma(u)$ is compact. Hence there is a $\delta > 0$, a finite set $\{\tau_1, \ldots, \tau_k\} \subset \beta_q \cap \gamma(u)$, and a corresponding family of open intervals $\{\mathcal{I}_1, \ldots, \mathcal{I}_k\}$ in $[0,1]$ such that

$$\text{(64a)} \qquad \mathcal{I} \overset{\text{def}}{=} \bigcup_{i=1}^k \mathcal{I}_i \supset \beta_q \cap \gamma(u),$$

$$\text{(64b)} \qquad \inf_{t \in \mathcal{I}_i} \Delta_{\tau_i}(u)(t) \geq \delta,$$

$$\text{(64c)} \qquad \forall i \quad \tau_i \in \mathcal{I}_i \subset \text{int } A_0^q,$$

$$\text{(64d)} \qquad \forall i \quad A_0[\mathcal{I}_i] = \{\mathcal{A}_j\}_{j \in \pi(u)(\tau)},$$

and thus

$$\text{(64e)} \qquad \forall t \in \mathcal{I}_i \quad q \in \mathcal{A}_{\nu_{\tau_i}} \subset A_0(u(t))$$

in view of (58). Now note that

$$\text{(65a)} \qquad \forall \lambda \in \mathbb{R}^l \quad \left\| \sum_{j=1}^l \lambda_j a^j \right\| \leq M_a \max_{1 \leq j \leq l} |\lambda_j|$$

with

$$\text{(65b)} \qquad M_a = \sum_{j=1}^{l} \|a^j\| > 0.$$

Since active constraint gradients are linearly independent, conditions (41b), (50), (58), (62), (64), and (65) imply that

$$\forall i \quad \inf_{t \in \mathcal{I}_i} \lambda_q(t) \geq \frac{\delta}{M_a}$$

and hence

$$\inf_{t \in \beta_q \cap \mathcal{I}} \lambda_q(t) \geq \frac{\delta}{M_a}.$$

Suppose that $\beta_q \cap \mathcal{I}^c \neq \emptyset$. By (64a), $\beta_q \cap \mathcal{I}^c$ is a compact set in $\gamma(u)^c = \cup_{i=1}^d \operatorname{int} \alpha_i(u)$, condition (60a) yields

$$\inf_{t \in \beta_q \cap \mathcal{I}^c} \Delta(u)(t) \overset{\text{def}}{=} \delta_c > 0,$$

and therefore (32d), (50), (62), and (65) give

$$\inf_{t \in \beta_q \cap \mathcal{I}^c} \lambda_q(t) \geq \frac{\delta_c}{M_a} > 0$$

since active constraint gradients are linearly independent and $q \in A_0(u(t))$ for all $t \in A_0^q \supset \beta$. Finally, if $\beta_q \cap \gamma(u) = \emptyset$, then β_q is a compact set in $\gamma(u)^c = \cup_{i=1}^d \operatorname{int} \alpha_i(u)$, and a repetition of the preceding argument shows that $\lambda_q(t)$ is again bounded away from 0 on β_q. This proves that (51) is implied by (60).

Conversely, suppose that (50), (51), and (60) hold at u and that the active constraint gradients in (49) are linearly independent at $\xi = u(t)$ for $t \in [0,1]$. The latter condition ensures that

$$\text{(66)} \qquad \exists m_a > 0 \; \forall t \in [0,1] \; \forall \lambda \in \mathbb{R}^l \quad \left\| \sum_{j \in A_0(u(t))} \lambda_j a^j \right\| \geq m_a \max_{j \in A_0(u(t))} |\lambda_j|.$$

Let β be a compact set in $\cup_{i=1}^d \operatorname{int} \alpha_i(u)$. Since the sets $\operatorname{int} \alpha_i(u)$ are open in $[0,1]$ and pairwise disjoint, the corresponding sets $\beta_i = \beta \cap \operatorname{int} \alpha_i(u)$ are compact sets in $\operatorname{int} \alpha_i(u)$, with $\beta = \cup_{i=1}^d \beta_i$. Therefore, by (55) and (51),

$$\forall i \; \forall j \in \mathcal{A}_i \quad \beta_i \text{ compact and } \beta_i \subset \operatorname{int} A_0^j,$$

and therefore

$$\exists \delta > 0 \; \forall i \quad \inf_{t \in \beta_i} \min_{j \in A_0(u(t))} \lambda_j(t) \geq \delta > 0.$$

It now follows from (12), (32d), (61), (62), and (66) that

$$\forall i \quad \inf_{t \in \beta_i} \Delta(u)(t) \geq \delta m_a > 0,$$

and therefore

$$\inf_{t\in\beta} \Delta(u)(t) \geq \delta m_a > 0.$$

This proves (60a). Furthermore, suppose that (58) also holds at u. Then for each $\tau \in \gamma(u)$ there is a $\delta_\tau > 0$ such that for all t

$$\begin{aligned} t \in (\tau - \delta_\tau, \tau + \delta_\tau) \cap [0,1] \quad &\Rightarrow \mathcal{A}_{\nu_\tau} \subset A_0\left(u(t)\right) \\ &\Rightarrow \forall j \in \mathcal{A}_{\nu_\tau} \quad g_j\left(u(t)\right) = 0. \end{aligned}$$

Thus $\tau \in \operatorname{int} A_0^j$ for all $j \in \mathcal{A}_{\nu_\tau}$, and condition (51) ensures that

$$\forall j \in \mathcal{A}_{\nu_\tau} \quad \lim_{t\longrightarrow\tau} \inf \lambda_j(t) > 0.$$

Assertion (60b) now follows from (41b), (50), (58), (62), and (66). □

Note 6.3. According to (56) and (59c), the pointwise coercivity condition (52) reduces to

$$\text{(67a)} \qquad \forall \tau \in \gamma(u)\, \exists c_\tau > 0\, \forall \xi \in \mathsf{T}_{\nu_\tau} \quad \langle \xi, S(\tau)\xi\rangle \geq c_\tau \left\|\xi\right\|^2,$$

$$\text{(67b)} \qquad S(\tau) = \nabla^2_{uu} H\left(\tau, \psi(\tau), x(\tau), u(\tau)\right)$$

when the minimal index set restriction (58) is satisfied. In [2], condition (52) is taken as a *hypothesis*. In the present analysis, the pointwise coercivity condition (67a) is *deduced* from (36), (37), and the coercivity condition (32c), and the relationship between (32c) and second-order *necessary* conditions for L^∞-local optimality is established in Theorems 5.2 and 5.3. On the other hand, the sufficient conditions in §4 of [2] impose the linear independence qualification on active constraint gradients but do not assume the existence of a minimal element in the family of index sets $\{\mathcal{A}_i\}_{i\in\pi(u)(\tau)}$ for τ in $\gamma(u)$; however, if the active constraint gradient sets in (49) *are* linearly independent at $\xi = u(t)$ for $t \in [0,1]$ and (58) is *not* satisfied, then (52b), (57), and (62) imply that $\overline{\mathsf{T}}_0(\tau)$ contains each subspace in the family $\{\mathsf{T}_i\}_{i\in\pi(u)(\tau)}$ as a *proper subset*, and the pointwise coercivity condition (52) cannot be inferred from the counterpart of (32c) in [2]. To put this another way, when (58) does not hold in the setting of [2], the gap between sufficient conditions and necessary conditions widens substantially.

In this connection, an extreme case of some interest occurs when u is a step function with range in the vertex set of the polyhedron U (e.g., a *bang-bang* control). Under these circumstances, it can be seen that for almost all t in $[0,1]$ and τ in $\gamma(u)$: $\mathsf{N}_U(u(t)) = \mathbb{R}^m$ and $\mathsf{T}_U(u(t)) = \{0\}$; $\{\mathsf{N}_i\}_{i\in\pi(u)} = \mathbb{R}^m$; $\mathsf{N}_{\nu_\tau} = \mathbb{R}^m$; $\Delta_\tau(u)(t) = \Delta(u)(t)$; *card* $A_0\left(u(t)\right) = m$; $\{\mathcal{A}_i\}_{i\in\pi(u)(\tau)}$ has no minimal element; $\overline{A}_0(\tau)$ is a proper subset of $\mathcal{A}_j$ for all $j \in \pi(u)(\tau)$; $1 \leq$ *dim* $\overline{\mathsf{T}}_0(\tau) \leq m$; condition (32c) is satisfied trivially; conditions (51) can hold when conditions (60) do not, but condition (52) is unrelated to the necessary condition (14b) and is not likely to hold for nonconvex objective functions.

The stationary control u in example 2 of [1] satisfies the L^∞ sufficient conditions in Theorem 6.4 with $\nu = 2$ but does not obey the necessary condition (17) for L^2-local optimality in Theorem 5.4. Hence, the L^∞ sufficient conditions do not imply L^2-local optimality, even if (2) and (3) hold with $\nu = 2$. On the other hand, it will now be shown that L^2-local optimality *is* implied by (42) and a strengthened variant of (17) when (2) and (3) hold with $\nu = 2$; this result and *any* sufficient conditions for (42) immediately yield sufficient conditions for L^2-local optimality.

THEOREM 6.6. *Let J be a twice directionally differentiable real-valued function on a domain $\mathcal{D} \subset \mathsf{L}^\infty_m[0,1]$ that is open relative to the norm $\|\cdot\|_2$, and suppose that the structure/continuity conditions* (2)–(3) *hold with $\nu = 2$. Assume that u is a proper L^∞-local minimizer of J in $\Omega \cap \mathcal{D}$ satisfying* (42). *In addition, suppose that for some $c_P > 0$, the condition*

$$(68)\ \forall \xi \in U \quad \langle \nabla J(u)(t), \xi - u(t)\rangle + \tfrac{1}{2}\langle \xi - u(t), S(u)(t)\,(\xi - u(t))\rangle \geq \tfrac{1}{2}c_P\|\xi - u(t)\|^2$$

holds for almost all t in $[0,1]$. Then u is a proper L^2-local minimizer of J in $\Omega \cap \mathcal{D}$, and for each c_2 in the interval $0 < c_2 < \min\{c_T, c_P\}$ there is a corresponding $\delta_2 > 0$ such that for all v

$$(69) \qquad v \in \Omega \text{ and } \|v - u\|_2 < \delta_2 \Rightarrow J(v) - J(u) \geq c_2\|v - u\|_2^2.$$

Proof. Fix c_2 in the interval $0 < c_2 < \min\{c_T, c_P\}$, fix c_∞ in the interval $\frac{1}{2}(\min\{c_T, c_P\} + c_2) < c_\infty < \min\{c_T, c_P\}$, and choose $\delta_\infty > 0$ so that (42) is satisfied. Note that if (2) and (3) are satisfied with $\nu = 2$, then Taylor's formula (4) holds with $\nu = 2$ *and* $\nu = \infty$. Consequently, there is a $\rho \in (0, \delta_\infty]$ such that for all v in Ω,

$$(70\text{a}) \qquad \|v - u\|_2 \leq \rho \Rightarrow v \in \mathcal{D} \cap \Omega \text{ and } |r(u; v)| \leq \tfrac{1}{8}\left(\min\{c_T, c_P\} - c_2\right)\|v - u\|_2^2$$

and

(70b)

$$\|v - u\|_\infty \leq \rho \Rightarrow \langle \nabla J(u), v - u\rangle_2 + \tfrac{1}{2}\langle v - u, \nabla^2 J(u)(v - u)\rangle_2 \geq \tfrac{1}{4}\left(\min\{c_T, c_P\} + c_2\right)\|v - u\|_2^2.$$

Furthermore, there is an $\epsilon \in (0,1]$ such that for all measurable sets $\theta \subset [0,1]$

$$(71) \qquad \mu[\theta] < \epsilon \Rightarrow \int\int_{(\theta^c \times \theta^c)^c} \|K(u)(t,s)\|^2 dtds \leq \tfrac{1}{8}\left(\min\{c_T, c_P\} - c_2\right).$$

For each $v \in \Omega$, let

$$(72\text{a}) \qquad \theta_v = \{t \in [0,1] : \|v(t) - u(t)\| > \rho\},$$

and define $w_v \in \Omega$ by the rule

$$(72\text{b}) \qquad w_v(t) = \begin{cases} v(t), & t \in \theta_v^c, \\ u(t), & t \in \theta_v. \end{cases}$$

By construction, θ_v is measurable, and for all v,

$$(73) \qquad v \in \Omega \text{ and } \|v - u\|_2 \leq \rho\epsilon^{\frac{1}{2}} \Rightarrow v \in \Omega \cap \mathcal{D},\ \mu[\theta_v] < \epsilon \text{ and } \|w_v - u\|_\infty \leq \rho.$$

Put $\delta_2 = \rho\epsilon^{\frac{1}{2}}$. Then, in view of (70)–(73) and Taylor's formula (4) with $\nu = 2$, it now follows that for all $v \in \Omega$ such that $\|v - u\|_2 \leq \delta_2$,

$$\begin{aligned} J(v) - J(u) = {} & \langle \nabla J(u), w_v - u\rangle_2 + \tfrac{1}{2}\langle w_v - u, \nabla^2 J(u)(w_v - u)\rangle_2 \\ & + \int_{\theta_v} \left[\langle \nabla J(u)(t), v(t) - u(t)\rangle + \tfrac{1}{2}\langle v(t) - u(t), S(u)(t)\,(v(t) - u(t))\rangle\right] dt \end{aligned}$$

$$
\begin{aligned}
&+ \int\int_{(\theta_v^c\times\theta_v^c)^c} \langle v(t)-u(t), K(u)(t,s)\,(v(t)-u(t))\rangle dtds + r(u;v)\\
&\geq \tfrac{1}{4}\left(\min\{c_T,c_P\}+c_2\right)\|w_v-u\|_2^2+\tfrac{1}{2}c_P\int_{\theta_v}\|v(t)-u(t)\|^2dt\\
&\quad -\tfrac{1}{4}\left(\min\{c_T,c_P\}-c_2\right)\|v-u\|_2^2\\
&\geq \tfrac{1}{4}\left(\min\{c_T,c_P\}+c_2\right)\|v-u\|_2^2-\tfrac{1}{4}\left(\min\{c_T,c_P\}-c_2\right)\|v-u\|_2^2\\
&= \tfrac{1}{2}c_2\|v-u\|_2^2. \qquad \square
\end{aligned}
$$

Note 6.4. Theorem 6.6 remains valid if the range set U in (1b) is an arbitrary subset of $\mathbb{R}^m$ (cf. Note 5.3).

Note 6.5. Theorem 4 in [1] gives sufficient conditions for L^∞-local optimality and L^2-local optimality in the set of L^p functions with range in $U=[0,\infty)\subset\mathbb{R}^1$. For $p=\infty$, these results are contained in Theorems 6.4 and 6.6, since (36), (37), and (41) hold trivially when $U=[0,\infty)$ and the set $\alpha(u)=\{t\in[0,1]:u(t)=0\}$ is closed, condition (68) is easily verified when $U=[0,\infty)$ and $S(u)(t)$ is essentially bounded away from 0 on $[0,1]$, and the condition $\|K(u)\|_\infty<\infty$ implies the weaker requirement $\|K(u)\|_2<\infty$ imposed here.

Note 6.6. If (2) and (3) hold with $\nu=2$ for Bolza objective functions (5), then the Hamiltonian H is quadratic in u (see Note 5.4), and hypothesis (68) is equivalent to the following strengthened version of the Pontryagin minimum principle: for some $c_P>0$ and almost all t in $[0,1]$,

$$\forall\xi\in U\quad H\left(t,\psi(t),x(t),\xi\right)-H\left(t,\psi(t),x(t),u(t)\right)\geq\tfrac{1}{2}c_P\|\xi-u(t)\|^2.$$

Note 6.7. For optimal control problems, coercivity conditions like (32c) can be derived from disconjugacy conditions of the Jacobi type [9]; however, the latter conditions are more stringent than the former and are further removed from the second-order necessary conditions for optimality (cf. remarks preceding Theorem 5.2 in [9]).

REFERENCES

[1] J. C. Dunn and T. Tian, *Variants of the Kuhn-Tucker sufficient conditions in cones of nonnegative functions*, SIAM J. Control Optim., 30 (1992), pp. 1361–1384.

[2] A. S. Dontchev, W. W. Hager, A. B. Poore, and B. Yang, *Optimality, stability and convergence in nonlinear control*, June 1, 1992, preprint.

[3] H. Maurer and J. Zowe, *First and second-order necessary and sufficient optimality conditions for infinite-dimensional programming problems*, Math. Programming, 16 (1979), pp. 98–110.

[4] A. Ben-Tal, *Second-order theory of extremum problems*, in Extremal Methods and Systems Analysis, A. V. Fiacco et al. eds, Springer-Verlag, New York, 1980.

[5] A. Ben-Tal and J. Zowe, *A unified theory of first and second order conditions for extremum problems in topological spaces*, Math. Programming Stud., 19 (1982), pp. 39–76.

[6] H. Kawasaki, *An envelope-like effect of infinitely many inequality constraints on second-order necessary conditions for minimization problems*, Math. Programming, 41 (1988), pp. 73–96.

[7] R. Cominetti, *Metric regularity, tangent sets, and second-order optimality conditions*, Appl. Math. Optim., 21 (1990), pp. 265–287.

[8] K. Malanowski, *Sensitivity analysis of optimization problems in Hilbert space, with application to optimal control*, Appl. Math. Optim., 21 (1990), pp. 1–20.

[9] H. Maurer, *First and second order sufficient optimality conditions in mathematical programming and optimal control*, Math. Programming Stud., 14 (1981), pp. 163–177.

[10] A. Ioffe, *On some recent developments in the theory of second order optimality conditions*, Lecture Notes in Mathematics 1405, S. Doleki, ed., Springer-Verlag, New York, 1988.

[11] A. L. DONTCHEV AND W. W. HAGER, *Lipschitzian stability in nonlinear control and optimization*, SIAM J. Control Optim., 31 (1993), pp. 569–603.

[12] T. TIAN, *Convergence Analysis of a Projected Gradient Method for a Class of Optimal Control Problems*, Ph.D. Dissertation, North Carolina State University, May 1992.

[13] T. TIAN AND J. C. DUNN, *On the gradient projection method for optimal control problems with nonnegative* L^2 *inputs*, SIAM J. Control Optim., 32 (1994), pp. 517–537.

[14] H. L. ROYDEN, *Real Analysis*, 2nd ed., Macmillan, New York, 1968.

[15] R. T. ROCKAFELLAR, *Convex Analysis*, Princeton University Press, Princeton, NJ, 1970.

[16] T. TIAN, Private communication.

[17] D. P. BERTSEKAS, *Constrained Optimization and Lagrange Multiplier Methods*, Academic Press, New York, 1982.

[18] V. M. ALEKSEEV, V. M. TIKHOMIROV, AND S. V. FOMIN, *Optimal Control*, Plenum, New York, 1987.

[19] J. C. DUNN, *Gradient-related constrained minimization algorithms in function spaces: Convergence properties and computational implications*, in Large Scale Optimization: State of the Art, W. Hager, D. Hearn, and P. Pardalos, eds., Kluwer Academic Publishers, Dordrecht, 1994.

[20] ———, *On the convergence of projected gradient processes to singular critical points*, J. Optim. Theory Appl., 55 (1987), pp. 203–215.

[21] M. GAWANDE AND J. C. DUNN, *Variable metric gradient projection processes in convex feasible sets defined by nonlinear inequalities*, Applied Math. Optim., 17 (1988), pp. 103–119.

[22] D. G. LUENBERGER, *Linear and Nonlinear Programming*, 2nd ed., Addison–Wesley, Reading, MA, 1984.

[23] ———, *Optimization by Vector Space Methods*, John Wiley, New York, 1969.

[24] R. P. BOAS, *A Primer of Real Functions*, 3rd ed., The Carus Mathematical Monographs 13, Mathematical Association of America, Washington, DC, 1981.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 33, No. 6, pp. 1637–1642, November 1995

© 1995 Society for Industrial and Applied Mathematics

001

RENDEZVOUS SEARCH ON THE LINE WITH INDISTINGUISHABLE PLAYERS*

EDWARD J. ANDERSON† AND SKANDER ESSEGAIER‡

Abstract. Alpern introduced a problem in which two players are placed on the real line at a distance drawn from a bounded distribution F known to both. They can move at maximum velocity one and wish to meet as soon as possible. Neither knows the direction of the other, nor do they have a common notion of a positive direction on the line. It is required to find the *symmetric rendezvous value* $R^s(F)$, which is the minimum expected meeting time achievable by players using the same mixed strategy. This corresponds to the case where the players are indistinguishable; they both take directions from a controller who does not know their names. In this paper we give a mixed strategy which has an expected meeting time of $1.78D + \mu/2$, where D is the maximum of F and μ its mean. This leads to an upper bound $R^s(F) \leq 1.78D + \mu/2$ on the symmetric rendezvous value, which is better than the upper bound $R^s(F) \leq 2D + \mu/2$ obtained by Alpern.

Key words. search games, rendezvous search

AMS subject classifications. 90B40, 90D26

1. Introduction.

The work on rendezvous problems in unbounded domains was initiated by Alpern [Alp], who introduced the following problem. Two players are placed on the real line. They can move at maximum velocity one and wish to meet as soon as possible. We assume that the players know only the probability distribution F of the distance between them at time 0 but not the direction of the other. This problem is related to the linear search problem; see Gal [Gal80] and Beck and Beck [BB92].

The strategy space for both players is the set of paths with maximum speed one:

$$P = \{f : R^+ \to R,\ f(0) = 0,\ |f(s) - f(t)| \leq |s - t|\}.$$

A player placed at point a who chooses strategy f will follow the trajectories $a + f(t)$ and $a - f(t)$ equiprobably.

Since the problem is translation invariant, we may assume that player I starts at point 0 and player II starts equiprobably at $+x$ or $-x$, the initial distance x between the players being drawn from the known cumulative probability distribution F. If player I chooses $f \in P$ and player II chooses $g \in P$, then the expected meeting time $\hat{T}(f, g)$ is given by

$$\hat{T}(f,g) = \int_0^\infty \frac{1}{4} \sum_{i,j=\pm 1} \min\{t : f(t) = ix + jg(t)\}\, dF(x).$$

The ambiguity of the i indicates the equiprobable placement of player II at $+x$ or $-x$, and the ambiguity of the j reflects the assumption that the players cannot tell left from right.

A mixed strategy f^* is a regular Borel probability measure on P. The set of mixed strategies will be denoted by P^*. The payoff $T^*(f^*, g^*)$ is the expected value of $\hat{T}$ in the cross product measure $f^* \times g^*$. The symmetric rendezvous value $R^s(F)$ is then defined as the minimum expected meeting time achievable by players using the same mixed strategies:

$$R^s(F) = \min_{f^* \in P^*} T^*(f^*, f^*).$$

Alpern and Gal [AG] considered the asymmetric case, where the players are distinguishable. This is to say that they have previously agreed which of the roles each will take, and

* Received by the editors December 29, 1993; accepted for publication (in revised form) June 27, 1994.

† Department of Engineering, Cambridge University, Cambridge CB2 1RX, United Kingdom.

‡ London School of Economics, Houghton Street, London WC2A 2AE, United Kingdom.

they are therefore allowed to use different strategies. The asymmetric case has some connection with Beck's work on the linear search problem (see [AG]). We are interested here in the symmetric version, where players are indistinguishable. There is no prior agreement on which of the roles each will take, and we therefore constrain the players to using the same search strategy. This is essentially a one-person decision problem. However, randomization over the set of available pure strategies is necessary in order to have a finite expected meeting time, since there is a probability of at least $1/2$ that players don't meet if they use the same pure strategy. A related problem in which rendezvous takes place on discrete locations is considered by Anderson and Weber [AW90] and in [Tho92].

In the following, we consider the case where the probability distribution F is bounded; i.e., there is a maximal initial distance between the players. Let $D = D_F = \min\{x : F(x) = 1\}$, and let $\mu = \mu_F$, the mean initial distance. Alpern [Alp] obtained an upper bound $R^s(F) \leq 2D + \mu/2$ in terms of D and μ.

In this paper, we describe how results for the symmetric case can be derived from those of the asymmetric one [AG]. We then show how the method can be generalized to obtain upper bounds significantly better than Alpern's $R^s(F) \leq 2D + \mu/2$. We actually give a strategy which has an expected meeting time of $1.78D + \mu/2$ and obtain the upper bound $R^s(F) \leq 1.78D + \mu/2$.

2. Relation with the asymmetric case. The asymmetric case, where the players are distinguishable, has been analyzed by Alpern and Gal [AG]. In this case, the players have previously agreed which of the two roles each will take, and they are therefore allowed to use different strategies. The asymmetric rendezvous value $R^a(F)$ is given by

$$R^a(F) = \min_{f,g \in P} \hat{T}(f, g).$$

In their analysis, Alpern and Gal [AG] showed that if F has a point distribution, then an optimal strategy pair (f_1, f_2) is defined by the following formulae:

$$f_1(t) = \begin{cases} t & \text{if } 0 \leq t \leq D, \\ 2D - t & \text{if } D \leq t \leq 3D, \end{cases} \qquad f_2(t) = \begin{cases} t & \text{if } 0 \leq t \leq D/2, \\ D - t & \text{if } D/2 \leq t \leq 2D, \\ t - 3D & \text{if } 2D \leq t \leq 3D. \end{cases}$$

Furthermore, they showed that this strategy pair guarantees meeting by time $3D$ and that the expected meeting time it achieves is $\hat{T}(f_1, f_2) = (4\mu + 9D)/8$. This provides us with an upper bound on the asymmetric rendezvous value $R^a(F)$. In the case where F is a point distribution ($\mu = D = d$), this upper bound is exactly the asymmetric rendezvous value $R^a = 13d/8$. It is therefore a good upper bound of the form $\alpha D + \beta\mu$.

In the symmetric case, although players use the same *mixed* strategy, there is a positive probability that they actually follow a different *pure* strategy. Suppose, for example, that the players have two pure strategies, s_1 and s_2, and that they choose one or the other equiprobably. Then there is a probability of $1/2$ that one player is using s_1 while the other is using s_2; in other words, there is one chance in two that players end up playing the asymmetric game with the strategy pair (s_1, s_2).

Therefore, one approach to the symmetric problem would be to have the players randomizing over a set of pure strategies carefully chosen to ensure that in the event where they are actually using different *pure* strategies, their expected meeting time is the least possible. We don't expect to solve the symmetric problem in this way, but this approach enables us to make good use of any result in the asymmetric case and to derive some general upper bounds on the symmetric rendezvous value.

Motivated by these remarks, we consider a mixed strategy where, for every period of $3D$ units of time, the players randomize between the strategies f_1 and f_2. More precisely, the strategy proceeds as follows: once and for all, pick a number p in [0, 1]; then, independently every $3D$ cycle, pick either direction to call forward equiprobably, and over the $3D$ units of time period of the cycle move forward by either using f_1 with probability p or using f_2 with probability $1 - p$. So p is picked once, and the forward direction is picked independently every $3D$ cycle.

Suppose that player II is placed at a distance x from player I, where x is drawn from the bounded distribution F. Let T^* be the expected meeting time achieved by this strategy. Because of the symmetry of the situation, we can carry out the calculations by assuming that player II is initially placed at $-x$.

Let A_{kl} represent the expected meeting time if players move in the same direction, with player I using f_l and player II using f_k. Because of the symmetry of the situation, we can carry out the calculations by assuming that player I begins by moving to the left. We have that

$$A = \begin{bmatrix} T^* + 3D & D/2 + x/2 \\ 5D/2 + x/2 & T^* + 3D \end{bmatrix}.$$

Similarly, let B_{kl} be the expected meeting time if the players move away from each other:

$$B = \begin{bmatrix} 2D + x/2 & 3D/2 + x/2 \\ 3D/2 + x/2 & D + x/2 \end{bmatrix}.$$

Finally, it is clear that if players move toward each other, they meet in time $x/2$, and therefore

$$\hat{T}(f_k, f_l) = \int_0^\infty \frac{1}{4}(x/2 + 2A_{kl} + B_{kl})\, dF(x).$$

Writing $p_1 = p$ and $p_2 = 1 - p$, T^* satisfies the equation

$$T^* = \sum_{1 \le k,l \le 2} p_k p_l \hat{T}(f_k, f_l).$$

Solving for T^* we obtain

$$T^*(p) = \frac{6p^2 - 5p + 7}{-4p^2 + 4p + 2} D + \mu/2 \quad \forall p \in [0, 1].$$

This provides us with a general upper bound on the symmetric rendezvous value of the form $\alpha(p)D + \mu/2$. It is easy to see that for $p \in (1/2, 3/7)$, $\alpha(p)$ is strictly less than 2, and therefore one can find a general upper bound of the form $\alpha D + \beta\mu$ better than $2D + \mu/2$.

The attractiveness of this approach lies in that it establishes a relation between the symmetric and the asymmetric case. If we actually minimize $\alpha(p)$ over p, we find that $T^*_{\text{opt}} \approx 1.99404D + \mu/2$. Therefore, the improvement achieved here is not really substantial. We leave for the next section the task of significantly improving Alpern's general upper bound.

3. A better estimate.

PROPOSITION 1. *For any bounded distribution F,*

$$R^s(F) \le 1.78388D + \mu/2.$$

Proof. The two trajectories considered in the previous section have the feature that if the players don't meet by time $3D$, then their distance at time $3D$ is the same as the initial distance.

Suppose that we restrict ourselves to a set of strategies which have this feature, and let the players randomize over this carefully chosen subset. The calculations can then be carried out in the same way as in the above section.

We introduce the following trajectories, defined on $[0, 3D]$:

$$\begin{cases} f_1 = 2F4B, \\ f_2 = 1F3B2F, \\ f_3 = 1F2B1F2B, \\ f_4 = 1F1B1F3B. \end{cases}$$

For example, $1F2B1F2B$ stands for *one step forward, two steps backward, one step forward, two steps backward*: pick either direction to call forward equiprobably; go a distance $D/2$ forward at unit speed; then go a distance $2D/2$ backward at unit speed; then go a distance $D/2$ forward at unit speed; then go a distance $2D/2$ backward at unit speed. More formally we have, for example,

$$1F2B1F2B(t) = \begin{cases} t & \text{if } 0 \le t \le D/2, \\ D - t & \text{if } D/2 \le t \le 3D/2, \\ t - 2D & \text{if } 3D/2 \le t \le 2D, \\ 2D - t & \text{if } 2D \le t \le 3D. \end{cases}$$

Figure 1 shows how the four trajectories appear when position as a function of time is plotted. (Note that a change in the direction labelled forward corresponds to reflection with respect to the horizontal axis.)

The set $\{f_1, f_2, f_3, f_4\}$ has the feature stated above, and we have that

$$A = \begin{bmatrix} T^* + 3D & D/2 + x/2 & D/2 + x/2 & D/2 + x/2 \\ 5D/2 + x/2 & T^* + 3D & 5D/2 + x/2 & 5D/2 + x/2 \\ T^* + 3D & 3D/2 + x/2 & T^* + 3D & T^* + 3D \\ T^* + 3D & D + x/2 & D + x/2 & T^* + 3D \end{bmatrix}$$

and

$$B = \begin{bmatrix} 2D + x/2 & 3D + x/2 & 2D + x/2 & 2D + x/2 \\ 3D/2 + x/2 & D + x/2 & D + x/2 & 3D2 + x/2 \\ 2D + x/2 & D + x/2 & D + x/2 & 2D + x/2 \\ 2D + x/2 & 3D/2 + x/2 & 2D + x/2 & 2D + x/2 \end{bmatrix},$$

and therefore

$$\hat{T}(f_k, f_l) = \int_0^\infty \frac{1}{4}(x/2 + 2A_{kl} + B_{kl})\, dF(x).$$

Suppose that players choose a probability distribution (p_1, p_2, p_3, p_4), and for each $3D$ units of time, follow f_i with probability p_i. T^* is then the solution of the equation

$$T^* = \sum_{1 \le k, l \le 4} p_k p_l \hat{T}(f_k, f_l).$$

Solving for T^* we obtain, for a given set of probabilities $(p_1, p_2, p_3, 1 - \sum_{1 \le k \le 3} p_k)$,

$$T^*(p_1, p_2, p_3) = \alpha(p_1, p_2, p_3)D + \mu/2,$$

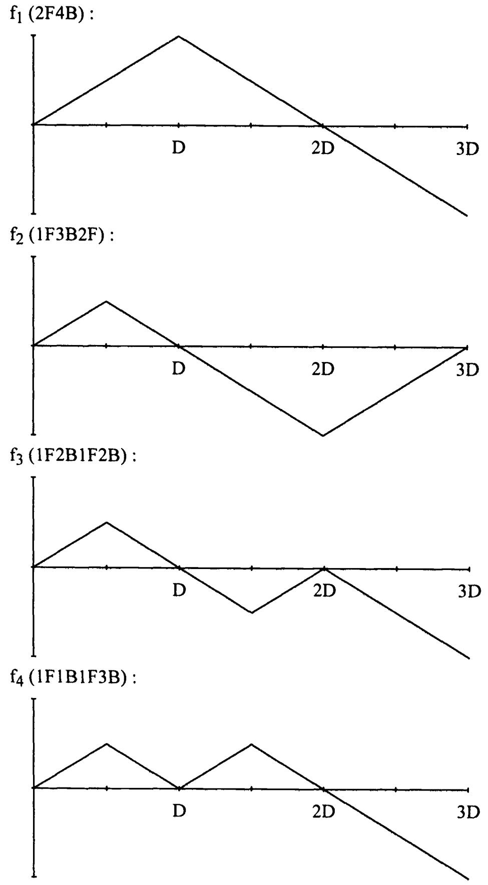


FIG. 1. *Four trajectories for the rendezvous search.*

where

$$(1)\quad \alpha(p_1,p_2,p_3)=\frac{-8+4p_1-3p_1^2+6p_2-4p_1p_2-5p_2^2+5p_3-4p_1p_3-4p_2p_3-5p_3^2}{2(-1-p_1+p_1^2-2p_2+p_1p_2+2p_2^2-p_3+p_1p_3+p_2p_3+p_3^2)}.$$

Taking $p_1 = 0.227917$, $p_2 = 0.404018$, and $p_3 = 0.212053$ and substituting these values in (1), we find that $\alpha(p_1,p_2,p_3) \approx 1.78388$, and this proves the result. We actually used Mathematica to minimize $\alpha(p_1,p_2,p_3)$ subject to the constraints $0 \le p_i \le 1$. The values of the p_i determine a (local) minimum. □

A particular simple version of the rendezvous linear search problem is when the two players know the initial distance (say, 1) between them but neither knows the direction of the other. In this case F is a point distribution and $D_F = \mu_F = 1$, so the above estimate gives $R^s \le 2.28388$. This refutes a conjecture by Alpern [Alp] that the rendezvous value for this problem is $5/2$.

REFERENCES

[AG] S. ALPERN AND S. GAL, *Rendezvous search on the line with distinguishable players*, SIAM J. Control Optim., 33 (1995), pp. 1271–1277.

[Alp] S. Alpern, *The rendezvous search problem*, SIAM J. Control Optim., 33 (1995), pp. 673–683.

[AW90] E. J. Anderson and R. R. Weber, *The rendezvous problem on discrete locations*, J. Appl. Probab., 28 (1990), pp. 839–851.

[BB92] A. Beck and M. Beck, *The revenge of the linear search problem*, SIAM J. Control Optim., 30 (1992), pp. 112–122.

[Gal80] S. Gal, *Search Games*, Academic Press, New York, 1980.

[Tho92] L. C. Thomas, *Finding your kids when they are lost*, J. Operational Research Society, 43 (1992), pp. 637–639.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 33, No. 6, pp. 1643–1660, November 1995

© 1995 Society for Industrial and Applied Mathematics
002

OPTIMAL PROGRAMS ON INFINITE HORIZON 1*

A. J. ZASLAVSKI†

Abstract. We consider the limit behavior, as $N \to \infty$, of the expression $\sum_{i=0}^{N-1} v(x_i, x_{i+1})$ for programs $\{x_i\}_{i=0}^{\infty}$ in a compact metric space K where v is a real-valued function defined on $K \times K$. We study the structure of (v)-good programs and establish the existence of a G_δ set F in $\mathbf{C}(K \times K)$ such that, for each $u \in F$, all (u)-good programs have the same limit points set and also, for every $x \in K$, there exists a (u)-optimal program $\{x_i\}_{i=0}^{\infty}$ satisfying $x_0 = x$.

Key words. good program, minimal-energy configuration, overtaking optimality, (v)-weakly optimal program

AMS subject classification. 49J99

Introduction. In this paper we consider the infinite-horizon problem of minimizing the expression $\sum_{i=0}^{N-1} v(x_i, x_{i+1})$ as N grows to infinity, where $\{x_i\}_{i=0}^{\infty}$ is a sequence in a compact metric space K and v is a continuous function defined on $K \times K$. This provides a convenient setting for the study of various optimization problems, e.g., continuous-time control systems which are represented by ordinary differential equations whose cost integrand contains a discounting factor [1], the infinite-horizon deterministic control problem of minimizing $\int_0^T L(z, \dot{z})\,dt$ as $T \to \infty$ [2], the analysis of a long slender bar of a polymeric material under tension [3], and the analysis of an infinite discrete model for crystals which undergo phase transitions [4], [5].

The continuous-time case can be reduced to this framework in the following manner. A control system is operating on an infinite-time interval $[0, \infty)$. We choose a sampling time interval, say, $[0, T]$. For any action that steers the state $x \in K$ at time $t = 0$ to state $y \in K$ at time $t = T$, there is an associated cost. The value $v(x, y)$ is the minimal cost possible. Any choice of a control generates a trajectory, say, $x(t)$, and we will refer to $z_k = x(kT)$ as a program. If the control action is chosen in an optimal way on finite intervals, the cost of the program $\{z_k\}$ at time $t = NT$ is $\sum_{i=0}^{N-1} v(z_i, z_{i+1})$.

A hidden assumption is that $v(x, y)$ is finitely defined on $K \times K$, namely, a controllability-type assumption. Another assumption is that v is time invariant; therefore, the original control problem is, in general, either stationary or T-periodic. The continuity of v holds for many problems. The same is true for the compactness assumption; namely, in many examples one can show that all the reasonable solutions occur in a prescribed compact set.

Let K be a compact metric space, $\mathbf{R}^n$ be the Euclidean n-dimensional space, $\mathbf{C}(K \times K)$ be the space of all continuous functions $v : K \times K \to \mathbf{R}^1$ with the topology of the uniform convergence ($\|v\| = \sup\{|v(x, y)| : x, y \in K\}$). Let $\mathbf{C}(K)$ be the space of all continuous functions $v : K \to \mathbf{R}^1$ with the topology of the uniform convergence ($\|v\| = \sup\{|v(x)| : x \in K\}$) and $B(K \times K)$ be the set of all bounded and lower semicontinuous functions $v : K \times K \to \mathbf{R}^1$ (i.e., $v(\lim(x_k, y_k)) \le \liminf v(x_k, y_k)$).

Consider any $v \in B(K \times K)$. We are interested in the limit behavior as $N \to \infty$ of the expression $\sum_{i=0}^{N-1} v(x_i, x_{i+1})$, where $\{x_i\}_{i=0}^{\infty}$ is an infinite sequence in K which we call a *program* (or a *configuration*) (see [1], [4], [5]) and which occasionally will be denoted by a boldface $\mathbf{x}$. (Similarly $\{y_i\}_{i=0}^{\infty}$ will be denoted by $\mathbf{y}$, etc.) A finite sequence $\{x_i\}_{i=0}^{N} \subset K$ $(N = 0, 1, \ldots)$ will be also called a program. We shall define three concepts of optimality.

* Received by the editors October 7, 1993; accepted for publication July 8, 1994. This research was supported in part by a grant from the Israeli Ministry of Science and the MA-AGARA (a special project for the absorption of new immigrants) in the Department of Mathematics, Technion.

† Department of Mathematics, Technion–Israel Institute of Technology, 32000 Haifa, Israel.

A program $\{x_i\}_{i=0}^{\infty}$ is a (v)-*overtaking optimal program* if for every program $\{z_i\}_{i=0}^{\infty}$ satisfying $z_0 = x_0$ the following inequality holds:

$$\limsup_{N\to\infty} \sum_{i=0}^{N-1} [v(x_i, x_{i+1}) - v(z_i, z_{i+1})] \leq 0.$$

This notion, known as the *overtaking optimality criterion*, was introduced in the economic literature by Gale [6] and von Weizsäcker [7] and was employed to study the infinite-horizon control problems [1], [8]–[10].

A program $\{x_i\}_{i=0}^{\infty}$ is (v)-*weakly optimal* [1], [6], [7] if for every program $\{z_i\}_{i=0}^{\infty}$ satisfying $z_0 = x_0$ the following inequality holds:

$$\liminf_{N\to\infty} \sum_{i=0}^{N-1} [v(x_i, x_{i+1}) - v(z_i, z_{i+1})] \leq 0.$$

A sequence $\{x_i\}_{i=-\infty}^{\infty} \subset K$ is called a (v)-*minimal energy configuration* (*program*) if for each N, $M > 0$ the inequality

$$\sum_{i=-N}^{M-1} v(x_i, x_{i+1}) \leq \sum_{i=-N}^{M-1} v(z_i, z_{i+1})$$

holds for every sequence $\{z_i\}_{i=-N}^{M} \subset K$ satisfying $x_{-N} = z_{-N}$, $x_M = z_M$ [3]–[5].

Of special interest is the *minimal long-run average cost growth rate*,

$$\mu(v) = \inf \left\{ \liminf_{N\to\infty} N^{-1} \sum_{i=0}^{N-1} v(z_i, z_{i+1}) : \{z_i\}_{i=0}^{\infty} \text{is a program} \right\}.$$

A program $\{z_i\}_{i=0}^{\infty}$ is called a (v)-*good program* [1] if the sequence $\{\sum_{i=0}^{N-1} [v(z_i, z_{i+1}) - \mu(v)]\}_{N=1}^{\infty}$ is bounded. It was proved in [1] that for every program $\{z_i\}_{i=0}^{\infty}$ the sequence $\{\sum_{i=0}^{N-1} [v(z_i, z_{i+1}) - \mu(v)]\}_{N=1}^{\infty}$ either is bounded or diverges to infinity and that for every initial value z there is a (v)-good program $\{z_i\}_{i=0}^{\infty}$ satisfying $z_0 = z$. In [1] the following representation formula valid for every $v \in \mathbf{C}(K \times K)$ was also established:

$$v(x,y) = \theta^v(x,y) + \mu(v) - \pi^v(y) + \pi^v(x) \quad (x, y \in K),$$

where π^v, θ^v are continuous functions, θ^v is nonnegative, and $E(x) = \{y \in K : \theta^v(x,y) = 0\}$ is nonempty for every $x \in K$. (π^v, θ^v are calculated directly through v and $\mu(v)$.)

In this paper we study the structure of (v)-good programs and establish for a generic $v \in \mathbf{C}(K \times K)$, for every given $x \in K$, the existence of (v)-optimal program $\{x_i\}_{i=0}^{\infty}$ satisfying $x_0 = x$. For the latter subject we should choose a proper optimality criterion. In spite of the example $v \in C([-1,1] \times [-1,1])$ given in [1] which demonstrates the nonexistence of (v)-weakly optimal programs, the notion of (v)-weak optimality is the most suitable one to meet our goal when we consider generic functions $v \in \mathbf{C}(K \times K)$.

We establish the existence of a set $F \subset \mathbf{C}(K \times K)$ which is a countable intersection of open everywhere dense sets in $\mathbf{C}(K \times K)$ such that for every $u \in F$ the following propositions hold:

a) There exist closed sets $H(u) \subset K \times K$, $H_0(u) \subset K$ such that for every (u)-good program $\{x_i\}_{i=0}^{\infty}$ the limit points set of $\{x_i\}_{i=0}^{\infty}$ is $H_0(u)$ and the limit points set of $\{(x_i, x_{i+1})\}_{i=0}^{\infty}$ is $H(u)$.

b) The set $H(u)$ is approximated by finite periodic programs.

c) For every initial point $x \in K$ there exists a (u)-weakly optimal program $\{x_i\}_{i=0}^{\infty}$ satisfying $x_0 = x$, $\theta^u(x_i, x_{i+1}) = 0$ $(i = 0, 1, \ldots)$.

The paper is organized as follows. In §1 we give the necessary definitions and state precisely our results. In §2 we prove the preliminary lemmas and develop the suitable technique which is used in §3 to prove the theorems. Two examples are given in §4: one concerning the nonexistence of a (v)-overtaking optimal program for each v belonging to some open set $D \subset C([0,1] \times [0,1])$ and the other showing the existence of a (v)-weakly optimal program $\{x_i\}_{i=0}^{\infty}$ for which the relation $\theta^v(x_i, x_{i+1}) = 0$ $(i = 0, 1, \ldots)$ does not hold.

1. Definitions and theorems. Let K be a compact metric space, $v \in B(K \times K)$. We define

$$(1.1)\qquad a(v) = \sup\{v(x,y) : x, y \in K\}, \qquad b(v) = \inf\{v(x,y) : x, y \in K\},$$

$$(1.2)\qquad \mu(v) = \inf\left\{\liminf N^{-1} \sum_{i=0}^{N-1} v(z_i, z_{i+1}) : \{z_i\}_{i=0}^{\infty} \text{ is a program}\right\},$$

(1.3)

$$\lambda(N, v) = \min\left\{N^{-1} \sum_{i=0}^{N-1} v(z_i, z_{i+1}) : \{z_i\}_{i=0}^{N} \subset K,\ z_0 = z_N\right\} \qquad (N = 1, 2, \ldots),$$

$$(1.4)\quad \rho(N, v) = \min\left\{N^{-1} \sum_{i=0}^{N-1} v(z_i, z_{i+1}) : \{z_i\}_{i=0}^{N} \subset K\right\} \qquad (N = 1, 2, \ldots).$$

The following two results established in [1] are very useful in the study of infinite-horizon control problems. They were mentioned and explained in the introduction, but we need their exact formulations.

THEOREM 1 [1]. 1. $\rho(N, v) \le \mu(v) \le \lambda(N, v)$, $N(\lambda(N, v) - \rho(N, v)) \le a(v) - b(v)$ $(N = 1, 2, \ldots)$.

2. *For every program* $\{z_i\}_{i=0}^{\infty}$

$$\sum_{i=0}^{N-1} [v(z_i, z_{i+1}) - \mu(v)] \ge b(v) - a(v) \qquad (N = 1, 2, \ldots).$$

3. *For every program* $\{z_i\}_{i=0}^{\infty}$ *the sequence* $\{\sum_{i=0}^{N-1} [v(z_i, z_{i+1}) - \mu(v)]\}_{N=1}^{\infty}$ *either is bounded or diverges to infinity.*

4. *For every initial value* z_0 *there is a program* $\{z_i\}_{i=0}^{\infty}$ *which satisfies*

$$\left|\sum_{i=0}^{N} [v(z_i, z_{i+1}) - \mu(v)]\right| \le 4|a(v) - b(v)| \qquad (N = 1, 2, \ldots).$$

THEOREM 2 [1]. *Let* $v \in \mathbf{C}(K \times K)$, *and define*

$$(1.5)\qquad \pi^v(x) = \inf\left\{\liminf \sum_{i=0}^{N-1} [v(z_i, z_{i+1}) - \mu(v)] : z \subset K,\ z_0 = x\right\},$$

$$\theta^v(x,y) = v(x,y) - \mu(v) + \pi^v(y) - \pi^v(x) \tag{1.6}$$

for x, $y \in K$. Then π^v, θ^v are continuous functions, θ^v is nonnegative, and

$$E(x) = \{y \in K : \theta^v(x,y) = 0\}$$

is nonempty for every $x \in K$.

In [1] these theorems were established when K was a compact in $\mathbf{R}^n$, but their proofs remain in force also when K is a compact metric space.

For a program $\mathbf{x}$ we denote by $\omega(\mathbf{x})$ the set of all points $z \in K$ such that some subsequence $\{x_{i_k}\}_{k=1}^{\infty}$ converges to z and by $\Omega(\mathbf{x})$ the set of all points $(z_1, z_2) \in K \times K$ such that some subsequence $\{(x_{i_k}, x_{i_k+1})\}_{k=1}^{\infty}$ converges to (z_1, z_2). Denote by $d(x,y)$ $(x, y \in K)$ the metric in K, and define the metric d_1 on $K \times K$ by

$$d_1((x_1,x_2),\ (y_1,y_2)) = d(x_1,y_1) + d(x_2,y_2)\ (x_1,x_2,y_1,y_2 \in K).$$

We denote $d(x,B) = \inf\{d(x,y) : y \in B\}$ for $x \in K$, $B \subset K$ and

$$d_1((x_1,x_2),A) = \inf\{d_1((x_1,x_2),\ (y_1,y_2)) : (y_1,y_2) \in A\}$$

for $(x_1,x_2) \in K \times K$ and $A \subset K \times K$.

Denote the Hausdorff metric for two sets $A \subset K$ and $B \subset K$ by $\operatorname{dist}(A,B)$ and the cardinality of a set A by $\operatorname{Card}(A)$.

A sequence $\{x_i\}_{i=-\infty}^{\infty} \subset K$ is called *almost periodic* if for every $\varepsilon > 0$ there exists an integer $m \geq 1$ such that the relation $d(x_i, x_{i+pm}) \leq \varepsilon$ holds for any i and any p.

A program $\{x_i\}_{i=0}^{\infty}$ is called *asymptotic almost periodic* if for every $\varepsilon > 0$ there exist integers $k \geq 1$, $m \geq 1$ such that $d(x_i, x_{i+mj}) \leq \varepsilon$ for any $i \geq k$ and any $j \geq 1$.

In this paper we prove the existence of a set $F \subset \mathbf{C}(K \times K)$ which is a countable intersection of open everywhere dense sets in $\mathbf{C}(K \times K)$ and for which the following theorems are valid.

THEOREM 3. 1. *For every $u \in F$ there are closed sets $H(u) \subset K \times K$, $H_0(u) \subset K$ such that for every (u)-good program $\mathbf{x}$ we have*

$$H(u) = \Omega(\mathbf{x}), \qquad H_0(u) = \omega(\mathbf{x}).$$

2. *Let $u \in F$. Then every (u)-good program $\mathbf{x}$ is asymptotic almost periodic.*

3. *Let $u \in F$ and δ be a positive number. Then there is a neighborhood $W(u)$ of u in $\mathbf{C}(K \times K)$ such that for every $w \in W(u)$ for every (w)-good program $\mathbf{x}$ we have* $\operatorname{dist}(H(u), \Omega(\mathbf{x})) \leq \delta$.

Assertion 1 of Theorem 3 establishes that for $u \in F$ all the sequences $\{(x_i, x_{i+1})\}_{i=0}^{\infty}$, where $\{x_i\}_{i=0}^{\infty}$ is a (u)-good program, have the same limit points set denoted by $H(u)$. Assertion 2 means that for $u \in F$ every (u)-good program is asymptotic almost periodic, and assertion 3 of Theorem 3 shows that for every w belonging to a small neighborhood of u, for every (w)-good program $\mathbf{x}$ the set $\Omega(\mathbf{x})$ is close enough to $H(u)$ in the Hausdorff metric (here $u \in F$). If we think of $H(u)$ as an analogue of a turnpike set (see [11], [12]), assertion 3 means stability of the turnpike phenomena.

THEOREM 4. *Let $u \in F$, $\{x_i\}_{i=0}^{\infty}$ be a program such that $\theta^u(x_i, x_{i+1}) = 0$ $(i = 0, 1, \ldots)$. Then $\{x_i\}_{i=0}^{\infty}$ is a (u)-weakly optimal program. Moreover, there exists a subsequence $\{x_{i_k}\}_{k=1}^{\infty}$ such that for every program $\{y_i\}_{i=0}^{\infty}$ satisfying $y_0 = x_0$ the inequality*

$$\liminf_{k \to \infty} \sum_{j=0}^{i_k - 1} [u(y_j, y_{j+1}) - u(x_j, x_{j+1})] \geq 0$$

holds, and if for some program $\{y_i\}_{i=0}^{\infty}$ *satisfying* $y_0 = x_0$,

$$\liminf_{k\to\infty} \sum_{j=0}^{i_k-1} [u(y_j, y_{j+1}) - u(x_j, x_{j+1})] = 0,$$

then $\theta^u(y_j, y_{j+1}) = 0\,(j = 0, 1, \ldots)$.

Theorem 4 establishes that for every $u \in F$, for every initial value $x \in K$, there exists a (u)-weakly optimal program $\{x_i\}_{i=0}^{\infty}$ satisfying $x_0 = x$, $\theta^u(x_i, x_{i+1}) = 0\,(i = 0, 1, \ldots)$. This theorem also implies that if $u \in F$ and $\{y_i\}_{i=0}^{\infty}$ is a (u)-overtaking optimal program, then $\theta^u(y_i, y_{i+1}) = 0\,(i = 0, 1, \ldots)$.

In [1] an example of $v \in C([-1, 1] \times [-1, 1])$ which demonstrates the nonexistence of (v)-weakly optimal programs was given. Here we will give an example of an open set $D \subset C([0, 1] \times [0, 1])$ such that for every $v \in D$ there is not any (v)-overtaking optimal program. We will also give an example of $v \in C([0, 1] \times [0, 1])$ for which there exist a (v)-weakly optimal program $\{x_i\}_{i=0}^{\infty}$ and a (v)-minimal energy configuration $\{y_i\}_{i=-\infty}^{\infty}$ such that

$$\sup\{\theta^v(x_i, x_{i+1}) : i = 0, 1, \ldots\} > 0, \qquad \theta^v(y_0, y_1) > 0.$$

2. Preliminary lemmas.

LEMMA 1. *Let* $u \in \mathbf{C}(K \times K)$, $r > 0$. *Then there exist a nonnegative function* $\theta \in \mathbf{C}(K \times K)$, *an integer* $m \geq 1$, *and a sequence* $\{x_i\}_{i=0}^{m} \subset K$ *such that* $x_0 = x_m$, $\theta(x_i, x_{i+1}) = 0\,(i = 0, \ldots, m-1)$, $\|v - u\| \leq r$, *where* $v(x, y) = \mu(u) + \pi^u(x) - \pi^u(y) + \theta(x, y)(x, y \in K)$.

Proof. The uniform continuity of θ^u on $K \times K$ implies the existence of a number $\delta > 0$ such that $|\theta^u(x_1, x_2) - \theta^u(y_1, y_2)| \leq 2^{-1}r$ for each (x_1, x_2) and $(y_1, y_2) \in K \times K$ satisfying $d_1((x_1, x_2), (y_1, y_2)) \leq 2\delta$. We consider a program $\{y_i\}_{i=0}^{\infty}$ such that $\theta^u(y_i, y_{i+1}) = 0\,(i = 0, 1, \ldots)$. It is easy to see that there exist $i \in \{0, 1, \ldots\}$ and $m \in \{1, 2, \ldots\}$ for which $d(y_i, y_{i+m}) < 2^{-1}\delta$. Without loss of generality, suppose that $i = 0$. There exists a continuous function $F : K \times K \to [0, 1]$ such that $F(y_{m-1}, y_0) = 0$, $F(y, z) = 1$ for y, $z \in K$, satisfying $d_1((y, z), (y_{m-1}, y_0)) \geq 3 \cdot 4^{-1}\delta$. We define $\theta(x, y) = \theta^u(x, y)F(x, y)\,(x, y \in K)$, $x_i = y_i\,(i = 0, \ldots, m-1)$, $x_m = y_0$. To complete the proof we should note only that $\|\theta - \theta^u\| \leq r$.

LEMMA 2. *Let* $u \in \mathbf{C}(K \times K)$, $r > 0$. *Then there are an integer* $m \geq 1$, *a sequence* $\{x_i\}_{i=0}^{m} \subset K$, *and a nonnegative function* $\theta \in \mathbf{C}(K \times K)$ *such that*

1. *for integers* i, j *satisfying* $0 \leq i < j \leq m$ *the equality* $x_i = x_j$ *holds if and only if* $i = 0$, $j = m$;
2. *for* x, $y \in K$ *the equality* $\theta(x, y) = 0$ *holds if and only if*

$$(x, y) \in \{(x_i, x_{i+1}) : i = 0, \ldots, m-1\};$$

3. $\|v - u\| \leq r$, *where*

$$v(x, y) = \mu(u) + \pi^u(x) - \pi^u(y) + \theta(x, y) \qquad (x, y \in K).$$

Proof. By Lemma 1, there exist a nonnegative function $\theta_1 \in \mathbf{C}(K \times K)$, an integer $m \geq 1$, and a sequence $\{x_i\}_{i=0}^{m} \subset K$ such that $x_0 = x_m$, $\theta_1(x_i, x_{i+1}) = 0\,(i = 0, \ldots, m-1)$, $\|\theta^u - \theta_1\| \leq 2^{-1}r$. Without loss of generality, we suppose that for integers i, j satisfying $0 \leq i < j \leq m$ the equality $x_i = x_j$ holds if and only if $i = 0$, $j = m$. Define

$$\phi(x, y) = \prod_{i=0}^{m-1} d_1((x, y), (x_i, x_{i+1})) \qquad (x, y \in K).$$

Choose $\gamma > 0$ for which $\gamma\|\phi\| \leq 4^{-1}r$, and set $\theta = \theta_1 + \gamma\phi$, $v(x,y) = \mu(u) + \pi^u(x) - \pi^u(y) + \theta(x,y)$ $(x, y \in K)$. The lemma is proved.

For a number $r > 0$ and a point x of some metric space we denote the open (closed) ball in this space which has the center x and the radius r by $B(x,r)(\bar{B}(x,r))$.

From now on in this section we consider a fixed function $v \in \mathbf{C}(K \times K)$ for which there exist an integer $m \geq 1$, a sequence $\{x_i^*\}_{i=0}^m \subset K$, a number μ, and functions $\pi \in \mathbf{C}(K)$, $\theta \in \mathbf{C}(K \times K)$ such that

1. for integers i, j satisfying $0 \leq i < j \leq m$, the equality $x_i^* = x_j^*$ holds if and only if $i = 0$, $j = m$;

2. $\|\theta\| > 0$, θ is nonnegative, and for each x and $y \in K$ the equality $\theta(x,y) = 0$ holds if and only if $(x,y) \in \{(x_i^*, x_{i+1}^*) : i = 0, \ldots, m-1\}$;

3. $v(x,y) = \mu + \pi(x) - \pi(y) + \theta(x,y)$ $(x, y \in K)$.

We denote by E the set of all such functions v. Lemma 2 implies that the set E is dense everywhere in $\mathbf{C}(K \times K)$. We define $x_i^* \in K$ for $i \in \{0, \pm 1, \ldots\} \setminus \{0, \ldots, m\}$ such that $x_{m+i}^* = x_i^*$ $(i = 0, \pm 1, \ldots)$. For every number $\delta > 0$ we define

$$C_1(\delta) = \sup\left\{\theta(x,y) : (x,y) \in \bigcup_{i=0}^{m-1} \bar{B}((x_i^*, x_{i+1}^*), \delta)\right\}, \tag{2.1}$$

$$C_2(\delta) = \inf\left\{\theta(x,y) : (x,y) \in (K \times K) \setminus \bigcup_{i=0}^{m-1} B((x_i^*, x_{i+1}^*), \delta)\right\}, \tag{2.2}$$

$$C_3(\delta) = \sup\{|\pi(x) - \pi(y)| : x, y \in K,\ d(x,y) \leq \delta\}. \tag{2.3}$$

We define

$$D_0 = 8^{-1} \inf\{d(x_i^*, x_j^*) : i, j \in \{0, \ldots, m-1\},\ i < j\} \tag{2.4}$$

(if $m = 1$, then $D_0 = +\infty$).

LEMMA 3. *Let* $\delta \in (0, D_0)$, $r \in (0, (48m)^{-1}C_2(\delta))$, $\phi \in \mathbf{C}(K \times K)$, $\|\phi\| \leq r$, $u = v + \phi$, $N \in \{1, 2, \ldots\}$, *and* $\{z_i\}_{i=0}^N$ *be a program such that for every program* $\{y_i\}_{i=0}^N$ *satisfying* $y_0 = z_0$, $y_N = z_N$ *the following relation holds*:

$$\sum_{i=0}^{N-1} u(z_i, z_{i+1}) \leq 2\|\theta\| + \sum_{i=0}^{N-1} u(y_i, y_{i+1}).$$

Let $i_0, j_0 \in \{1, 2, \ldots\}$, $0 < i_0 < j_0 < N$, $j_0 - i_0 \geq 240\|\theta\|C_2(\delta)^{-1}m$. *Then there exists an integer* $k \in \{i_0, \ldots, j_0 - 6m\}$ *for which* $d(z_{k+i}, x_i^*) \leq \delta$ $(i = 0, \ldots, 3m)$.

Proof. Assume that there exists an integer $k \in \{i_0, \ldots, j_0 - 6m\}$ for which

$$(z_j, z_{j+1}) \in \bigcup_{p=0}^{m-1} B((x_p^*, x_{p+1}^*), \delta) \qquad (j = k, \ldots, k + 4m - 1).$$

Then for every integer $j \in \{k, \ldots, k + 4m - 1\}$ there exists an integer $p(j) \in \{0, \ldots, m-1\}$ such that

$$(z_j, z_{j+1}) \in B((x_{p(j)}^*, x_{p(j)+1}^*), \delta).$$

Let $j \in \{k, \dots, k+4m-2\}$. It is easy to see that

$$d(z_{j+1}, x^*_{p(j)+1}) \le d_1((z_j, z_{j+1}), (x^*_{p(j)}, x^*_{p(j)+1})) \le \delta,$$
$$d(z_{j+1}, x^*_{p(j+1)}) \le d_1((z_{j+1}, z_{j+2}), (x^*_{p(j+1)}, x^*_{p(j+1)+1})) \le \delta,$$
$$d(x^*_{p(j)+1}, x^*_{p(j+1)}) \le 2\delta.$$

It follows from the definition of δ and D_0 (see (2.4) and the conditions of Lemma 3) that

$$x^*_{p(j)+1} = x^*_{p(j+1)} \quad \text{for all } j \in \{k, \dots, k+4m-2\}.$$

It follows from the definition of v (see condition 1) that for every integer $j \in \{k, \dots, k+4m-2\}$

$$p(j+1) = p(j) + 1 \quad \text{if } p(j) < m-1$$

and

$$p(j+1) = 0 \quad \text{if } p(j) = m-1.$$

Together with the definition of $\{x^*_i\}_{i=-\infty}^{\infty}$ and $\{p(j)\}_{j=k}^{k+4m-1}$ this implies that for every $i \in \{0, \dots, 4m-1\}$ the number $m^{-1}[p(k+i) - p(k) - i]$ is an integer and

$$(z_{k+i}, z_{k+i+1}) \in B((x^*_{p(k)+i}, x^*_{p(k)+i+1}), \delta).$$

To prove the lemma it remains now to show that there is an integer $k \in \{i_0, \dots, j_0 - 6m\}$ for which

$$(z_j, z_{j+1}) \in \bigcup_{p=0}^{m-1} B((x^*_p, x^*_{p+1}), \delta) \qquad (j = k, \dots, k+4m-1).$$

Let us assume the opposite. Then for every $k \in \{i_0, \dots, j_0 - 6m\}$ there is $j(k) \in \{k, \dots, k+4m\}$ such that

$$(z_{j(k)}, z_{j(k)+1}) \notin \bigcup_{i=0}^{m-1} B((x^*_i, x^*_{i+1}), \delta),$$

and by (2.2)

$$\theta(z_{j(k)}, z_{j(k)+1}) \ge C_2(\delta) \qquad (k \in \{i_0, \dots, j_0 - 6m\}). \tag{2.5}$$

Consider a program $\{y_i\}_{i=0}^N$ such that $y_i = z_i$ $(i \in \{0, \dots, i_0 - 1\} \cup \{j_0 + 1, \dots, N\})$, $y_i = x^*_i (i \in [i_0, j_0])$. It follows from the conditions of Lemma 3, the definition of v and $\{y_i\}_{i=0}^N$, and (2.2), (2.5) that

$$-2\|\theta\| \le \sum_{i=0}^{N-1} [u(y_i, y_{i+1}) - u(z_i, z_{i+1})] = \sum_{i=0}^{N-1} [(v+\phi)(y_i, y_{i+1}) - (v+\phi)(z_i, z_{i+1})]$$
$$= \sum_{i=0}^{N-1} [(\theta+\phi)(y_i, y_{i+1}) - (\theta+\phi)(z_i, z_{i+1})]$$
$$\le \sum_{i=i_0}^{j_0-1} [(\theta+\phi)(y_i, y_{i+1}) - (\theta+\phi)(z_i, z_{i+1})] + 2\|\theta\| + 4\|\phi\|$$

$$\begin{aligned}
&\le 2\|\theta\| + 4\|\phi\| + 2\|\phi\|(j_0 - i_0) - \sum_{i=i_0}^{j_0-1} \theta(z_i, z_{i+1}) \\
&\le 2\|\theta\| + 2\|\phi\|(2 + j_0 - i_0) - C_2(\delta)(j_0 - i_0 - 12m)(6m)^{-1} \\
&\le 3\|\theta\| + (2\|\phi\| - C_2(\delta)(12m)^{-1})(j_0 - i_0) \\
&\le 3\|\theta\| - (j_0 - i_0)C_2(\delta)(24m)^{-1} \le -7\|\theta\|.
\end{aligned}$$

The contradiction obtained proves the lemma.

For an integer $N \ge 1$, $\phi \in \mathbf{C}(K \times K)$ we define

$$\ell(\phi, N) = \inf\left\{\sum_{i=0}^{N-1} (\phi + \theta)(y_i, y_{i+1}) : \{y_i\}_{i=0}^{N} \subset K\right\}. \tag{2.6}$$

LEMMA 4. *Let $\delta \in (0, D_0)$. Then there exists a number $r > 0$ such that for every $\phi \in \mathbf{C}(K \times K)$, every integer $N \ge 1$ and every program $\{y_i\}_{i=0}^{N}$ satisfying $\|\phi\| \le r$, $\sum_{i=0}^{N-1} (\phi + \theta)(y_i, y_{i+1}) \le \ell(\phi, N) + r$, there are integers $k(1)$, $k(2)$ for which the following relations hold*:

$$d_1((y_i, y_{i+1}), (x^*_{i-N+k(1)}, x^*_{i-N+k(1)+1})) \le \delta \qquad (i \in [N - 3m, N - 1] \cap \{0, 1, \dots\}), \tag{2.7}$$

$$d_1((y_i, y_{i+1}), (x^*_{i+k(2)}, x^*_{i+k(2)+1})) \le \delta \qquad (i \in [0, \min\{3m, N - 1\}]). \tag{2.8}$$

Proof. We choose numbers $\delta_0 \in (0, \delta)$, $r > 0$ such that

$$8^{-1}C_2(\delta) \ge C_1(\delta_0), \qquad r < (4 \cdot 10^3 m\|\theta\|)^{-1} C_2(2^{-1}\delta_0) C_2(\delta_0).$$

Let $\theta \in \mathbf{C}(K \times K)$, $\|\phi\| \le r$, $N \ge 1$, $\{y_i\}_{i=0}^{N} \subset K$, and

$$\sum_{i=0}^{N-1} (\phi + \theta)(y_i, y_{i+1}) \le \ell(\phi, N) + r. \tag{2.9}$$

There are two cases:

1. $N \le 500m\|\theta\|C_2(2^{-1}\delta_0)^{-1}$;
2. $N > 500m\|\theta\|C_2(2^{-1}\delta_0)^{-1}$.

Consider the first case. Our choice of r and (2.9) imply

$$\begin{aligned}
\sum_{i=0}^{N-1} \theta(y_i, y_{i+1}) &\le \sum_{i=0}^{N-1} (\phi + \theta)(y_i, y_{i+1}) + Nr \le \ell(\phi, N) + (N + 1)r \\
&\le (N + 1)r + \sum_{i=0}^{N-1} (\phi + \theta)(x^*_i, x^*_{i+1}) \le (2N + 1)r \le 2^{-1}C_2(\delta_0),
\end{aligned}$$

$$(y_i, y_{i+1}) \in \bigcup_{j=0}^{m-1} B((x^*_j, x^*_{j+1}), \delta_0) \qquad (i = 0, \dots, N - 1).$$

Now it is easy to see that in the first case the lemma is valid.

Consider the second case. Applying Lemma 3 with $\delta = 2^{-1}\delta_0$ and r to the program $\{y_i\}_{i=0}^N$ we can easily see that there exist integers $i(1)$, $i(2) \in [0, N]$, $q(1)$, and $q(2)$ such that

(2.10)
$$3m < i(1) \le 300m\|\theta\|C_2(2^{-1}\delta_0)^{-1}, \quad d(y_{i(1)-p}, x^*_{q(1)-p}) \le 2^{-1}\delta_0 \quad (p = 0, \dots, 3m),$$

$$\begin{aligned} &N - 3m > i(2) \ge N - 300m\|\theta\|C_2(2^{-1}\delta_0)^{-1}, \\ &d(y_{i(2)+p}, x^*_{q(2)+p}) \le 2^{-1}\delta_0 \quad (p = 0, \dots, 3m). \end{aligned} \tag{2.11}$$

Consider programs $\{\bar{y}_i\}_{i=0}^N$, $\{\bar{\bar{y}}_i\}_{i=0}^N$, where

$$\begin{aligned} &\bar{y}_i = y_i \quad (i = 0, \dots, i(2) + 3m), \quad \bar{y}_i = x^*_{i-i(2)+q(2)} \quad (i = i(2) + 3m + 1, \dots, N), \\ &\bar{\bar{y}}_i = y_i \quad (i = i(1) - 3m, \dots, N), \quad \bar{\bar{y}}_i = x^*_{i-i(1)+q(1)} \quad (i = 0, \dots, i(1) - 3m - 1). \end{aligned}$$

(2.9) implies

$$\begin{aligned} -r &\le \sum_{i=0}^{N-1} [(\phi + \theta)(\bar{y}_i, \bar{y}_{i+1}) - (\phi + \theta)(y_i, y_{i+1})] \\ &\le - \sum_{i=i(2)+3m}^{N-1} \theta(y_i, y_{i+1}) + 2r(N - i(2)) + C_1(\delta_0), \end{aligned}$$

$$\sum_{i=i(2)+3m}^{N-1} \theta(y_i, y_{i+1}) \le C_1(\delta_0) + 601m\|\theta\|C_2(2^{-1}\delta_0)^{-1}r,$$

$$\begin{aligned} -r &\le \sum_{i=0}^{N-1} [(\phi + \theta)(\bar{\bar{y}}_i, \bar{\bar{y}}_{i+1}) - (\phi + \theta)(y_i, y_{i+1})] \\ &\le - \sum_{i=0}^{i(1)-3m-1} \theta(y_i, y_{i+1}) + 2ri(1) + C_1(\delta_0), \end{aligned}$$

$$\sum_{i=0}^{i(1)-3m-1} \theta(y_i, y_{i+1}) \le C_1(\delta_0) + 601m\|\theta\|C_2(2^{-1}\delta_0)^{-1}r.$$

By the choice of δ_0, r

$$\begin{aligned} &\theta(y_i, y_{i+1}) \le 3 \cdot 8^{-1}C_2(\delta), \qquad (y_i, y_{i+1}) \in \bigcup_{j=0}^{m-1} \bar{B}((x^*_j, x^*_{j+1}), \delta) \\ &(i \in \{i(2) + 3m, \dots, N - 1\} \cup \{0, \dots, i(1) - 3m - 1\}). \end{aligned}$$

The last relation and (2.10) and (2.11) imply the validity of the lemma.

LEMMA 5. *Let $\delta \in (0, D_0)$. Then there exists a number $r > 0$ such that for every $\phi \in \mathbf{C}(K \times K)$ satisfying $\|\phi\| \le r$, every integer $N \ge 1$, and every integer k there are programs $\{z_i\}_{i=0}^N$, $\{t_i\}_{i=0}^N$ for which $z_0 = x^*_k$,*

$$d_1((z_i, z_{i+1}), (x^*_{k+i}, x^*_{k+i+1})) \le \delta \qquad (i = 0, \dots, \min\{N - 1, 3m - 1\}),$$

$$(z_{N-1}, z_N) \in \bigcup_{j=0}^{m-1} \bar{B}((x_j^*, x_{j+1}^*), \delta),$$

$$t_N = x_k^*, d_1((t_{N-i-1}, t_{N-i}), \qquad (x_{k-i-1}^*, x_{k-i}^*)) \le \delta$$

$$(i = 0, \ldots, \min\{3m-1, N-1\}), \qquad (t_0, t_1) \in \bigcup_{j=0}^{m-1} \bar{B}((x_j^*, x_{j+1}^*), \delta),$$

$$\text{(2.12)} \quad \sup\left\{\sum_{i=0}^{N-1} (\phi+\theta)(z_i, z_{i+1}), \sum_{i=0}^{N-1} (\phi+\theta)(t_i, t_{i+1})\right\} \le \ell(\phi, N) + 2C_1(\delta).$$

Proof. We choose a number $r \in (0, (16m)^{-1}C_1(\delta))$ such that Lemma 4 is valid for δ, r. Let $\phi \in \mathbf{C}(K \times K)$, $\|\phi\| \le r$, N be a natural number, and k be an integer. There are two cases: 1. $N \le 4m$; 2. $N > 4m$.

Consider the first case. Let q be an integer. We define $y_i^q = x_{q+i}^*$ $(i = 0, \ldots, N)$. It is easy to see that

$$\sum_{i=0}^{N-1} (\phi+\theta)(y_i^q, y_{i+1}^q) \le \ell(\phi, N) + 2rN \le \ell(\phi, N) + C_1(\delta).$$

Set $z_i = y_i^k$, $t_i = y_i^{k-N} (i = 0, \ldots, N)$. For the first case the lemma is proved. Consider the second case. There is a program $\{y_0, \ldots, y_N\}$ such that $\sum_{i=0}^{N-1} (\phi+\theta)(y_i, y_{i+1}) = \ell(\phi, N)$. By Lemma 4 there are integers $k(1)$ and $k(2)$ such that relations (2.7) and (2.8) hold.

Define $y_i = x_{k(2)+i}^*$ $(i = -1, -2, \ldots)$, $y_{i+N} = x_{k(1)+i}^*$ $(i = 1, 2, \ldots)$. There exist $g, q \in \{1, \ldots, m\}$ such that $y_{-g} = x_k^* = y_{N+q}$. We set $z_i = y_{i-g} (i = 0, \ldots, N)$, $t_i = y_{i+q}$ $(i = 0, \ldots, N)$. To complete the proof we should only verify the validity of (2.12). By our choice of r we have

$$\sum_{i=0}^{N-1} [(\phi+\theta)(z_i, z_{i+1}) - (\phi+\theta)(y_i, y_{i+1})]$$
$$= \sum_{i=0}^{g-1} (\phi+\theta)(y_{-g+i}, y_{-g+i+1}) - \sum_{i=N-g}^{N-1} (\theta+\phi)(y_i, y_{i+1}) \le 2gr + C_1(\delta),$$

$$\sum_{i=0}^{N-1} (\theta+\phi)(z_i, z_{i+1}) \le \ell(\phi, N) + 2C_1(\delta).$$

Similarly

$$\sum_{i=0}^{N-1} [(\theta+\phi)(t_i, t_{i+1}) - (\theta+\phi)(y_i, y_{i+1})]$$
$$= \sum_{i=N}^{q+N-1} (\theta+\phi)(y_i, y_{i+1}) - \sum_{i=0}^{q-1} (\theta+\phi)(y_i, y_{i+1}) \le 2qr + C_1(\delta),$$

$$\sum_{i=0}^{N-1} (\theta+\phi)(t_i, t_{i+1}) \le \ell(\phi, N) + 2C_1(\delta).$$

The validity of relation (2.12) is proved. This completes the proof of the lemma.

LEMMA 6. *Let $\varepsilon \in (0, D_0)$. Then there exists $R > 0$ such that for every $\phi \in \mathbf{C}(K \times K)$ satisfying $\|\phi\| \le R$, every integer $N \ge 1$, and every program $\{y_i\}_{i=0}^{N}$ satisfying $\sum_{i=0}^{N-1}(\theta + \phi)(y_i, y_{i+1}) \le \ell(\phi, N) + R$ there is an integer k such that $d_1((y_i, y_{i+1}), (x^*_{k+i}, x^*_{k+i+1})) \le \varepsilon$ $(i = 0, \dots, N-1)$.*

Proof. Choose $r_0 > 0$ such that Lemma 4 holds with $\delta = \varepsilon$, $r = r_0$, and choose $\delta_0 \in (0, \varepsilon)$ such that $8C_1(\delta_0) < r_0$. We choose $R \in (0, r_0)$ such that $16mR \le C_1(\delta_0)$ and Lemma 5 holds with $\delta = \delta_0$, $r = R$. Let $\phi \in \mathbf{C}(K \times K)$, $\|\phi\| \le R$, N be a natural number, and $\{y_i\}_{i=0}^{N}$ be a program satisfying $\sum_{i=0}^{N-1}(\theta + \phi)(y_i, y_{i+1}) \le \ell(\phi, N) + R$.

We suppose that for ϕ, N, $\{y_i\}_{i=0}^{N}$ the lemma does not hold. Then there exists $q \in \{0, \dots, N-1\}$ such that $(y_q, y_{q+1}) \notin \cup_{j=0}^{m-1} \bar{B}((x^*_j, x^*_{j+1}), \varepsilon)$. Lemma 4 implies that $N - 3m > q > 3m$,

$$\sum_{i=0}^{q}(\theta + \phi)(y_i, y_{i+1}) > \ell(\phi, q+1) + r_0. \tag{2.13}$$

Lemma 5 implies the existence of programs $\{z_i\}_{i=0}^{q+1}$, $\{h_i\}_{i=0}^{N-q-1}$ such that

$$z_{q+1} = x^*_0 = h_0, \tag{2.14}$$

$$\sum_{i=0}^{q}(\theta + \phi)(z_i, z_{i+1}) \le \ell(\phi, q+1) + 2C_1(\delta_0), \tag{2.15}$$

$$\sum_{i=0}^{N-q-2}(\theta + \phi)(h_i, h_{i+1}) \le \ell(\phi, N-q-1) + 2C_1(\delta_0). \tag{2.16}$$

Consider a program $\{t_i\}_{i=0}^{N}$, where $t_i = z_i$ $(i = 0, \dots, q+1)$, $t_i = h_{i-q-1}$ $(i = q+2, \dots, N)$. Then using (2.13) and (2.15), (2.16) we have

$$\begin{aligned}
-R &\le \sum_{i=0}^{N-1}[(\theta + \phi)(t_i, t_{i+1}) - (\theta + \phi)(y_i, y_{i+1})] \\
&= \sum_{i=0}^{q}[(\theta + \phi)(z_i, z_{i+1}) - (\theta + \phi)(y_i, y_{i+1})] \\
&\quad + \sum_{i=0}^{N-q-2}(\theta + \phi)(h_i, h_{i+1}) - \sum_{i=q+1}^{N-1}(\phi + \theta)(y_i, y_{i+1}) < \ell(\phi, q+1) \\
&\quad + 2C_1(\delta_0) - (\ell(\phi, q+1) + r_0) + (\ell(\phi, N-q-1) + 2C_1(\delta_0)) \\
&\quad - \ell(\phi, N-q-1) \le 4C_1(\delta_0) - r_0, \qquad r_0 \le R + 4C_1(\delta_0) \le 5C_1(\delta_0) < r_0.
\end{aligned}$$

The obtained contradiction proves the lemma.

LEMMA 7. *Let $\varepsilon \in (0, D_0)$. Then there exists a number $r > 0$ such that for every integer $N \ge 1$, every integer k, and every $\phi \in \mathbf{C}(K \times K)$, satisfying $\|\phi\| \le r$ there is a program $\{z_i\}_{i=0}^{N}$ such that $z_0 = x^*_k$, $z_N = x^*_{k+N}$,*

$$d_1((z_i, z_{i+1}), (x^*_{k+i}, x^*_{k+i+1})) \le \varepsilon \qquad (i = 0, \dots, N-1), \tag{2.17}$$

$$\sum_{i=0}^{N-1}(\phi + \theta)(z_i, z_{i+1}) \le \ell(\phi, N) + 6C_1(\varepsilon). \tag{2.18}$$

Proof. We choose a number $r_0 > 0$ such that Lemma 6 holds for $R = r_0$ and ε. We choose numbers $\delta \in (0, \varepsilon)$, $r \in (0, r_0)$ such that $8C_1(\delta) \le r_0$, $16mr \le C_1(\delta)$, and Lemma 5 holds for r, δ. Let N be a natural number, k be an integer, $\phi \in \mathbf{C}(K \times K)$, and $\|\phi\| \le r$. The validity of Lemma 5 for r, δ implies the existence of a program $\{y_i\}_{i=0}^N$ such that $y_0 = x_k^*$,

$$d_1((y_i, y_{i+1}), (x_{k+i}^*, x_{k+i+1}^*)) \le \delta \qquad (i = 0, \dots, \min\{N-1, 3m-1\}),$$

$$(y_{N-1}, y_N) \in \bigcup_{j=0}^{m-1} \bar{B}((x_j^*, x_{j+1}^*), \delta),$$

$$\sum_{i=0}^{N-1} (\phi + \theta)(y_i, y_{i+1}) \le \ell(\phi, N) + 2C_1(\delta) \le \ell(\phi, N) + r_0. \tag{2.19}$$

The validity of Lemma 6 for $R = r_0$ and ε implies that

$$d_1((y_i, y_{i+1}), (x_{k+i}^*, x_{k+i+1}^*)) \le \varepsilon \quad \text{for}\, i = 0, \dots, N-1.$$

Now to complete the proof of the lemma we should only set $z_i = y_i$ $(i = 0, \dots, N-1)$, $z_N = x_{N+k}^*$ and note (see (2.19)) that

$$\sum_{i=0}^{N-1} (\phi + \theta)(z_i, z_{i+1}) \le \sum_{i=0}^{N-1} (\phi + \theta)(y_i, y_{i+1}) + 2r + C_1(\varepsilon)$$
$$\le \ell(\phi, N) + 2C_1(\delta) + 2r + C_1(\varepsilon) \le \ell(\phi, N) + 6C_1(\varepsilon).$$

LEMMA 8. *Let $\varepsilon_0 \in (0, D_0)$. Then there exists a number $r > 0$ such that for every $u \in \mathbf{C}(K \times K)$ satisfying $\|u - v\| \le r$ and every (u)-good program $\{z_i\}_{i=0}^\infty$ there are an integer $N > 1$ and an integer k such that*

$$d_1((z_{N+i}, z_{N+i+1}), \quad (x_{k+i}^*, x_{k+i+1}^*)) \le \varepsilon_0 \quad (i = 0, 1, \dots).$$

Proof. Choose a number $r_0 > 0$ such that Lemma 6 holds with $\varepsilon = \varepsilon_0$ and $R = r_0$, and choose $\delta \in (0, \varepsilon_0)$ and $r \in (0, r_0)$ such that $8C_1(\delta) \le r_0$, $16mr \le C_1(\delta)$, and Lemma 7 holds with $\varepsilon = \delta$ and r. Let $u \in \mathbf{C}(K \times K)$, $\|u - v\| \le r$, and $\{z_i\}_{i=0}^\infty$ be a (u)-good program. We define $\phi = u - v$. Suppose that the lemma doesn't hold for u, $\{z_i\}_{i=0}^\infty$. Then there exists a sequence of integers $\{i_k\}_{k=1}^\infty$ such that $i_1 \ge 8m$, $i_{k+1} - i_k \ge 8m$:

$$(z_{i_k}, z_{i_k+1}) \notin \bigcup_{j=0}^{m-1} \bar{B}((x_j^*, x_{j+1}^*), \varepsilon_0) \qquad (k = 1, 2, \dots). \tag{2.20}$$

Set $i_0 = -1$. Let $k \in \{0, 1, \dots\}$. Consider a program $\{z_i : i = i_k + 1, \dots, i_{k+1} + 1\}$. By (2.20) and Lemma 6, which holds for $\varepsilon = \varepsilon_0$ and $R = r_0$, we have

$$\sum_{i=i_k+1}^{i_{k+1}} (\phi + \theta)(z_i, z_{i+1}) > \ell(\phi, i_{k+1} - i_k) + r_0. \tag{2.21}$$

Using Lemma 7 with r and $\varepsilon = \delta$ we obtain a program $\{y_i\}_{i=0}^\infty$ such that $y_{i_k+1} \in \{x_j^* : j = 0, \dots, m-1\}$ $(k = 0, 1, \dots)$,

$$\sum_{i=i_k+1}^{i_{k+1}} (\theta + \phi)(y_i, y_{i+1}) \le \ell(\phi, i_{k+1} - i_k) + 6C_1(\delta) \qquad (k = 0, 1, \dots). \tag{2.22}$$

Relations (2.21) and (2.22) and Theorem 1 imply

$$\begin{aligned}
&\sum_{i=0}^{i_N} [u(z_i, z_{i+1}) - u(y_i, y_{i+1})] \\
&\quad \geq -4\|\pi\| + \sum_{i=0}^{i_N} [(\phi + \theta)(z_i, z_{i+1}) - (\theta + \phi)(y_i, y_{i+1})] \\
&\quad \geq -4\|\pi\| + \sum_{k=0}^{N-1} \sum_{i=i_k+1}^{i_{k+1}} [(\phi + \theta)(z_i, z_{i+1}) - (\theta + \phi)(y_i, y_{i+1})] \\
&\quad \geq -4\|\pi\| + N(r_0 - 6C_1(\delta)) \to \infty, \qquad N \to \infty, \\
&\sum_{i=0}^{i_N} [u(z_i, z_{i+1}) - \mu(u)] \\
&\quad = \sum_{i=0}^{i_N} [u(z_i, z_{i+1}) - u(y_i, y_{i+1})] + \sum_{i=0}^{i_N} [u(y_i, y_{i+1}) - \mu(u)] \\
&\quad \geq -4\|\pi\| + N(r_0 - 6C_1(\delta)) + b(u) - a(u) \to \infty, \qquad N \to \infty
\end{aligned}$$

But $\{z_i\}_{i=0}^{\infty}$ is a (u)-good program. The obtained contradiction proves the lemma.

3. Proof of Theorems 3 and 4. We consider the set E of all functions $v \in \mathbf{C}(K \times K)$ for which there exist an integer $m(v) \geq 1$; a sequence $\{x_i^*(v)\}_{i=0}^{m(v)} \subset K$; continuous functions $\pi_v : K \to \mathbf{R}^1$, $\theta_v : K \times K \to \mathbf{R}^1$; and a number μ_v such that the following conditions hold.

1. for i, $j \in \{0, \ldots, m(v)\}$ satisfying $i < j$ the equality $x_i^*(v) = x_j^*(v)$ holds if and only if $i = 0$, $j = m(v)$;

2. $v(x, y) = \mu_v + \pi_v(x) - \pi_v(y) + \theta_v(x, y)(x, y \in K)$;

3. $\|\theta_v\| > 0$, the function θ_v is nonnegative, and $\theta_v(x, y) = 0$ if and only if

$$(x, y) \in \{(x_i^*(v), x_{i+1}^*(v)) : i = 0, \ldots, m(v) - 1\}.$$

At the beginning of §2 we already noted that E is everywhere dense in $\mathbf{C}(K \times K)$ by Lemma 2. It is easy to see that $\mu(v) = \mu_v(v \in E)$, and for every $v \in E$ we can apply Lemmas 3–8. For every $v \in E$ and every $i \in \{0, \pm 1, \ldots\} \backslash \{0, \ldots, m(v)\}$ define $x_i^*(v) \in K$ such that $x_{i+m(v)}^*(v) = x_i^*(v)$ $(i = 0, \pm 1, \ldots)$. Let $v \in E$. We set

$$D(v) = 8^{-1} \inf \{d(x_i^*(v), x_j^*(v)) : i, j \in \{0, \ldots, m(v) - 1\},\ i \neq j\}. \tag{3.1}$$

If $m(v) = 1$, then $D(v) = +\infty$.

Let $p \in \{1, 2, \ldots\}$. We define

$$\delta(v, p) = \inf \{2^{-1} D(v), p^{-1}\}. \tag{3.2}$$

It is easy to see that there exist numbers $\Gamma(v, p) \in (0, \delta(v, p))$, $d(v, p) \in (0, \Gamma(v, p))$ such that Lemma 8 holds for

$$\varepsilon_0 = \Gamma(v, p), \quad r = d(v, p), \quad x_i^* = x_i^*(v) \quad (i = 0, \pm 1, \ldots).$$

Now we define $F = \bigcap_{p=1}^{\infty} \bigcup_{v \in E} B(v, d(v, p))$. For this set F we will prove the theorems.

Proof of Theorem 3. First we will prove that assertion 1 of Theorem 3 holds. Let $u \in F$, $\mathbf{x}, \mathbf{y}$ be (u)-good programs, $p \in \{1, 2, \ldots\}$. There is $v \in E$ such that $u \in B(v, d(v, p))$. It follows from the definition of $d(v, p)$ and $\Gamma(v, p)$ and Lemma 8 that

$$\begin{aligned} \operatorname{dist}(\Omega(\mathbf{x}), \{(x_i^*(v), x_{i+1}^*(v)) : i = 0, \ldots, m(v) - 1\}) &\le \Gamma(v, p), \\ \operatorname{dist}(\Omega(\mathbf{y}), \{(x_i^*(v), x_{i+1}^*(v)) : i = 0, \ldots, m(v) - 1\}) &\le \Gamma(v, p), \\ \operatorname{dist}(\omega(\mathbf{x}), \{x_i^*(v) : i = 0, \ldots, m(v) - 1\}) &\le \Gamma(v, p), \\ \operatorname{dist}(\omega(\mathbf{y}), \{x_i^*(v) : i = 0, \ldots, m(v) - 1\}) &\le \Gamma(v, p), \\ \operatorname{dist}(\Omega(\mathbf{x}), \Omega(\mathbf{y})) \le 2p^{-1}, \ \operatorname{dist}(\omega(\mathbf{x}), \omega(\mathbf{y})) &\le 2p^{-1}. \end{aligned}$$

This completes the proof of assertion 1.

Let us prove assertion 2. Let $u \in F$, $\varepsilon > 0$, and $\{x_i\}_{i=0}^\infty$ be a (u)-good program. Choose an integer $p \ge 1$, $v \in E$ such that $4p^{-1} < \varepsilon$, $u \in B(v, d(v, p))$. Now the validity of assertion 2 follows from the definition of $d(v, p)$, $\Gamma(v, p)$ and Lemma 8.

Now we will prove assertion 3. Let $u \in F$, $\delta > 0$. There is an integer $p \ge 1$, $v \in E$ such that $4p^{-1} < \delta$, $u \in B(v, d(v, p))$. Set $W(u) = B(v, d(v, p))$. The validity of assertion 3 follows from the definition of $d(v, p)$, $\Gamma(v, p)$ and Lemma 8.

We have the following result.

PROPOSITION 1. 1. $\sup\{\pi^u(x) : x \in H_0(u)\} = 0 \ (u \in F)$.

2. *Let $u \in F$, δ be a positive number. Then there are an integer $m \ge 1$ and a sequence $\{x_i^*\}_{i=0}^m$ such that*

$$x_0^* = x_m^*;$$

$$\textit{if } 0 \le i < j \le m, \ x_i^* = x_j^*, \ \textit{then } i = 0, \ j = m;$$

$$\operatorname{dist}(H(u), \{(x_i^*, x_{i+1}^*) : i = 0, \ldots, m - 1\}) \le \min\{\delta, 8^{-1} d(x_i, x_j) \ (0 \le i < j \le m - 1)\}.$$

Proof. It is easy to see that assertion 1 of Proposition 1 follows from Assertion 1 of Theorem 3. Let us prove Assertion 2. Let $\delta > 0$, $u \in F$. Choose an integer $p \ge 1$ satisfying $8p^{-1} < \delta$. There is $v \in E$ such that $u \in B(v, d(v, p))$. By Lemma 8, assertion 1 of Theorem 3, and our choice of $\delta(v, p)$, $\Gamma(v, p)$ we have

$$\Gamma(v, p) < \delta(v, p) \le 16^{-1} \inf\{d(x_i^*(v), x_j^*(v)) : i, j \in \{0, \ldots, m(v) - 1\}, \ i \ne j\},$$

$$\operatorname{dist}(H(u), \{(x_i^*(v), x_{i+1}^*(v)) : i = 0, \ldots, m(v) - 1\}) \le \Gamma(v, p).$$

The validity of assertion 2 follows easily from these relations.

Assertion 2 of Proposition 1 establishes that for $u \in F$ the set $H(u)$ is approximated by finite periodic programs.

Proof of Theorem 4. We will prove the following lemma first.

LEMMA 9. *Assume that $u \in \mathbf{C}(K \times K)$ and there is a closed set $H_0(u) \subset K$ such that for every (u)-good program $\{x_i\}_{i=0}^\infty$ we have $\omega(\{x_i\}_{i=0}^\infty) = H_0(u)$. Then the following assertions hold*:

1. *Let $\{x_i\}_{i=0}^\infty$ be a program satisfying $\theta^u(x_i, x_{i+1}) = 0 \ (i = 0, 1, \ldots)$. Then $\{x_i\}_{i=0}^\infty$ is a (u)-weakly optimal program and there is a subsequence $\{x_{i_k}\}_{k=1}^\infty$ such that for every program $\{y_i\}_{i=0}^\infty$ satisfying $y_0 = x_0$, the relation*

$$\liminf_{k \to \infty} \sum_{j=0}^{i_k - 1} [u(y_j, y_{j+1}) - u(x_j, x_{j+1})] \ge 0$$

holds, and moreover, if we have an equality

$$\liminf_{k\to\infty} \sum_{j=0}^{i_k-1} [u(y_j, y_{j+1}) - u(x_j, x_{j+1})] = 0,$$

then $\theta^u(y_i, y_{i+1}) = 0$ $(i = 0, 1, \ldots)$.

2. *Let* $\{x_i\}_{i=0}^{\infty}$ *be a program. Then the relation*

$$\pi^u(x_0) = \liminf \sum_{i=0}^{N-1} [u(x_i, x_{i+1}) - \mu(u)]$$

holds if and only if $\theta^u(x_i, x_{i+1}) = 0$ $(i = 0, 1, \ldots)$.

Proof. First let us prove assertion 1. There is $x^* \in H_0(u)$ satisfying $\pi^u(x^*) = \sup\{\pi^u(y)\colon y \in H_0(u)\}$. To prove the assertion we should only note that there exists a subsequence $\{x_{i_k}\}_{k=1}^{\infty}$ satisfying $x_{i_k} \xrightarrow[k\to\infty]{} x^*$.

Next we will prove assertion 2. There is $x^* \in H_0(u)$ satisfying $\pi^u(x^*) = \sup\{\pi^u(y)\colon y \in H_0(u)\}$. To prove assertion 2 we should only note that for every (u)-good program $\{x_i\}_{i=0}^{\infty}$ the following relation holds:

$$\liminf \sum_{i=0}^{N-1} [u(x_i, x_{i+1}) - \mu(u)] = \sum_{i=0}^{\infty} \theta^u(x_i, x_{i+1}) + \pi^u(x_0) - \pi^u(x^*). \tag{3.3}$$

Theorem 3 and Lemma 9 imply the validity of Theorem 4.

4. Examples. We will give an example of an open set $D \subset C([0,1] \times [0,1])$ such that for every $v \in D$ there is not any (v)-overtaking optimal program.

Example 1. Let $K = [0,1]$, $0 < x_0^* < x_1^* < 1$, $\pi\colon [0,1] \to \mathbf{R}^1$, and $\theta\colon [0,1] \times [0,1] \to \mathbf{R}^1$ be continuous functions such that θ is nonnegative, $\theta(x,y) = 0$ if and only if $(x,y) \in \{(x_0^*, x_1^*), (x_1^*, x_0^*)\}$, $\pi(x_0^*) - \pi(x_1^*) - \theta(x_0^*, x_0^*) > 0$. Set $v(x,y) = \pi(x) - \pi(y) + \theta(x,y) (x, y \in [0,1])$.

Choose numbers

$$\varepsilon \in (0, 14^{-1}[\pi(x_0^*) - \pi(x_1^*) - \theta(x_0^*, x_0^*)]), \tag{4.1}$$

$\delta \in (0, 8^{-1}|x_0^* - x_1^*|)$ such that

$$\begin{aligned} \sup\{\, &|\pi(z_1) - \pi(z_2)|, |\theta(y_1, y_2) - \theta(y_3, y_4)|\colon z_1, z_2, y_1, y_2, y_3, y_4 \\ &\in [0,1], |z_1 - z_2| \le \delta, |y_1 - y_3| \le \delta, |y_2 - y_4| \le \delta\} \le \varepsilon, \end{aligned} \tag{4.2}$$

and choose $r \in (0, \varepsilon)$ such that Lemma 8 holds for $\varepsilon_0 = \delta$, r, and v.

Let $u \in C([0,1] \times [0,1])$, $\|u - v\| \le r$, and $\phi = u - v$. We prove that there is no (u)-overtaking optimal program.

Suppose that $\{x_i\}_{i=0}^{\infty}$ is a (u)-overtaking optimal program. By Lemma 8 there exists an integer $N \ge 1$ such that

$$|x_{N+2i} - x_0^*| \le \delta, \quad |x_{N+2i+1} - x_1^*| \le \delta \quad (i = 0, 1, \ldots). \tag{4.3}$$

Consider a program $\{y_i\}_{i=0}^{\infty}$, where $y_i = x_i$ $(i = 0, \ldots, N)$, $y_{i+1} = x_i$ $(i = N, N+1, \ldots)$. For every integer $k \ge 1$ by (4.1)–(4.3) we have

$$\begin{aligned}
&\sum_{i=0}^{N+2k} [u(x_i, x_{i+1}) - u(y_i, y_{i+1})] \\
&\quad = \sum_{i=0}^{N+2k} u(x_i, x_{i+1}) \\
&\qquad - \left[\sum_{i=0}^{N-1} u(x_i, x_{i+1}) + u(x_N, x_N) + \sum_{i=N}^{N+2k-1} u(x_i, x_{i+1})\right] \\
&\quad = u(x_{N+2k}, x_{N+2k+1}) - u(x_N, x_N) \\
&\quad \geq \pi(x_{N+2k}) - \pi(x_{N+2k+1}) + \theta(x_{N+2k}, x_{N+2k+1}) - \theta(x_N, x_N) - 2r \\
&\quad \geq -2r + (\pi(x_0^*) - \varepsilon) - (\pi(x_1^*) + \varepsilon) - \theta(x_0^*, x_0^*) - \varepsilon \\
&\quad \geq \pi(x_0^*) - \pi(x_1^*) - \theta(x_0^*, x_0^*) - 3\varepsilon - 2r \\
&\quad \geq \pi(x_0^*) - \pi(x_1^*) - \theta(x_0^*, x_0^*) - 5\varepsilon \geq 6\varepsilon, \\
&\sum_{i=0}^{N+2k} [u(x_i, x_{i+1}) - u(y_i, y_{i+1})] \geq 6\varepsilon \qquad (k = 1, 2 \ldots).
\end{aligned}$$

We thus arrive at a contradiction. Hence, there is no (u)-overtaking optimal program for $u \in \bar{B}(v, r)$.

Next we will give an example of $v \in C([0,1] \times [0,1])$ for which there exist a (v)-weakly optimal program $\{x_i\}_{i=0}^{\infty}$ and a (v)-minimal energy configuration $\{y_i\}_{i=-\infty}^{\infty}$ such that $\sup\{\theta^v(x_i, x_{i+1}) : i = 0, 1, \ldots\} > 0$, $\sup\{\theta^v(y_i, y_{i+1}) : i = 0, \pm 1, \ldots\} > 0$.

Example 2. Let $K = [0,1]$, $0 < x_0^* < x_1^* < 1$, $0 < \varepsilon < 8^{-1}|x_0^* - x_1^*|$, and

$$\theta(x, y) = \min\{1, 64\varepsilon^{-4}[(x - x_0^*)^2 + (y - x_1^*)^2][(x - x_1^*)^2 + (y - x_0^*)^2]\} \\ (x, y \in [0,1]).$$

Let π: $[0,1] \to \mathbf{R}^1$ be a continuous function such that

$$\begin{aligned} &\pi(x_0^*) - \pi(x_1^*) > 1, \\ &v(x, y) = \pi(x) - \pi(y) + \theta(x, y) \qquad (x, y \in [0,1]). \end{aligned} \tag{4.4}$$

We define x_i^* for all integers $i \notin \{0, 1\}$ so that $x_{i+2}^* = x_i^*$ $(i = 0, \pm 1, \ldots)$. It is easy to see that $\mu(v) = 0$. By Lemma 8 $\omega(\mathbf{x}) = \{x_0^*, x_1^*\}$ for every (v)-good configuration $\mathbf{x}$. It is easy to verify that

$$\pi^v(x_0^*) = 0, \quad \pi^v(x_1^*) = \pi(x_1^*) - \pi(x_0^*), \quad \theta^v(x, y) = \theta(x, y)(x, y \in \{x_0^*, x_1^*\}).$$

Fix an integer $k \geq 0$ and consider a program $\mathbf{Y}^k = \{y_i^k\}_{i=0}^{\infty}$, where

$$y_i^k = x_0^* \; (i = 0, \ldots, k), \quad y_i^k = x_{i-1}^* \quad (i \text{ is an integer}, \; i \geq k + 1). \tag{4.5}$$

We shall show that $\{y_i^k\}_{i=0}^{\infty}$ is a (v)-weakly overtaking program.

Let $\mathbf{z}$ be a (v)-good program, $z_0 = x_0^*$. It is easy to see that $\Omega(\mathbf{z}) = \{(x_0^*, x_1^*), (x_1^*, x_0^*)\}$. Hence, there exists an integer $Q \geq 1$ such that

$$|z_{Q+i} - x_i^*| \leq 4^{-1}\varepsilon \qquad (i = 0, 1, \ldots). \tag{4.6}$$

There are two cases: 1. $Q = 2i + 1$ for some integer i; 2. $Q = 2i$ for some integer i. Clearly,

$$\sum_{i=0}^{2N} v(y_i^k, y_{i+1}^k) = \theta(x_k^*, x_k^*) = 1 \qquad (N > k + 1). \tag{4.7}$$

Consider case 1. There exists an integer $q \geq 0$ such that $Q = 2q + 1$. By (4.6)

$$|z_{2q+1} - x_0^*| \leq 4^{-1}\varepsilon. \tag{4.8}$$

Suppose that

$$(z_i, z_{i+1}) \in \bigcup_{j=0}^{1} \bar{B}((x_j^*, x_{j+1}^*), \varepsilon) \qquad (i = 0, \dots, 2q).$$

This relation implies that $|z_i - x_i^*| \leq \varepsilon \, (i = 0, \dots, 2q + 1)$ and we obtained a contradiction (see (4.8)). Hence, there exists $p \in \{0, \dots, 2q\}$ such that

$$(z_p, z_{p+1}) \notin \bigcup_{j=0}^{1} \bar{B}((x_j^*, x_{j+1}^*), \varepsilon).$$

It follows from the definition of θ that $\theta(z_p, z_{p+1}) = 1$. Together with (4.7) the relation $\Omega(\mathbf{z}) = \{(x_0^*, x_1^*), (x_1^*, x_0^*)\}$ implies that

$$\begin{aligned}\liminf_{N\to\infty} \sum_{i=0}^{2N} [v(z_i, z_{i+1}) - v(y_i^k, y_{i+1}^k)] &\geq \sum_{i=0}^{\infty} \theta(z_i, z_{i+1}) - 1\\ &\geq \theta(z_p, z_{p+1}) - 1 = 0,\\ \liminf_{N\to\infty} \sum_{i=0}^{2N} [v(z_i, z_{i+1}) - v(y_i^k, y_{i+1}^k)] &\geq 0.\end{aligned} \tag{4.9}$$

Consider case 2. Relation (4.6) implies that $z_{2i} \to x_0^*$, $z_{2i+1} \to x_1^*$ as $i \to \infty$,

$$\begin{aligned}&\liminf_{N\to\infty} \sum_{i=0}^{2N} [v(z_i, z_{i+1}) - v(y_i^k, y_{i+1}^k)]\\ &\geq \liminf_{N\to\infty} [\pi(x_0^*) - \pi(z_{2N+1}) - 1] = \pi(x_0^*) - \pi(x_1^*) - 1,\end{aligned}$$

and by (4.4) relation (4.9) holds also for case 2. We have proved that $\{y_i^k\}_{i=0}^{\infty}$ is a (v)-weakly overtaking program. For an integer $i < 0$ we set $y_i^k = x_i$. It is easy to see that $\{y_i^k\}_{i=-\infty}^{\infty}$ is a (v)-minimal energy configuration.

Acknowledgments. The author thanks Dr. Arie Leizarowitz for fruitful discussions.

REFERENCES

[1] A. Leizarowitz, *Infinite horizon autonomous systems with unbounded cost*, Appl. Math. Optim., 13 (1985), pp. 19–43.

[2] ———, *Optimal trajectories of infinite-horizon deterministic control systems*, Appl. Math. Optim., 19 (1989), pp. 11–32.

[3] A. Leizarowitz and V. J. Mizel, *One-dimensional infinite-horizon variational problems arising in continuum mechanics*, Arch. Rational Mech. Anal., 106 (1989), pp. 161–194.

[4] S. Aubry and P. Y. Le Daeron, *The discrete Frenkel–Kontorova model and its extensions* I, Phys. D, 8 (1983), pp. 381–442.

[5] A. J. Zaslavski, *Ground states in Frenkel–Kontorova models*, Izv. Akad. Nauk SSSR Ser. Mat., 50 (1986), pp. 969–999. (In English.)

[6] D. Gale, *On optimal development in multisector economy*, Rev. Econom. Stud., 34 (1967), pp. 1–19.

[7] C. C. von Weizsäcker, *Existence of optimal programs of accumulation for an infinite horizon*, Rev. Econom. Stud., 32 (1965), pp. 85–104.

[8] W. A. Brock and A. Haurie, *On existence of overtaking optimal trajectories over an infinite time horizon*, Math. Oper. Res., 1 (1976), pp 337–346.

[9] D. Carlson, A. Haurie, and A. Leizarowitz, *Infinite Horizon Optimal Control*, Springer-Verlag, Berlin, 1991.

[10] Z. Artstein and A. Leizarowitz, *Tracking periodic signals with the overtaking criterion*, IEEE Trans. Automat. Control, AC-30 (1985), pp. 1123–1126.

[11] V. L. Makarov and A. M. Rubinov, *Mathematical Theory of Economic Dynamics and Equilibria*, Nauka, Moscow, 1973. (English translation: Springer-Verlag, New York, 1977.)

[12] A. M. Rubinov, *Superlinear Multivalued Mappings and their Applications to Economic Mathematical Problems*, Nauka, Leningrad, 1980.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 33, No. 6, pp. 1661–1686, November 1995

© 1995 Society for Industrial and Applied Mathematics
003

OPTIMAL PROGRAMS ON INFINITE HORIZON 2*

A. J. ZASLAVSKI†

Abstract. We consider the limit behavior, as $N \to \infty$, of the expression $\sum_{i=0}^{N-1} v(x_i, x_{i+1})$ for programs $\{x_i\}_{i=0}^{\infty}$ in a compact metric space K, where v is a real-valued function defined on $K \times K$. We study the structure of (v)-weakly optimal programs and establish a turnpike theorem for a generic continuous function v. We also prove the existence of an almost periodic (v)-optimal program for a generic continuous function v.

Key words. good program, minimal-energy configuration, overtaking optimality, (v)-weakly optimal program

AMS subject classification. 49J99

Introduction. The study of optimization problems defined on infinite intervals has recently been a rapidly growing area of research. In this paper we are concerned with the infinite-horizon problem of minimizing the expression $\sum_{i=0}^{N-1} v(x_i, x_{i+1})$ as N grows to infinity, where $\{x_i\}_{i=0}^{\infty}$ is a sequence in a compact metric space K and v is a continuous function defined on $K \times K$. The interest in this discrete-time infinite-horizon problem stems from the recent study of various optimization problems which can be reduced to this framework, e.g., continuous-time control systems which are represented by ordinary differential equations whose cost integrand contains a discounting factor [1], the infinite-horizon deterministic control problem of minimizing $\int_0^T L(z, \dot{z})\, dt$ as $T \to \infty$ [2], and the analysis of a long slender bar of a polymeric material under tension [3].

Let K be a compact metric space, $\mathbf{R}^n$ be the Euclidean n-dimensional space, $\mathbf{C}(K \times K)$ be the space of all continuous functions $v : K \times K \to \mathbf{R}^1$ with the topology of the uniform convergence ($\|v\| = \sup\{|v(x, y)| : x, y \in K\}$). Let $\mathbf{C}(K)$ be the space of all continuous functions $v : K \to \mathbf{R}^1$ with the topology of the uniform convergence ($\|v\| = \sup\{|v(x)| : x \in K\}$) and $B(K \times K)$ be the set of all bounded and lower semicontinuous functions $v : K \times K \to \mathbf{R}^1$ (i.e., $v(\lim(x_k, y_k)) \leq \liminf v(x_k, y_k)$).

Consider any $v \in B(K \times K)$. We are interested in the limit behavior as $N \to \infty$ of the expression $\sum_{i=0}^{N-1} v(x_i, x_{i+1})$, where $\{x_i\}_{i=0}^{\infty}$ is an infinite sequence in K which we call a *program* (or a *configuration*) (see [1], [4], [5]) and which occasionally will be denoted by a boldface $\mathbf{x}$. (Similarly $\{y_i\}_{i=0}^{\infty}$ will be denoted by $\mathbf{y}$, etc.) A finite sequence $\{x_i\}_{i=0}^{N} \subset K$ $(N = 0, 1, \ldots)$ will be also called a program. We shall define three concepts of optimality.

A program $\{x_i\}_{i=0}^{\infty}$ is a (v)-*overtaking optimal program* if for every program $\{z_i\}_{i=0}^{\infty}$ satisfying $z_0 = x_0$ the following inequality holds:

$$\limsup_{N \to \infty} \sum_{i=0}^{N-1} [v(x_i, x_{i+1}) - v(z_i, z_{i+1})] \leq 0.$$

This notion known as the *overtaking optimality criterion* was introduced in the economic literature by Gale [6] and von Weizsäcker [7] and was employed to study infinite-horizon control problems [1], [8]–[10].

A program $\{x_i\}_{i=0}^{\infty}$ is (v)-weakly optimal [1], [6], [7] if for every program $\{z_i\}_{i=0}^{\infty}$

* Received by the editors October 7, 1993; accepted for publication July 8, 1994. This research was supported in part by a grant from the Israeli Ministry of Science and the MA-AGARA (a special project for the absorption of new immigrants) in the Department of Mathematics, Technion.

† Department of Mathematics, Technion–Israel Institute of Technology, 32000 Haifa, Israel.

satisfying $z_0 = x_0$ the following inequality holds:

$$\liminf_{N\to\infty} \sum_{i=0}^{N-1} [v(x_i, x_{i+1}) - v(z_i, z_{i+1})] \le 0.$$

A sequence $\{x_i\}_{i=0}^{\infty} \subset K$ is called a (v)-*minimal energy configuration* (*program*) if for each N, $M > 0$ the inequality

$$\sum_{i=-N}^{M-1} v(x_i, x_{i+1}) \le \sum_{i=-N}^{M-1} v(z_i, z_{i+1})$$

holds for every sequence $\{z_i\}_{i=-N}^{M} \subset K$ satisfying $x_{-N} = z_{-N}$, $x_M = z_M$ [3]–[5].

Of special interest is the *minimal long-run average cost growth rate*,

$$\mu(v) = \inf\left\{\liminf_{N\to\infty} N^{-1} \sum_{i=0}^{N-1} v(z_i, z_{i+1}) : \{z_i\}_{i=0}^{\infty} \text{is a program}\right\}.$$

A program $\{z_i\}_{i=0}^{\infty}$ is called a (v)-*good program* [1] if the sequence $\{\sum_{i=0}^{N-1} [v(z_i, z_{i+1}) - \mu(v)]\}_{N=1}^{\infty}$ is bounded. It was proved in [1] that for every program $\{z_i\}_{i=0}^{\infty}$ the sequence $\{\sum_{i=0}^{N-1} [v(z_i, z_{i+1}) - \mu(v)]\}_{N=1}^{\infty}$ either is bounded or diverges to infinity and that for every initial value z there is a (v)-good program $\{z_i\}_{i=0}^{\infty}$ satisfying $z_0 = z$. In [1] the following representation formula valid for every $v \in \mathbf{C}(K \times K)$ was also established:

$$v(x, y) = \theta^v(x, y) + \mu(v) - \pi^v(y) + \pi^v(x) \qquad (x, y \in K),$$

where π^v and θ^v are continuous functions,

$$\pi^v(x) = \inf\left\{\liminf \sum_{i=0}^{N-1} [v(z_i, z_{i+1}) - \mu(v)] : \mathbf{z} \subset K,\ z_0 = x\right\},$$

θ^v is nonnegative, and $E(x) = \{y \in K : \theta^v(x, y) = 0\}$ is nonempty for every $x \in K$.

In [13] we studied the structure of (v)-good programs and proved for a generic $v \in \mathbf{C}(K \times K)$, for every given $x \in K$, the existence of a (v)-weakly optimal program $\{x_i\}_{i=0}^{\infty}$ satisfying $x_0 = x$. We established the existence of a set $F_0 \subset \mathbf{C}(K \times K)$ which is a countable intersection of open and everywhere dense sets in $\mathbf{C}(K \times K)$ such that for every $u \in F_0$ the following propositions hold:

a) there are closed sets $H(u) \subset K \times K$, $H_0(u) \subset K$ such that for every (u)-good program $\{x_i\}_{i=0}^{\infty}$ the limit points set of $\{x_i\}_{i=0}^{\infty}$ is $H_0(u)$ and the limit points set of $\{(x_i, x_{i+1})\}_{i=0}^{\infty}$ is $H(u)$;

b) the set $H(u)$ is approximated by finite periodic programs;

c) for every initial point $x \in K$ there is a (u)-weakly optimal program $\{x_i\}_{i=0}^{\infty}$ satisfying $x_0 = x$, $\theta^u(x_i, x_{i+1}) = 0 (i = 0, 1, \ldots)$.

By proposition c), programs $\{x_i\}_{i=0}^{\infty}$ satisfying $\theta^u(x_i, x_{i+1}) = 0$ $(i = 0, 1, \ldots)$ are of great interest. The existence of such a program for any initial value x_0 follows from the properties of θ^u. In this paper we will establish the existence of a set $F \subset F_0$ which is a countable intersection of sets dense open and everywhere in $\mathbf{C}(K \times K)$ such that for every $u \in F$ the following propositions hold:

d) u is a continuity point of the mapping $v \to (\mu(v), \pi^v) \in \mathbf{R}^1 \times \mathbf{C}(K)$;

e) given any $\varepsilon > 0$, there exists $\delta > 0$ such that for every program $\{x_i\}_{i=0}^{\infty}$ satisfying $\theta^u(x_i, x_{i+1}) = 0 (i = 0, 1 \ldots)$ the sequence $\{(x_i, x_{i+1})\}_{i=0}^{\infty}$ is contained in an ε-neighborhood of $H(u)$ if x_0 belongs to a δ-neighborhood of $H_0(u)$ (stability condition);

f) all the sequences $\{(x_i, x_{i+1})\}_{i=0}^{\infty} \subset K \times K$ satisfying $\theta^u(x_i, x_{i+1}) = 0\,(i = 0, 1, \ldots)$ converge uniformity to $H(u)$;

g) every sequence $\{x_i\}_{i=-\infty}^{\infty} \subset K$ satisfying $\theta^u(x_i, x_{i+1}) = 0\,(i = 0, \pm 1, \ldots)$ is almost periodic.

One can see that the set $H(u)$ $(u \in F)$ is an analogue of a turnpike set [11], [12], and we prove for it a weak turnpike theorem.

The paper is organized as follows. In §1 we give the necessary definitions and state precisely our main results. Necessary auxiliary results obtained in [13] are stated in §2. In §§3 and 4 we prove the preliminary lemmas and develop the suitable technique which is used in §5 to prove the theorems.

1. Definitions and theorems. Let K be a compact metric space, $v \in \mathbf{C}(K \times K)$. We define

$$\mu(v) = \inf\left\{\liminf N^{-1} \sum_{i=0}^{N-1} v(z_i, z_{i+1}) : \{z_i\}_{i=0}^{\infty} \text{ is a program}\right\}, \tag{1.1}$$

$$\pi^v(x) = \inf\left\{\liminf \sum_{i=0}^{N-1} [v(z_i, z_{i+1}) - \mu(v)] : \mathbf{z} \subset K,\ z_0 = x\right\}, \tag{1.2}$$

$$\theta^v(x, y) = v(x, y) - \mu(v) + \pi^v(y) - \pi^v(x) \tag{1.3}$$

for x, $y \in K$. It was proved in [1] that π^v and θ^v are continuous functions, θ^v is nonnegative, and $E(x) = \{y \in K : \theta^v(x, y) = 0\}$ is nonempty for every $x \in K$.

In [1] this result was established when K was a compact in $\mathbf{R}^n$, but its proof also remains in force when K is a compact metric space.

For a program $\mathbf{x}$ we denote by $\omega(\mathbf{x})$ the set of all points $z \in K$ such that some subsequence $\{x_{i_k}\}_{k=1}^{\infty}$ converges to z and by $\Omega(\mathbf{x})$ the set of all points $(z_1, z_2) \in K \times K$ such that some subsequence $\{(x_{i_k}, x_{i_k+1})\}_{k=1}^{\infty}$ converges to (z_1, z_2). Denote the metric in K by $d(x, y)$ $(x, y \in K)$, and define the metric d_1 on $K \times K$ by

$$d_1((x_1, x_2), (y_1, y_2)) = d(x_1, y_1) + d(x_2, y_2) \qquad (x_1, x_2, y_1, y_2 \in K).$$

We denote $d(x, B) = \inf\{d(x, y) : y \in B\}$ for $x \in K$, $B \subset K$, and

$$d_1((x_1, x_2), A) = \inf\{d_1((x_1, x_2), (y_1, y_2)) : (y_1, y_2) \in A\}$$

for $(x_1, x_2) \in K \times K$ and $A \subset K \times K$.

Denote the Hausdorff metric for two sets $A \subset K$ and $B \subset K$ by $\operatorname{dist}(A, B)$ and the cardinality of a set A by $\operatorname{Card}(A)$.

A sequence $\{x_i\}_{i=-\infty}^{\infty} \subset K$ is called *almost periodic* if for every $\varepsilon > 0$ there exists an integer $m \geq 1$ such that the relation $d(x_i, x_{i+pm}) \leq \varepsilon$ holds for any i and any p.

A program $\{x_i\}_{i=0}^{\infty}$ is called *asymptotic almost periodic* if for every $\varepsilon > 0$ there exist integers $k \geq 1$, $m \geq 1$, such that $d(x_i, x_{i+mj}) \leq \varepsilon$ for any $i \geq k$ and any $j \geq 1$.

For every $u \in \mathbf{C}(K \times K)$, every number $\Delta > 0$, and every integer $N \geq 1$ we define

$$A(u,N,\Delta) = \left\{ \{y_i\}_{i=0}^{N} \subset K : \text{for every sequence } \{z_i\}_{i=0}^{N} \subset K \text{ satisfying } z_0 = y_0,\ z_N = y_N \text{ the following inequality holds: } \sum_{i=0}^{N-1} [u(y_i, y_{i+1}) - u(z_i, z_{i+1})] \leq \Delta \right\}.$$

In [13] we proved the existence of a set $F_0 \subset \mathbf{C}(K \times K)$, which is a countable intersection of open everywhere dense sets in $\mathbf{C}(K \times K)$ and such that every $u \in F_0$ satisfies assertions a), b), and c) stated in the introduction. In this paper we will establish the existence of a set $F \subset F_0$ which is a countable intersection of open everywhere dense sets in $\mathbf{C}(K \times K)$ and for which the following theorems are valid.

THEOREM 1. *We define* $L : \mathbf{C}(K \times K) \to \mathbf{R}^1 \times \mathbf{C}(K) \times \mathbf{C}(K \times K)$ *by*

$$L(v) = (\mu(v), \pi^v, \theta^v)(v \in \mathbf{C}(K \times K)).$$

Then the set of continuity points of the operator L contains F.

By assertion a), for every $u \in F_0$ there exist compact sets $H(u) \subset K \times K$ and $H_0(u) \subset K$ such that $\Omega(\mathbf{x}) = H(u)$, $\omega(\mathbf{x}) = H_0(u)$ for every (u)-good program $\mathbf{x}$. Theorem 2 is an analogue of the weak turnpike theorem [9], [11], [12], showing that for $u \in F$, for every "nice" finite program $\{z_i\}_{i=0}^{N}$, the number of integers $i \in \{0, \dots N-1\}$ such that (x_i, x_{i+1}) does not belong to a fixed neighborhood of $H(u)$ is bounded by some constant which depends on the neighborhood and does not depend on N.

THEOREM 2. *Let* $u \in F$ *and* δ *be a positive number. Then there are a neighborhood* $W(u)$ *of* u *in* $\mathbf{C}(K \times K)$ *and positive numbers* Q_1, Q_2 *such that for every* $w \in W(u)$, *for every integer* $N \geq 1$, *for every number* $M > 0$ *and every program* $\{y_i\}_{i=0}^{N} \in A(w, N, M)$ *the following relation holds*:

$$\text{Card}\{i \in \{0, \dots, N-1\} : d_1((y_i, y_{i+1}), H(u)) > \delta\} \leq Q_1 + MQ_2.$$

Let $u \in F$. Theorem 3 establishes that for every w belonging to some small neighborhood of u in $\mathbf{C}(K \times K)$, for every program $\{x_i\}_{i=0}^{\infty}$ satisfying $\theta^w(x_i, x_{i+1}) = 0\,(i = 0, 1, \dots)$, the elements of the sequence $\{(x_i, x_{i+1})\}_{i=0}^{\infty}$ belong to a small neighborhood of $H(u)$ for all integers $i \geq N$, where N is a constant which depends on the neighborhoods and does not depend either on w or on x_0. In addition, for every w belonging to some small neighborhood of u in $\mathbf{C}(K \times K)$, for every program $\{x_i\}_{i=0}^{\infty}$ satisfying $\theta^w(x_i, x_{i+1}) = 0$, $(i = 0, 1, \dots)$ whose initial value belongs to some small neighborhood of $H_0(u)$, the sequence $\{(x_i, x_{i+1})\}_{i=0}^{\infty}$ is contained in a small neighborhood of $H(u)$.

Theorem 4 implies that for $w \in F$ every sequence $\{y_i\}_{i=-\infty}^{\infty}$ satisfying $\theta^w(y_i, y_{i+1}) = 0\,(i = 0, \pm 1, \dots)$ is almost periodic and shows that having any ε we can choose m (see the definition of an almost periodic program) uniformly for all w belonging to some small neighborhood of $u \in F$.

THEOREM 3. 1. *Let* $u \in F$ *and* ε *be a positive number. Then there exist a neighborhood* $W(u)$ *of* u *in* $\mathbf{C}(K \times K)$ *and* $\delta > 0$ *such that for every* $w \in W(u)$, *for every program* $\{x_i\}_{i=0}^{\infty}$ *satisfying* $\theta^w(x_i, x_{i+1}) = 0\,(i = 0, 1, \dots)$, $d(x_0, H_0(u)) \leq \delta$, *the relation* $d_1((x_i, x_{i+1}), H(u)) \leq \varepsilon$ *holds for* $i = 0, 1, \dots$.

2. *Let* $u \in F$, ε *be a positive number. Then there exist a neighborhood* $W(u)$ *of* u *in* $\mathbf{C}(K \times K)$ *and an integer* $N \geq 1$ *such that for every* $w \in W(u)$, *for every program* $\{x_i\}_{i=0}^{\infty}$ *satisfying* $\theta^w(x_i, x_{i+1}) = 0\,(i = 0, 1, \dots)$, *the following relation holds*: $d_1((x_i, x_{i+1}), H(u)) \leq \varepsilon$ (i *is an integer*, $i \geq N$).

COROLLARY 1. *Let $u \in F$, $\{x_i\}_{i=-\infty}^{\infty}$ be a program such that $\theta^u(x_i, x_{i+1}) = 0$ $(i = 0, \pm 1, \ldots)$. Then $(x_i, x_{i+1}) \in H(u)$ $(i = 0, \pm 1, \ldots)$.*

COROLLARY 2. *Let $u \in F$, $\varepsilon > 0$. Then there exists a neighborhood $W(u)$ of u in $\mathbf{C}(K \times K)$ such that for every $w \in W(u)$, for every program $\{x_i\}_{i=-\infty}^{\infty}$ satisfying $\theta^w(x_i, x_{i+1}) = 0$ $(i = 0, \pm 1, \ldots)$, the following relation holds: $d_1((x_i, x_{i+1}), H(u)) \leq \varepsilon$ $(i = 0, \pm 1, \ldots)$.*

THEOREM 4. *Let $u \in F$. Then every sequence $\{y_i\}_{i=-\infty}^{\infty}$ satisfying $\theta^u(y_i, y_{i+1}) = 0$ $(i = 0, \pm 1 \ldots)$ is almost periodic. Moreover, for every $\varepsilon > 0$ there exist a neighborhood $W(u)$ of u in $\mathbf{C}(K \times K)$ and an integer $m \geq 1$ such that for every $w \in W(u)$, for every program $\{y_i\}_{i=-\infty}^{\infty}$ satisfying $\theta^w(y_i, y_{i+1}) = 0$ $(i = 0, \pm 1, \ldots)$, the relation $d(y_i, y_{i+pm}) \leq \varepsilon$ holds for any integers i and p.*

2. Auxiliary results. We have the following result.

PROPOSITION 1 [1]. *Let $v \in \mathbf{C}(K \times K)$,*

$$\lambda(N, v) = \min \left\{ N^{-1} \sum_{i=0}^{N-1} v(z_i, z_{i+1}) : \{z_i\}_{i=0}^{N} \subset K,\ z_0 = z_N \right\},$$

$$\rho(N, v) = \min \left\{ N^{-1} \sum_{i=0}^{N-1} v(z_i, z_{i+1}) : \{z_i\}_{i=0}^{N} \subset K \right\} \qquad (N = 1, 2, \ldots).$$

Then $\rho(N, v) \leq \mu(v) \leq \lambda(N, v)$, $N(\lambda(N, v) - \rho(N, v)) \leq 2\|v\|$ $(N = 1, 2, \ldots)$, and for every program $\{z_i\}_{i=0}^{\infty}$ the sequence $\{\sum_{i=0}^{N-1} [v(z_i, z_{i+1}) - \mu(v)]\}_{N=1}^{\infty}$ either is bounded or diverges to infinity.

For a number $r > 0$ and a point x of some metric space we denote by $B(x, r)$ $(\bar{B}(x, r))$ the open (closed) ball in this space which has the center x and the radius r.

From here on in §§2–4 we consider a fixed function $v \in \mathbf{C}(K \times K)$ for which there exist an integer $m \geq 1$, a sequence $\{x_i^*\}_{i=0}^{m} \subset K$, a number μ, and functions $\pi \in \mathbf{C}(K)$, $\theta \in \mathbf{C}(K \times K)$ such that

1. for integers i, j satisfying $0 \leq i < j \leq m$ the equality $x_i^* = x_j^*$ holds if and only if $i = 0$, $j = m$;
2. $\|\theta\| > 0$, θ is nonnegative and for each x and $y \in K$ the equality $\theta(x, y) = 0$ holds if and only if $(x, y) \in \{(x_i^*, x_{i+1}^*) : i = 0, \ldots, m-1\}$;
3. $v(x, y) = \mu + \pi(x) - \pi(y) + \theta(x, y)(x, y \in K)$.

We denote by E the set of all such functions v. Lemma 2 in [13] implies that the set E is dense everywhere in $\mathbf{C}(K \times K)$. We define $x_i^* \in K$ for $i \in \{0, \pm 1, \ldots\} \backslash \{0, \ldots m\}$ such that $x_{m+i}^* = x_i^*$ $(i = 0, \pm 1, \ldots)$. For every number $\delta > 0$ we define

$$C_1(\delta) = \sup \left\{ \theta(x, y) : (x, y) \in \bigcup_{i=0}^{m-1} \bar{B}((x_i^*, x_{i+1}^*), \delta) \right\}, \tag{2.1}$$

$$C_2(\delta) = \inf \left\{ \theta(x, y) : (x, y) \in (K \times K) \backslash \bigcup_{i=0}^{m-1} B((x_i^*, x_{i+1}^*), \delta) \right\}, \tag{2.2}$$

$$C_3(\delta) = \sup \{ |\pi(x) - \pi(y)| : x, y \in K,\ d(x, y) \leq \delta \}. \tag{2.3}$$

We define

$$D_0 = 8^{-1} \inf\{d(x_i^*, x_j^*) : i, j \in \{0, \ldots, m-1\},\ i < j\} \qquad (\text{if } m = 1, \text{ then } D_0 = +\infty). \tag{2.4}$$

In order to establish Theorems 1 – 4 we need the following results, which were proved in [13].

LEMMA 1 [13, Lem. 3]. *Let $\delta \in (0, D_0)$, $r \in (0,(48m)^{-1}C_2(\delta))$, $\phi \in \mathbf{C}(K \times K)$, $\|\phi\| \le r$, $u = v + \phi$, $N \in \{1,2,\dots\}$, and $\{z_i\}_{i=0}^N$ be a program such that for every program $\{y_i\}_{i=0}^N$ satisfying $y_0 = z_0$, $y_N = z_N$, the relation $\sum_{i=0}^{N-1} u(z_i, z_{i+1}) \le 2\|\theta\| + \sum_{i=0}^{N-1} u(y_i, y_{i+1})$ holds. Let i_0, $j_0 \in \{1,2,\dots\}$, $0 < i_0 < j_0 < N$, $j_0 - i_0 \ge 240\|\theta\|C_2(\delta)^{-1}m$. Then there exists an integer $k \in \{i_0, \dots, j_0 - 6m\}$ for which $d(z_{k+i}, x_i^*) \le \delta\,(i = 0, \dots, 3m)$.*

For an integer $N \ge 1$, $\phi \in \mathbf{C}(K \times K)$ we define

$$\ell(\phi, N) = \inf\left\{ \sum_{i=0}^{N-1} (\phi + \theta)(y_i, y_{i+1}) : \{y_i\}_{i=0}^N \subset K \right\}. \tag{2.5}$$

LEMMA 2 [13, Lem. 6]. *Let $\varepsilon \in (0, D_0)$. Then there exists $R > 0$ such that for each $\phi \in \mathbf{C}(K \times K)$ satisfying $\|\phi\| \le R$, each integer $N \ge 1$, and each program $\{y_i\}_{i=0}^N$ satisfying $\sum_{i=0}^{N-1} (\theta + \phi)(y_i, y_{i+1}) \le \ell(\phi, N) + R$ there is an integer k such that $d_1((y_i, y_{i+1}), (x_{k+i}^*, x_{k+i+1}^*)) \le \varepsilon(i = 0, \dots, N-1)$.*

LEMMA 3 [13, Lem. 7]. *Let $\varepsilon \in (0, D_0)$. Then there is a number $r > 0$ such that for each integer $N \ge 1$, each integer k, and each $\phi \in \mathbf{C}(K \times K)$ satisfying $\|\phi\| \le r$ there is a program $\{z_i\}_{i=0}^N$ such that $z_0 = x_k^*$, $z_N = x_{k+N}^*$,*

$$d_1((z_i, z_{i+1}), (x_{k+i}^*, x_{k+i+1}^*)) \le \varepsilon(i = 0, \dots, N-1),$$
$$\sum_{i=0}^{N-1} (\phi + \theta)(z_i, z_{i+1}) \le \ell(\phi, N) + 6C_1(\varepsilon).$$

LEMMA 4 [13, Lem. 8]. *Let $\varepsilon_0 \in (0, D_0)$. Then there exists a number $r > 0$ such that for every $u \in \mathbf{C}(K \times K)$ satisfying $\|u - v\| \le r$, for every (u)-good program $\{z_i\}_{i=0}^\infty$, there are an integer $N > 1$ and an integer k such that*

$$d_1((z_{N+i}, z_{N+i+1}), \quad (x_{k+i}^*, x_{k+i+1}^*)) \le \varepsilon_0 \quad (i = 0, 1, \dots).$$

3. Preliminary lemmas for Theorem 2.

LEMMA 5. *Let $\Delta \in (0, D_0)$. Then there exist a number $r(\Delta) > 0$ and a integer $Q(\Delta) \ge 1$ such that for every $u \in \mathbf{C}(K \times K)$ satisfying $\|u - v\| \le r(\Delta)$, for every integer $N \ge 1$, and every program $\{y_i\}_{i=0}^N \in A(u, N, 2\|\theta\|)$ the following relation holds*:

$$\mathrm{Card}\left\{ i \in \{0, \dots, N-1\} : (y_i, y_{i+1}) \notin \bigcup_{j=0}^{m-1} \bar{B}((x_j^*, x_{j+1}^*), \Delta) \right\} \le Q(\Delta).$$

Proof. We choose positive numbers r_0, δ_1, and $r(\Delta)$ such that $r_0 < 1$, Lemma 2 holds with $\varepsilon = \Delta$ and $R = r_0$, $16\delta_1 \in (0, \Delta)$, $32C_1(2\delta_1) \le r_0$, $8r(\Delta) \in (0, \min\{\|\theta\|, r_0\})$; Lemma 1 holds with $\delta = \delta_1$ and $r = r(\Delta)\,(r(\Delta) < (48m)^{-1}C_2(\delta_1))$; and Lemma 3 holds with $\varepsilon = \delta_1$, $r = r(\Delta)$. Set

$$Q(\Delta) = \inf\{i \in \{1,2,\dots\} : i > (1 + \|\theta\|)m\|\theta\|(C_2(\delta_1)r_0)^{-1}14 \cdot 10^3\}.$$

Let $u \in \mathbf{C}(K \times K)$, $\|u - v\| \le r(\Delta)$, N be an integer, $N \ge 1$, $X = \{x_i\}_{i=0}^N \in A(u, N, 2\|\theta\|)$, and $\phi = u - v$. By Lemma 1, for each integers i_0 and j_0 such that $0 < i_0 < j_0 < N$, $j_0 - i_0 \ge 240\|\theta\|C_2(\delta_1)^{-1}m$, there exists $k \in \{i_0, \dots, j_0 - 6m\}$ such that $d(x_{k+i}, x_i^*) \le \delta_1\,(i = 0, \dots, 3m)$.

For $Y = \{y_i\}_{i=0}^{k} \subset K$ define $G(Y) = k$,

$$P(Y) = \mathrm{Card}\left\{ i \in \{0, \dots, k-1\} : (y_i, y_{i+1}) \notin \bigcup_{j=0}^{m-1} \bar{B}((x_j^*, x_{j+1}^*), \Delta) \right\}.$$

We shall say that a program $Y = \{y_i\}_{i=0}^{G(Y)}$ has a property (A) if for each integers i_0 and j_0 such that $0 < i_0 < j_0 < G(Y)$, $j_0 - i_0 \geq 240\|\theta\|C_2(\delta_1)^{-1}m$, there exists an integer $k \in \{i_0, \dots, j_0 - 3m\}$ for which $d(y_{k+i}, x_i^*) \leq \delta_1$ $(i = 0, \dots, 3m)$.

Consider a program $Y = \{y_i\}_{i=0}^{G(Y)}$ which has the property (A) and for which there is an integer $p \geq 1$ such that

$$\begin{aligned} &800m\|\theta\|C_2(\delta_1)^{-1} < p < G(Y) - 800m\|\theta\|C_2(\delta_1)^{-1}, \\ &(y_p, y_{p+1}) \notin \bigcup_{j=0}^{m-1} \bar{B}((x_j^*, x_{j+1}^*), \Delta). \end{aligned} \tag{3.1}$$

The property (A) which holds for Y implies the existence of integers g and q such that

$$\begin{aligned} &[g, g+3m] \subset [p+1, p+250m\|\theta\|C_2(\delta_1)^{-1}], \\ &[q, q+3m] \subset [p - 250m\|\theta\|C_2(\delta_1)^{-1}, p-1], \\ &d(y_{g+i}, x_i^*) \leq \delta_1, \quad d(y_{q+i}, x_i^*) \leq \delta_1 \qquad (i = 0, \dots, 3m). \end{aligned}$$

Consider a program $\Gamma(Y) = \{t_i\}_{i=0}^{G(\Gamma(Y))}$, where $G(\Gamma(Y)) = G(Y) - g + q + 3m$, $t_i = y_i$ $(i = 0, \dots, q+3m)$, $t_i = y_{i+g-q-3m}$ $(i = q+3m+1, \dots, G(\Gamma(Y)))$. Then

$$G(Y) > G(\Gamma(Y)) \geq G(Y) - 500m\|\theta\|C_2(\delta_1)^{-1}, \; P(\Gamma(Y)) < P(Y), \tag{3.2}$$

and the program $\Gamma(Y)$ has the property (A). We have

$$\begin{aligned} \sum_{i=0}^{G(\Gamma(Y))-1} [u(t_i, t_{i+1}) - \mu(u)] \leq &\sum_{i=0}^{G(Y)-1} [u(y_i, y_{i+1}) - \mu(u)] \\ &- \sum_{i=q+3m}^{g-1} (\theta + \phi)(y_i, y_{i+1}) + 2r(\Delta) + C_1(2\delta_1). \end{aligned} \tag{3.3}$$

By (3.1) and Lemma 2, which holds with $\varepsilon = \Delta$ and $R = r_0$, we have

$$\begin{aligned} \sum_{i=q+3m}^{g-1} (\theta + \phi)(y_i, y_{i+1}) &\geq \ell(\phi, g - q - 3m) + r_0 \\ &= \ell(\phi, G(Y) - G(\Gamma(Y))) + r_0, \end{aligned}$$

and by (3.3)

$$\begin{aligned} \sum_{i=0}^{G(\Gamma(Y))-1} [u(t_i, t_{i+1}) - \mu(u)] \leq &\sum_{i=0}^{G(Y)-1} [u(y_i, y_{i+1}) - \mu(u)] \\ &- \ell(\phi, G(Y) - G(\Gamma(Y))) - r_0 + 2r(\Delta) + C_1(2\delta_1). \end{aligned} \tag{3.4}$$

Consider again the program $X = \{x_i\}_{i=0}^{N}$. If there is no integer $p \geq 1$ satisfying (3.1) with $Y = X$, then $P(X) \leq 2 \cdot 10^3 m \|\theta\| C_2(\delta_1)^{-1}$ and for X the lemma holds. Assume that there is an integer $p \geq 1$ satisfying (3.1) with $Y = X$. We set $X^0 = X$. The program X^0 has the property (A), and we define $X^1 = \Gamma(X^0)$. Clearly, X^1 has the property (A). By induction we define a sequence of programs $\{X^i\}$. Suppose that we have already defined $X^i (i = 0, \dots, k)$ such that for $i = 0, \dots, k-1$ the following conditions hold:

1. X^i has the property (A);
2. there exists an integer $p \geq 1$ such that relation (3.1) holds with $Y = X^i$;
3. $X^{i+1} = \Gamma(X^i)$.

If there is no integer $p \geq 1$ such that for $Y = X^k$ relation (3.1) holds, then X^k is the last element of the sequence and its construction is completed. Otherwise, we define $X^{k+1} = \Gamma(X^k)$. By (3.2), the construction of the sequence will be completed in a finite number of steps. Let X^q be the last element of the sequence. It is easy to see that

$$G(X^q) \geq 10^3 m \|\theta\| C_2(\delta_1)^{-1}, \qquad P(x^q) \leq 2 \cdot 10^3 m \|\theta\| C_2(\delta_1)^{-1}. \tag{3.5}$$

We define $g_i = G(X^0) - G(X^i)\,(i = 0, \dots, q), X^q = \{x_{iq}\}_{i=0}^{G(X^q)}$. Lemma 3, which holds for $\varepsilon = \delta_1$ and $r = \delta_1$, implies the existence of a program $\{z_i : i = 0, \dots, g_q\}$ such that

$$\begin{aligned}\sum_{i=g_s}^{g_{s+1}-1} (\theta + \phi)(z_i, z_{i+1}) &\leq \ell(\phi, g_{s+1} - g_s) + 6C_1(\delta_1) \\ &\leq \ell(\phi, G(X^s) - G(X^{s+1})) + 6C_1(\delta_1) \qquad (s = 0, \dots, q-1).\end{aligned}$$

By this inequality, relation (3.4), and the choice of r_0, $r(\Delta)$, and δ_1,

$$\begin{aligned}
&\sum_{i=0}^{G(X^q)-1} [u(x_{iq}, x_{(i+1)q}) - \mu(u)] \\
&\quad \leq \sum_{i=0}^{G(X)-1} [u(x_i, x_{i+1}) - \mu(u)] \\
&\quad\quad - \sum_{s=0}^{q-1} [\ell(\phi, G(X^s) - G(X^{s+1})) + r_0 - 2r(\Delta) - C_1(2\delta_1)] \\
&\quad \leq \sum_{i=0}^{G(X)-1} [u(x_i, x_{i+1}) - \mu(u)] - q(r_0 - 2r(\Delta) - C_1(2\delta_1)) \qquad (3.6)\\
&\quad\quad - \sum_{s=0}^{q-1} \left[-6C_1(\delta_1) + \sum_{i=g_s}^{g_{s+1}-1} (\phi + \theta)(z_i, z_{i+1}) \right] \\
&\quad \leq \sum_{i=0}^{G(X)-1} [u(x_i, x_{i+1}) - \mu(u)] - q(r_0 - 2r(\Delta) - 8C_1(2\delta_1)) \\
&\quad\quad - \sum_{i=0}^{G(X)-G(X^q)-1} (\phi + \theta)(z_i, z_{i+1}) \\
&\quad \leq \sum_{i=0}^{G(X)-1} [u(x_i, x_{i+1}) - \mu(u)] - \sum_{i=0}^{G(X)-G(X^q)-1} (\phi + \theta)(z_i, z_{i+1}) - 2^{-1} q r_0.
\end{aligned}$$

Consider a program $(f_0, \ldots, f_{G(X)})$, where $f_i = x_{iq}$ $(i = 0, \ldots, G(X^q) - 1)$, $f_{G(X^q)+i} = z_{i+1}$ $(i = 0, \ldots, G(X) - G(X^q) - 1)$, $f_{G(X)} = x_{G(X)}$. Evidently, $f_0 = x_0$, $x_{G(X^q),q} = x_{G(X)}$. Relation (3.6) implies that

$$\begin{aligned}
\sum_{i=0}^{G(X)-1} [u(x_i, x_{i+1}) - \mu(u)] - 2\|\theta\| \leq \sum_{i=0}^{G(X)-1} [u(f_i, f_{i+1}) - \mu(u)]& \\
= \sum_{i=0}^{G(X^q)-2} [u(x_{iq}, x_{(i+1)q}) - \mu(u)]& \\
+ u(x_{(G(X^q)-1)q}, z_1) - \mu(u)& \\
+ \sum_{i=1}^{G(X)-G(X^q)-1} [u(z_i, z_{i+1}) - \mu(u)]& \\
+ u(z_{G(X)-G(X^q)}, x_{G(X)}) - \mu(u)& \\
= \sum_{i=0}^{G(X^q)-1} [u(x_{iq}, x_{(i+1)q}) - \mu(u)]& \\
+ \sum_{i=0}^{G(X)-G(X^q)-1} (\phi + \theta)(z_i, z_{i+1})& \\
+ [u(x_{(G(X^q)-1)q}, z_1) - u(x_{(G(X^q)-1)q}, x_{G(X^q)q})]& \\
- (\phi + \theta)(z_0, z_1) + \pi(z_1) - \pi(z_{G(X)-G(X^q)})& \\
+ (\theta + \phi)(z_{G(X)-G(X^q)}, x_{G(X)})& \\
+ \pi(z_{G(X)-G(X^q)}) - \pi(x_{G(X)})& \\
\leq \sum_{i=0}^{G(X^q)-1} [u(x_{iq}, x_{(i+1)q}) - \mu(u)]& \\
+ \sum_{i=0}^{G(X)-G(X^q)-1} (\phi + \theta)(z_i, z_{i+1}) + 8\|\theta\| + 8r(\Delta)& \\
\leq \sum_{i=0}^{G(X)-1} [u(x_i, x_{i+1}) - \mu(u)] - qr_0 2^{-1} + 8\|\theta\| + 8r(\Delta),&
\end{aligned}$$

$$2^{-1}qr_0 \leq 10\|\theta\| + 8r(\Delta), \qquad q \leq (24\|\theta\|)r_0^{-1}.$$

Relations (3.2) and (3.5) imply that

$$\begin{aligned}
P(X) &\leq P(X^q) + G(X) - G(X^q) + 1 \\
&\leq q500m\|\theta\|C_2(\delta_1)^{-1} + 2 \cdot 10^3 m\|\theta\|C_2(\delta_1)^{-1} + 1 \\
&\leq m\|\theta\|C_2(\delta_1)^{-1}(2 \cdot 10^3 + 12 \cdot 10^3\|\theta\|r_0^{-1}) + 1 \leq Q(\Delta).
\end{aligned}$$

The lemma is proved.

LEMMA 6. *Let $\Delta \in (0, D_0)$. Then there exist positive numbers $r(\Delta)$, $Q_1(\Delta)$, and $Q_2(\Delta)$ such that for every natural number N, for every positive number M, for every $u \in \mathbf{C}(K \times K)$ satisfying $\|u - v\| \leq r(\Delta)$, and for every program $\{y_i\}_{i=0}^N \in A(u, N, M)$, the following relation holds*:

$$\operatorname{Card}\left\{ i \in \{0, \dots, N-1\} : (y_i, y_{i+1}) \notin \bigcup_{j=0}^{m-1} \bar{B}((x_j^*, x_{j+1}^*), \Delta) \right\} \leq Q_1(\Delta) + Q_2(\Delta)M.$$

Proof. We choose an integer $Q(\Delta) \geq 1$ and a positive number $r(\Delta)$ such that for Δ, $Q(\Delta)$, and $r(\Delta)$ Lemma 5 holds, and define $Q_1(\Delta) = 1 + Q(\Delta)$, $Q_2(\Delta) = (1 + Q(\Delta))(2\|\theta\|)^{-1}$. Let N be an integer, $N \geq 1$, $M > 0$, $u \in \mathbf{C}(K \times K)$, $\|u - v\| \leq r(\Delta)$, $\{y_i\}_{i=0}^N \in A(u, N, M)$. If $M \leq 2\|\theta\|$, then Lemma 6 follows from Lemma 5. Consider the case $M > 2\|\theta\|$. By induction we obtain a sequence of integers $\{i_p\}_{p=0}^k$ such that $i_0 = 0 < \cdots < i_p < \cdots < i_k = N$; for every $p \in \{0, \dots, k-1\}$ the relation

$$\{y_j\}_{j=i_p}^q \in A(u, q - i_p, 2\|\theta\|) \qquad (i_p < q < i_{p+1}) \tag{3.7}$$

holds; and for every integer p satisfying $0 \leq p \leq k-2$ the relation

$$\{y_j\}_{j=i_p}^{i_{p+1}} \notin A(u, i_{p+1} - i_p, 2\|\theta\|)$$

holds. This relation implies that $(k-1)2\|\theta\| \leq M$. It is easy to see that $i_{p+1} - i_p \geq 2$ for every integer p satisfying $0 \leq p \leq k-2$. Relation (3.7) and Lemma 5, which holds for Δ, $Q(\Delta)$, and $r(\Delta)$, imply that

$$\begin{aligned}&\operatorname{Card}\left\{ i \in \{0, \dots, N-1\} : (y_i, y_{i+1}) \notin \bigcup_{j=0}^{m-1} \bar{B}((x_j^*, x_{j+1}^*), \Delta) \right\} \\ &\quad \leq k(1 + Q(\Delta)) \leq (1 + M(2\|\theta\|)^{-1})(1 + Q(\Delta)) \leq Q_1(\Delta) + MQ_2(\Delta).\end{aligned}$$

The lemma is proved.

4. Preliminary lemmas for proof of Theorems 1, 3, and 4.

LEMMA 7. *Let $\varepsilon \in (0, D_0)$. Then there exists a number $r > 0$ such that for every $\phi \in \mathbf{C}(K \times K)$ satisfying $\|\phi\| \leq r$ and for every integer k there is a program $\{z_i\}_{i=0}^\infty$ for which the following relations hold*:

$$z_0 = x_k^*,\ d_1((z_i, z_{i+1}), (x_{k+i}^*, x_{k+i+1}^*)) \leq \varepsilon \qquad (i = 0, 1, \dots),$$

$$\sum_{i=p}^{q-1} (\theta + \phi)(z_i, z_{i+1}) \leq \ell(\phi, q - p) + 18C_1(2\varepsilon) \qquad (p,\ q \textit{ are integers}, 0 \leq p < q).$$

Proof. Choose $r \in (0, C_1(\varepsilon))$ such that Lemma 3 holds for ε, r. Let k be an integer, $\phi \in \mathbf{C}(K \times K)$, $\|\phi\| \leq r$. By Lemma 3, for any integer $N \geq 1$ there exists a program $\{z_i(N)\}_{i=0}^N$ such that $z_0(N) = x_k^*$, $z_N(N) = x_{k+N}^*$,

$$\begin{aligned}&d_1((z_i(N), z_{i+1}(N)), (x_{k+i}^*, x_{k+i+1}^*)) \leq \varepsilon \qquad (i = 0, \dots, N-1), \\ &\sum_{i=0}^{N-1} (\phi + \theta)(z_i(N), z_{i+1}(N)) \leq \ell(\phi, N) + 6C_1(\varepsilon).\end{aligned} \tag{4.1}$$

We fix an integer $N \geq 1$, and let p and q be integers satisfying $0 \leq p < q < N$. We will show that

$$\sum_{i=p}^{q-1} (\phi + \theta)(z_i(N), z_{i+1}(N)) \leq \ell(\phi, q-p) + 18C_1(2\varepsilon). \tag{4.2}$$

By Lemma 3 there is a program $\{y_i\}_{i=0}^{q-p}$ such that $y_0 = x^*_{k+p}$, $y_{q-p} = x^*_{k+q}$,

$$\begin{gathered}\sum_{i=0}^{q-p-1} (\phi + \theta)(y_i, y_{i+1}) \leq \ell(\phi, q-p) + 6C_1(\varepsilon), \\ d_1((y_i, y_{i+1}), (x^*_{i+k+p}, x^*_{i+k+p+1})) \leq \varepsilon \qquad (i = 0, \dots, q-p-1).\end{gathered} \tag{4.3}$$

Consider a program $\{t_i\}_{i=0}^{N}$, where $t_i = z_i(N)$ $(i \in \{0, \dots, p\} \cup \{q, \dots, N\})$, $t_i = y_{i-p}$ $(p < i < q)$. By (4.1) and (4.3)

$$\begin{aligned} 6C_1(\varepsilon) &\geq \sum_{i=0}^{N-1} [(\phi + \theta)(z_i(N), z_{i+1}(N)) - (\phi + \theta)(t_i, t_{i+1})] \\ &= \sum_{i=p}^{q-1} [(\phi + \theta)(z_i(N), z_{i+1}(N)) - (\phi + \theta)(t_i, t_{i+1})], \end{aligned}$$

$$\begin{aligned} \sum_{i=p}^{q-1} (\phi + \theta)(t_i, t_{i+1}) &\leq \sum_{i=0}^{q-p-1} (\phi + \theta)(y_i, y_{i+1}) + 4r + 2C_1(2\varepsilon) \\ &\leq \ell(\phi, q-p) + 12C_1(2\varepsilon), \end{aligned}$$

$$\begin{aligned} \sum_{i=p}^{q-1} [(\phi + \theta)(z_i(N), z_{i+1}(N))] &\leq 6C_1(\varepsilon) + \sum_{i=p}^{q-1} (\phi + \theta)(t_i, t_{i+1}) \\ &\leq \ell(\phi, q-p) + 18C_1(2\varepsilon). \end{aligned}$$

Relation (4.2) is proved. It is easy to see that there is a sequence of integers $\{N_j\}_{j=1}^{\infty}$ such that $1 \leq N_1 < \cdots < N_j < N_{j+1} < \cdots$, and for every integer $i \geq 0$ the sequence $z_i(N_j) \to z_i \in K$ as $j \to \infty$. By (4.1) and (4.2) the program $\{z_i\}_{i=0}^{\infty}$ satisfies the lemma conditions. The lemma is proved.

From now on we assume that

$$\pi(x^*_0) = \sup\{\pi(x^*_i) : i = 0, \dots, m\}. \tag{4.4}$$

Let $x \in K$, $u \in \mathbf{C}(K \times K)$, $\varepsilon > 0$, $Q \in \{1, 2, \dots\}$. Set

$$\begin{aligned} H_0(x) &= \{\{z_i\}_{i=0}^{\infty} \subset K : z_0 = x\}, \\ H_1(x, u, Q) &= \{\{z_i\}_{i=0}^{\infty} \in H_0(x) : \{z_i\}_{i=0}^{Q} \in A(u, Q, 2\|\theta\|)\}, \\ H_2(x, Q, \varepsilon) &= \{\{z_i\}_{i=0}^{\infty} \in H_0(x) : \text{there exists an integer } g \in \{0, \dots, Q\} \\ &\qquad \text{such that } d(z_g, x^*_{m-1}) \leq \varepsilon\}, \\ H_3(x, u, Q, \varepsilon) &= \{\{z_i\}_{i=0}^{\infty} \in H_2(x, Q, \varepsilon) : \{z_i\}_{i=0}^{\infty} \text{is a } (u)\text{-good program}\}. \end{aligned}$$

LEMMA 8. *Let $\delta \in (0, D_0)$, $r \in (0, (48m)^{-1}C_2(\delta))$, Q be an integer, and $Q \geq 400m\|\theta\|C_2(\delta)^{-1}$. Then for every $u \in \mathbf{C}(K \times K)$ satisfying $\|u - v\| \leq r$, for every $x \in K$, the following relation holds*:

$$\pi^u(x) = \inf\left\{\liminf \sum_{i=0}^{N-1} [u(z_i, z_{i+1}) - \mu(u)] : \{z_i\}_{i=0}^{\infty} \in H_3(x, u, Q, \delta)\right\}.$$

Proof. Let $u \in \mathbf{C}(K \times K)$, $\|u - v\| \le r$, $x \in K$, $\phi = u - v$. Then

$$\pi^u(x) = \inf \left\{ \liminf \sum_{i=0}^{N-1} [u(z_i, z_{i+1}) - \mu(u)] : \{z_i\}_{i=0}^{\infty} \in H_1(x, u, Q) \right\}.$$

By Lemma 1, $H_1(x, u, Q) \subset H_2(x, Q, \delta) \subset H_0(x)$. This relation implies that

$$\pi^u(x) = \inf \left\{ \liminf \sum_{i=0}^{N-1} [u(z_i, z_{i+1}) - \mu(u)] : \{z_i\}_{i=0}^{\infty} \in H_2(x, Q, \delta) \right\}.$$

The lemma now follows from Proposition 1 and the definition of a (u)-good program.

LEMMA 9. *Let $\varepsilon \in (0, D_0)$. Then there exists a number $r > 0$ such that for every $\phi \in \mathbf{C}(K \times K)$ satisfying $\|\phi\| \le r$ the following conditions hold*:

$$\mu(\phi + v) = \mu + \mu(\phi + \theta);$$

there is a $(\phi + v)$-good program $\{z_i^\phi\}_{i=0}^{\infty}$ such that

$$z_0^\phi = x_0^*, \quad d_1((z_i^\phi, z_{i+1}^\phi), (x_i^*, x_{i+1}^*)) \le \varepsilon \quad (i = 0, 1, \ldots), \tag{4.5}$$

$$\sum_{i=p}^{q-1} (\phi + \theta)(z_i^\phi, z_{i+1}^\phi) \le \ell(\phi, q - p) + 18C_1(2\varepsilon) \tag{4.6}$$
$$(p \textit{ and } q \textit{ are integers}, 0 \le p < q),$$

$$\left| N\mu(\phi + \theta) - \sum_{i=0}^{N-1} (\phi + \theta)(z_i^\phi, z_{i+1}^\phi) \right| \le 30C_1(2\varepsilon) \qquad (N = 1, 2, \ldots), \tag{4.7}$$

$$\left| \liminf \sum_{i=0}^{N-1} [(\phi + v)(z_i^\phi, z_{i+1}^\phi) - \mu(\phi + v)] \right| \le C_3(\varepsilon) + 30C_1(2\varepsilon). \tag{4.8}$$

Proof. We choose a number $r \in (0, (2m)^{-1}C_1(\varepsilon))$ such that Lemma 7 holds for ε, r. Let $\phi \in \mathbf{C}(K \times K)$, $\|\phi\| \le r$. By Lemma 7 there is a program $\{z_i^\phi\}_{i=0}^{\infty}$ for which $z_0^\phi = x_0^*$ and relations (4.5) and (4.6) hold. It is easy to see that $\ell(\phi, N) = N\rho(N, \phi + \theta)$ $(N = 1, 2, \ldots)$ (see Proposition 1 and (2.5)). Let q be an integer, $q \ge 1$, and $\{y_i\}_{i=0}^{qm}$ be a program such that $y_i = z_i^\phi$ $(i = 0, \ldots, qm - 1)$, $y_{qm} = x_0^*$. It is easy to see that

$$\begin{aligned}
qm\lambda(qm, \phi + \theta) &\le \sum_{i=0}^{qm-1} (\phi + \theta)(y_i, y_{i+1}) \\
&\le \sum_{i=0}^{qm-1} (\phi + \theta)(z_i^\phi, z_{i+1}^\phi) + 2r + C_1(2\varepsilon) \\
&\le 2r + C_1(2\varepsilon) + \ell(\phi, qm) + 18C_1(2\varepsilon) \qquad (4.9)\\
&= 2r + 19C_1(2\varepsilon) + qm\rho(qm, \phi + \theta),
\end{aligned}$$

$$\begin{aligned}
qm\lambda(qm, \phi + \theta) - 3C_1(2\varepsilon) &\le \sum_{i=0}^{qm-1} (\phi + \theta)(z_i^\phi, z_{i+1}^\phi) \\
&\le 18C_1(2\varepsilon) + qm\rho(qm, \phi + \theta).
\end{aligned}$$

Let N be an integer, $N > m$. There is an integer $p \geq 1$ such that $pm \leq N < pm + m$. We have by (4.6), Proposition 1, and (4.9)

$$\sum_{i=0}^{N-1} (\phi + \theta)(z_i^\phi, z_{i+1}^\phi) \leq 18C_1(2\varepsilon) + \ell(\phi, N) \leq 18C_1(2\varepsilon) + N\mu(\phi + \theta),$$

$$\begin{aligned}(p+1)m\mu(\phi+\theta) &\leq (p+1)m\lambda((p+1)m,\ \phi+\theta)\\ &\leq \sum_{i=0}^{(p+1)m-1} (\phi+\theta)(z_i^\phi, z_{i+1}^\phi) + 3C_1(2\varepsilon)\\ &\leq \sum_{i=0}^{N-1} (\phi+\theta)(z_i^\phi, z_{i+1}^\phi) + 3C_1(2\varepsilon) + \ell(\phi, (p+1)m - N) + 18C_1(2\varepsilon),\end{aligned}$$

$$\begin{aligned}&\sum_{i=0}^{N-1} (\phi+\theta)(z_i^\phi, z_{i+1}^\phi)\\ &\qquad \geq (p+1)m\mu(\phi+\theta) - 21C_1(2\varepsilon) - \ell(\phi, (p+1)m - N)\\ &\qquad \geq -21C_1(2\varepsilon) + N\mu(\phi+\theta).\end{aligned}$$

Now it is easy to see that (4.7) holds for every $N > m$.

Consider the case with $N \leq m$. We have

$$|\rho(\phi+\theta, N)| = |\rho(\phi+\theta, N) - \rho(\theta, N)| \leq \|\phi\| \leq r, \qquad |\ell(\phi, N)| \leq mr,$$

$$|\mu(\phi+\theta)| = |\mu(\phi+\theta) - \mu(\theta)| \leq \|\phi\| \leq r,$$

and thus the validity of (4.7) for N follows from these relations, from relation (4.6), and from the inequality $2mr < C_1(\varepsilon)$. Thus, we have proved that (4.7) holds for every integer $N \geq 1$.

We will show that $\{z_i^\phi\}_{i=0}^\infty$ is a $(\phi + v)$-good program. Assume the contrary. Then by Proposition 1

$$\sum_{i=0}^{N-1} [(\phi+v)(z_i^\phi, z_{i+1}^\phi) - \mu(\phi+v)] \to +\infty \quad \text{as } N \to \infty.$$

Let $\{y_i\}_{i=0}^\infty$ be a $(\phi + v)$-good program. Then

$$\sum_{i=0}^{N-1} [(\phi+v)(z_i^\phi, z_{i+1}^\phi) - (\phi+v)(y_i, y_{i+1})] \to +\infty \quad \text{as } N \to \infty.$$

On the other hand, relation (4.6) implies that

$$\begin{aligned}&\sum_{i=0}^{N-1} [(\phi+v)(z_i^\phi, z_{i+1}^\phi) - (\phi+v)(y_i, y_{i+1})]\\ &\qquad \leq \sum_{i=0}^{N-1} [(\phi+\theta)(z_i^\phi, z_{i+1}^\phi) - (\phi+\theta)(y_i, y_{i+1})] + 4\|\pi\|\\ &\qquad \leq 18C_1(2\varepsilon) + 4\|\pi\| \qquad (N = 1, 2, \ldots).\end{aligned}$$

The contradiction obtained proves that $\{z_i^\phi\}_{i=0}^\infty$ is a $(\phi+v)$-good program. Then relation (4.7) implies that

$$\begin{aligned}\mu(\phi+v) &= \lim_{N\to\infty} N^{-1}\sum_{i=0}^{N-1}(\phi+v)(z_i^\phi,z_{i+1}^\phi)\\ &= \mu + \lim_{N\to\infty} N^{-1}\sum_{i=0}^{N-1}(\phi+\theta)(z_i^\phi,z_{i+1}^\phi) = \mu+\mu(\phi+\theta).\end{aligned}$$

Now, to complete the proof, we should only prove (4.8). We have for $N=1,2,\dots$ using (4.5), (4.7), and (2.3),

$$\begin{aligned}&\sum_{i=0}^{N-1}[(\phi+v)(z_i^\phi,z_{i+1}^\phi)-\mu(\phi+v)]\\ &\quad = \pi(x_0^*)-\pi(z_N^\phi)+\sum_{i=0}^{N-1}[(\phi+\theta)(z_i^\phi,z_{i+1}^\phi)-\mu(\phi+\theta)],\end{aligned}$$

$$\begin{aligned}&\left|\sum_{i=0}^{N-1}[(\phi+v)(z_i^\phi,z_{i+1}^\phi)-\mu(\phi+v)]-\pi(x_0^*)+\pi(x_N^*)\right|\\ &\quad\le C_3(\varepsilon)+30C_1(2\varepsilon).\end{aligned}$$

This relation and (4.4) imply that (4.8) holds. The lemma is proved.

LEMMA 10. *Let $\Delta\in(0,D_0)$. Then there is $h>0$ such that for every $u\in\mathbf{C}(K\times K)$ satisfying $\|u-v\|\le h$ the relation $\|\pi^u-\pi^v\|\le\Delta$ holds.*

Proof. Choose $\delta_0\in(0,\Delta)$ satisfying $4(220C_1(2\delta_0)+4C_3(2\delta_0))\le\Delta$, an integer $Q\ge 400m\|\theta\|C_2(\delta_0)^{-1}$, and $h\in(0,(48m)^{-1}C_2(\delta_0))$ such that $(Q+4)8h\le\Delta$, Lemma 9 holds for $\varepsilon=\delta_0$, $r=h$, Lemma 4 holds for $\varepsilon_0=\delta_0$, and $r=h$. Let $\phi\in\mathbf{C}(K\times K)$, $\|\phi\|\le h$, $u=v+\phi$, and $x\in K$. Lemma 8 implies that

$$\pi^u(x)=\inf\left\{\liminf\sum_{i=0}^{N-1}[u(z_i,z_{i+1})-\mu(u)] : \{z_i\}_{i=0}^\infty\in H_3(x,u,Q,\delta_0)\right\}.$$

Lemma 9, which holds for $\varepsilon=\delta_0$, $r=h$, implies the existence of a $(\phi+v)$-good program $\{z_i^\phi\}_{i=0}^\infty$ such that $z_0^\phi=x_0^*$ and relations (4.5)-(4.8) hold for $\varepsilon=\delta_0$. By H_4 denote the set of all programs $\{z_i\}_{i=0}^\infty$ such that $z_0=x$, $d(z_g,x_{m-1}^*)\le\delta_0$, $z_{g+i}=z_{i-1}^\phi$ $(i=1,2,\dots)$ for some $g\in\{0,\dots,Q\}$. It is easy to see that $H_4\subset H_3(x,u,Q,\delta_0)$. Let $\{y_i\}_{i=0}^\infty\in H_3(x,u,Q,\delta_0)$. It is a (u)-good program, and there is $q\in\{0,\dots,Q\}$ satisfying $d(y_q,x_{m-1}^*)\le\delta_0$.. Consider a program $\{a_i\}_{i=0}^\infty$, where $a_i=y_i(i=0,\dots,q)$, $a_i=z_{i-q-1}^\phi(i\ge q+1)$. It is easy to see that $\{a_i\}_{i=0}^\infty\in H_4$. So $\{y_i\}_{i=0}^\infty$ is a (u)-good program. By Lemma 4, which is valid for $\varepsilon_0=\delta_0$, $r=h$, there exists an integer $N\ge1$ such that

$$(4.10)\qquad d_1((y_{N+i},y_{N+i+1}),(x_i^*,x_{i+1}^*))\le\delta_0\qquad(i=0,1\dots).$$

We suppose, without loss of generality, that $N\ge 4q+4$. There is an integer $p\ge0$ such that $q+1+pm\le N<q+1+(p+1)m$. For $g=0,1,\dots$ define $s(g)=(p+1)m+g$. Fix integer $g\ge0$. Lemma 9, which holds for $\varepsilon=\delta_0$, $r=h$, implies that $\mu(u)=\mu(\phi+\theta)+\mu$,

$$\begin{aligned}
&\sum_{i=0}^{N+g-1}[u(y_i,y_{i+1})-\mu(u)]-\sum_{i=0}^{q+s(g)}[u(a_i,a_{i+1})-\mu(u)]\\
&\quad=\sum_{i=q}^{N+g-1}[u(y_i,y_{i+1})-\mu(u)]-\sum_{i=q}^{q+s(g)}[u(a_i,a_{i+1})-\mu(u)]\\
&\quad=\sum_{i=q}^{N+g-1}[(\phi+\theta)(y_i,y_{i+1})-\mu(\phi+\theta)]-\sum_{i=q}^{q+s(g)}[(\phi+\theta)(a_i,a_{i+1})-\mu(\phi+\theta)]\\
&\qquad+\pi(z^{\phi}_{s(g)})-\pi(y_{N+g})\\
&\quad=\sum_{i=q}^{N+g-1}[(\phi+\theta)(y_i,y_{i+1})-(\phi+\theta)(a_i,a_{i+1})]\\
&\qquad-\sum_{i=N+g}^{q+s(g)}[(\phi+\theta)(a_i,a_{i+1})-\mu(\phi+\theta)]+\pi(z^{\phi}_{s(g)})-\pi(y_{N+g}).
\end{aligned}\tag{4.11}$$

We will estimate the last expression and each of its terms. By (4.5), which holds for $\varepsilon=\delta_0$, and (4.10) we have

$$d(z^{\phi}_{s(g)},x^*_g)\le\delta_0,\qquad d(y_{N+g},x^*_g)\le\delta_0,\qquad |\pi(z^{\phi}_{s(g)})-\pi(y_{N+g})\,|\le C_3(2\delta_0).\tag{4.12}$$

It follows from the definition of $\{a_i\}_{i=0}^{\infty}$ and (4.7) which holds for $\varepsilon=\delta_0$, that

$$\begin{aligned}
&\left|\sum_{i=N+g}^{q+s(g)}[(\phi+\theta)(a_i,a_{i+1})-\mu(\phi+\theta)]\right|\\
&\quad=\left|\sum_{i=N+g-q-1}^{s(g)-1}[(\phi+\theta)(z^{\phi}_i,z^{\phi}_{i+1})-\mu(\phi+\theta)]\right|\\
&\quad\le\left|\sum_{i=0}^{s(g)-1}[(\phi+\theta)(z^{\phi}_i,z^{\phi}_{i+1})-\mu(\phi+\theta)]\right|\\
&\qquad+\left|\sum_{i=0}^{N+g-q-2}[(\phi+\theta)(z^{\phi}_i,z^{\phi}_{i+1})-\mu(\phi+\theta)]\right|\le 60C_1(2\delta_0).
\end{aligned}\tag{4.13}$$

It follows from the definition of $\{a_i\}_{i=0}^{\infty}$ and $\{z^{\phi}_i\}_{i=0}^{\infty}$ and relation (4.6), which holds for $\varepsilon=\delta_0$, that

$$\begin{aligned}
&\sum_{i=q}^{N+g-1} [(\phi+\theta)(y_i, y_{i+1}) - (\phi+\theta)(a_i, a_{i+1})] \\
&\quad \geq \sum_{i=q+1}^{N+g-1} [(\phi+\theta)(y_i, y_{i+1}) - (\phi+\theta)(a_i, a_{i+1})] - 2h - C_1(\delta_0) \\
&\quad \geq -2h - C_1(\delta_0) + \ell(\phi, N+g-q-1) - \sum_{i=0}^{N+g-q-2} (\phi+\theta)(z_i^\phi, z_{i+1}^\phi) \\
&\quad \geq -2h - 19C_1(2\delta_0).
\end{aligned} \tag{4.14}$$

Relations (4.11)–(4.14) imply that

$$\begin{aligned}
&\sum_{i=0}^{N+g-1} [u(y_i, y_{i+1}) - \mu(u)] - \sum_{i=0}^{q+s(g)} [u(a_i, a_{i+1}) - \mu(u)] \\
&\quad \geq -2h - 19C_1(2\delta_0) - 60C_1(2\delta_0) - C_3(2\delta_0) \qquad (g = 0, 1, \ldots).
\end{aligned}$$

This relation implies

$$\begin{aligned}
\liminf_{S\to\infty} \sum_{i=0}^{S-1} [u(a_i, a_{i+1}) - \mu(u)] &\leq \liminf_{S\to\infty} \sum_{i=0}^{S-1} [u(y_i, y_{i+1}) - \mu(u)] \\
&\quad + 2h + 80C_1(2\delta_0) + C_3(2\delta_0).
\end{aligned}$$

Since $\{a_i\}_{i=0}^\infty \in H_4 \subset H_3(x, u, Q, \delta_0)$ and $\{y_i\}_{i=0}^\infty$ is an arbitrary element of $H_3(x, u, Q, \delta_0)$

$$\begin{aligned}
\pi^u(x) &\leq \inf\left\{\liminf_{S\to\infty} \sum_{i=0}^{S-1} [u(z_i, z_{i+1}) - \mu(u)] : \{z_i\}_{i=0}^\infty \in H_4\right\} \\
&\leq \pi^u(x) + 2h + 80C_1(2\delta_0) + C_3(2\delta_0).
\end{aligned} \tag{4.15}$$

Denote by $H_5(x)$ the set of all programs $\{z_i\}_{i=0}^q (q \in \{0, \ldots, Q\})$ such that $z_0 = x$, $d(z_q, x_{m-1}^*) \leq \delta_0$. There exists a mapping $P : H_5(x) \to H_4$, where

$$\begin{aligned}
&P(\{z_i\}_{i=0}^q) = \{a_i\}_{i=0}^\infty, \quad a_i = z_i \quad (i = 0, \ldots, q), \quad a_i = z_{i-q-1}^\phi \quad (i \geq q+1) \\
&\text{for } q \in \{0, \ldots, Q\}, \ \{z_i\}_{i=0}^q \in H_5(x).
\end{aligned}$$

It is easy to see that $P(H_5(x)) = H_4$. Let $q \in \{0, \ldots, Q\}$, $\{z_i\}_{i=0}^q \in H_5(x)$, $\{a_i\}_{i=0}^\infty = P(\{z_i\}_{i=0}^q)$. Relation (4.8), which holds for $\varepsilon = \delta_0$, implies

$$\begin{aligned}
&\left| u(z_q, x_0^*) - \mu(u) + \sum_{i=0}^{q-1} [u(z_i, z_{i+1}) - \mu(u)] - \liminf_{N\to\infty} \sum_{i=0}^{N-1} [u(a_i, a_{i+1}) - \mu(u)] \right| \\
&\quad = \left| \liminf_{N\to\infty} \sum_{i=q+1}^{N-1} [u(a_i, a_{i+1}) - \mu(u)] \right| \\
&\quad \leq \left| \liminf_{N\to\infty} \sum_{i=0}^{N-1} [u(z_i^\phi, z_{i+1}^\phi) - \mu(u)] \right| \leq C_3(\delta_0) + 30C_1(2\delta_0).
\end{aligned}$$

This relation implies that

$$\left| \inf\left\{ \liminf \sum_{i=0}^{N-1} [u(z_i, z_{i+1}) - \mu(u)] : \{z_i\}_{i=0}^{\infty} \in H_4 \right\} \right.$$
$$\left. - \inf\left\{ \sum_{i=0}^{q-1} [u(z_i, z_{i+1}) - \mu(u)] + u(z_q, x_0^*) - \mu(u) : \{z_i\}_{i=0}^{q} \in H_5(x),\ 0 \le q \le Q \right\} \right|$$
$$\le C_3(\delta_0) + 30C_1(2\delta_0).$$

This relation and (4.15) imply that

(4.16)
$$\left| \pi^u(x) - \inf\left\{ \sum_{i=0}^{q-1} [u(z_i, z_{i+1}) - \mu(u)] + u(z_q, x_0^*) - \mu(u) : \{z_i\}_{i=0}^{q} \in H_5(x) \right\} \right|$$
$$\le 2h + 110C_1(2\delta_0) + 2C_3(2\delta_0).$$

(4.16) holds for every $x \in K$ and for every $u \in \mathbf{C}(K \times K)$ satisfying $\|u - v\| \le h$. Let $x \in K$, $w \in \mathbf{C}(K \times K)$, and $\|w - v\| \le h$. By Proposition 1 $|\mu(w) - \mu(v)| \le h$. It is easy to see that $\mu(v) = \mu$. It follows from the choice of δ_0 and h and relation (4.16), which holds for $u = w$ and for $u = v$, that

$$|\pi^w(x) - \pi^v(x)| \le 4h + 220C_1(2\delta_0) + 4C_3(2\delta_0)$$
$$+ \left| \inf\left\{ \sum_{i=0}^{q-1} [w(z_i, z_{i+1}) - \mu(w)] \right.\right.$$
$$\left. + w(z_q, x_0^*) - \mu(w) : \{z_i\}_{i=0}^{q} \in H_5(x),\ 0 \le q \le Q \right\}$$
$$- \inf\left\{ \sum_{i=0}^{q-1} [v(z_i, z_{i+1}) - \mu(v)] \right.$$
$$\left.\left. + v(z_q, x_0^*) - \mu(v) : \{z_i\}_{i=0}^{q} \in H_5(x),\ 0 \le q \le Q \right\} \right|$$
$$\le 4h + 220C_1(2\delta_0) + 4C_3(2\delta_0) + 2h(Q + 1) \le \Delta.$$

This completes the proof of the lemma.

LEMMA 11. $\mu(v) = \mu$, $\pi^v(x_g^*) = \pi(x_g^*) - \pi(x_0^*)(g = 0, \dots, m - 1)$.

Proof. The equality $\mu(v) = \mu$ follows from the definition of $\mu(v)$ (see (1.1)). Let $g \in \{0, \dots, m - 1\}$ and $y_i = x_{g+i}^*$ $(i = 0, 1, \dots)$. Then

$$\liminf \sum_{i=0}^{N-1} [v(y_i, y_{i+1}) - \mu] = \pi(x_g^*) - \pi(x_0^*), \tag{4.17}$$
$$\pi^v(x_g^*) \le \pi(x_g^*) - \pi(x_0^*),$$

(4.18)
$$\pi^v(x_g^*) = \inf\left\{ \liminf_{N\to\infty} \sum_{i=0}^{N} [v(y_i, y_{i+1}) - \mu] : \{y_i\}_{i=0}^{\infty} \text{ is a } (v)\text{-good program}, y_0 = x_g^* \right\}.$$

Let $\{y_i\}_{i=0}^{\infty}$ be a (v)-good program, $y_0 = x_g^*$, and $\varepsilon \in (0, D_0)$. Lemma 4 implies the existence of an integer $Q \geq 1$ such that

$$d_1((y_{Q+i}, y_{Q+i+1}),\ (x_i^*, x_{i+1}^*)) \leq \varepsilon \quad (i = 0, 1, \ldots). \tag{4.19}$$

Relation (4.19) implies that for every integer $j \geq 0$

$$\begin{aligned}\sum_{i=0}^{Q+j-1} [v(y_i, y_{i+1}) - \mu] &\geq \pi(x_g^*) - \pi(y_{Q+j}) \geq \pi(x_g^*) - \pi(x_j^*) - C_3(\varepsilon)\\ &\geq \pi(x_g^*) - \pi(x_0^*) - C_3(\varepsilon).\end{aligned}$$

This relation and (4.18) imply that $\pi^v(x_g^*) \geq \pi(x_g^*) - \pi(x_0^*)$. The lemma then follows from this inequality and (4.17).

For every number $\varepsilon > 0$ we set

$$C_4(\varepsilon) = \sup\{|\pi^v(z_1) - \pi^v(z_2)| : z_1, z_2 \in K,\ d(z_1, z_2) \leq \varepsilon\}. \tag{4.20}$$

LEMMA 12. *Let $\Delta \in (0, D_0)$. Then there exists a number $R(\Delta) \in (0, \Delta)$ such that for every $u \in \mathbf{C}(K \times K)$ satisfying $\|u - v\| \leq R(\Delta)$, for every integer $g \in \{0, \ldots, m-1\}$, and for every program $\{y_i\}_{i=0}^{\infty}$ satisfying $d(y_0, x_g^*) \leq R(\Delta)$, $\theta^u(y_i, y_{i+1}) = 0\,(i = 0, 1, \ldots)$, the following relation holds: $d_1((y_i, y_{i+1}),\ (x_{g+i}^*, x_{g+i+1}^*)) \leq \Delta\,(i = 0, 1, \ldots)$.*

Proof. We choose $r_0 \in (0, \Delta)$ such that Lemma 2 holds for $R = r_0$, $\varepsilon = \Delta$. We choose $\delta_0 \in (0, r_0)$ satisfying

$$50C_1(2\delta_0) + 2C_4(\delta_0) + 2C_3(\delta_0) \leq 4^{-1}r_0$$

and choose $R(\Delta) \in (0, \delta_0)$ such that Lemma 4 holds for $\varepsilon_0 = \delta_0$, $r = R(\Delta)$, Lemma 9 holds for $\varepsilon = \delta_0$, $r = R(\Delta)$, and the following relation is fulfilled:

$$\|\pi^v - \pi^u\| \leq 4^{-1}r_0 \qquad (u \in \mathbf{C}(K \times K),\ \|u - v\| \leq R(\Delta))$$

(see Lemma 10). Let $u \in \mathbf{C}(K \times K)$, $\|u - v\| \leq R(\Delta)$, $g \in \{0, \ldots, m-1\}$, $\{y_i\}_{i=0}^{\infty}$ be a program satisfying $d(y_0, x_g^*) \leq R(\Delta)$, and $\theta^u(y_i, y_{i+1}) = 0\,(i = 0, 1, \ldots)$. By Lemma 4, which holds for $\varepsilon_0 = \delta_0$, $r = R(\Delta)$, there is an integer $Q \geq 1$ such that

$$d_1((y_{Q+i}, y_{Q+i+1}),\ (x_i^*, x_{i+1}^*)) \leq \delta_0 \qquad (i = 0, 1, \ldots). \tag{4.21}$$

Set $\phi = u - v$. For every integer $j \geq 0$

$$\sum_{i=0}^{Q+j-1} [u(y_i, y_{i+1}) - \mu(u)] = \pi^u(y_0) - \pi^u(y_{Q+j}).$$

Lemma 11 and relation (4.21) imply

$$\begin{aligned}&\left|\sum_{i=0}^{Q+j-1} [u(y_i, y_{i+1}) - \mu(u)] - (\pi(x_g^*) - \pi(x_j^*))\right|\\ &= \left|\sum_{i=0}^{Q+j-1} [u(y_i, y_{i+1}) - \mu(u)] - (\pi^v(x_g^*) - \pi^v(x_j^*))\right|\\ &\leq |\pi^u(y_0) - \pi^u(y_{Q+j}) - (\pi^v(x_g^*) - \pi^v(x_j^*))|\\ &\leq 2\|\pi^v - \pi^u\| + |\pi^v(y_0) - \pi^v(y_{Q+j}) - (\pi^v(x_g^*) - \pi^v(x_j^*))|\\ &\leq 2\|\pi^v - \pi^u\| + 2C_4(\delta_0) \qquad (j = 0, 1, \ldots).\end{aligned} \tag{4.22}$$

It follows from Lemma 9 and the choice of δ_0, $R(\Delta)$ that

$$\mu(\phi+v)=\mu+\mu(\phi+\theta), \tag{4.23}$$

$$\ell(\phi,N)\le N\mu(\phi+\theta)\le \ell(\phi,N)+50C_1(2\delta_0) \qquad (N=1,2,\dots). \tag{4.24}$$

For every integer $j\ge 0$ using (4.21) – (4.23) we have

$$\sum_{i=0}^{Q+j-1}[u(y_i,y_{i+1})-\mu(u)]$$
$$=\pi(y_0)-\pi(y_{Q+j})+\sum_{i=0}^{Q+j-1}[(\theta+\phi)(y_i,y_{i+1})-\mu(\phi+\theta)];$$

$$\left|\sum_{i=0}^{Q+j-1}[(\theta+\phi)(y_i,y_{i+1})-\mu(\phi+\theta)]\right|$$
$$=\left|\sum_{i=0}^{Q+j-1}[u(y_i,y_{i+1})-\mu(u)]-(\pi(y_0)-\pi(y_{Q+j}))\right|$$
$$\le|\pi(x_g^*)-\pi(x_j^*)-(\pi(y_0)-\pi(y_{Q+j}))|+2\|\pi^v-\pi^u\|+2C_4(\delta_0)$$
$$\le 2\|\pi^v-\pi^u\|+2C_4(\delta_0)+2C_3(\delta_0) \qquad (j=0,1,\dots),$$

$$\left|\sum_{i=0}^{N-1}[(\theta+\phi)(y_i,y_{i+1})-\mu(\phi+\theta)]\right|$$
$$\le 2\|\pi^v-\pi^u\|+2C_4(\delta_0)+2C_3(\delta_0) \qquad (N=Q,Q+1,\dots).$$

The last relation, relation (4.24), and our choice of $R(\Delta)$ and δ_0 imply that

$$\sum_{i=0}^{N-1}(\theta+\phi)(y_i,y_{i+1})\le 2\|\pi^v-\pi^u\|+2C_4(\delta_0)+2C_3(\delta_0)+\ell(\phi,N)$$
$$+50C_1(2\delta_0)\le \ell(\phi,N)+r_0 \qquad (N=Q,Q+1,\dots).$$

The validity of the lemma follows from Lemma 2, which holds for $\varepsilon=\Delta$, $R=r_0$.

LEMMA 13. *Let $\Delta\in(0,D_0)$. Then there exist $R\in(0,\Delta)$, an integer $Q\ge 1$ such that for every $u\in \mathbf{C}(K\times K)$ satisfying $\|u-v\|\le R$, for every program $\{y_i\}_{i=0}^{\infty}$ satisfying $\theta^u(y_i,y_{i+1})=0\,(i=0,1,\dots)$, there is $q\in\{1,\dots,Q\}$ such that $d_1((y_{q+i},y_{q+i+1}),(x_i^*,x_{i+1}^*))\le\Delta\,(i=0,1,\dots)$.*

Proof. There exists $R(\Delta)\in(0,\Delta)$ such that Lemma 12 holds for Δ, $R(\Delta)$. We choose $R\in(0,R(\Delta))$ satisfying $R\le(48m)^{-1}C_2(R(\Delta))$ and an integer $Q\ge 400\|\theta\|C_2(R(\Delta))^{-1}m$. Let $u\in\mathbf{C}(K\times K)$, $\|u-v\|\le R$ and $\{y_i\}_{i=0}^{\infty}$ be a program satisfying $\theta^u(y_i,y_{i+1})=0\,(i=0,1,\dots)$. By Lemma 1, which holds for $\delta=R(\Delta)$, there is $q\in\{1,\dots,Q\}$ satisfying $d(y_q,x_0^*)\le R(\Delta)$. The lemma now follows from Lemma 12 and the definition of $R(\Delta)$.

LEMMA 14. *Let $m=1$, $\delta>0$. Then there exist $R_0\in(0,\delta)$ and an integer $Q\ge 1$ such that for every $u\in\mathbf{C}(K\times K)$ satisfying $\|u-v\|\le R_0$, for every integer $N\ge 2Q$, and for every program $\{y_i\}_{i=0}^{N}\in A(u,N,R_0)$ the relation $d(y_i,x_0^*)\le\delta$ holds for $i=Q,\dots,N-Q$. Moreover, if $d(y_0,x_0^*)\le R_0$, then $d(y_i,x_0^*)\le\delta$ for $i=0,\dots,N-Q$.*

Proof. We choose $R\in(0,\min\{\delta,\|\theta\|\})$ such that Lemma 2 holds for $\varepsilon=\delta$ and R. We choose $\delta_0\in(0,8^{-1}R)$ satisfying $64C_1(4\delta_0)\le R$ and choose $R_0\in(0,\delta_0)$ such that $R_0\le(48m)^{-1}C_2(\delta_0)$ and Lemma 3 holds for $\varepsilon=\delta_0$, $r=R_0$. We set $Q=400\|\theta\|C_2(\delta_0)^{-1}m$.

Let $u \in \mathbf{C}(K \times K)$, $\|u - v\| \leq R_0$, integer $N \geq 2Q$, and $\{y_i\}_{i=0}^N \in A(u, N, R_0)$. Lemma 1, which holds for $\delta = \delta_0$, $\phi = u - v$, and $\{y_i\}_{i=0}^N$ by the choice of R_0, implies the existence of integers p and q such that $0 < p < q < N$, $Q \geq p$, $N - q \leq Q$, $d(y_{p-1}, x_0^*) \leq \delta_0$, and $d(y_{q+1}, x_0^*) \leq \delta_0$. If $d(x_0^*, y_0) \leq R_0$, then we assume that $p = 1$. By Lemma 3, which holds for $\varepsilon = \delta_0$, $r = R_0$, there exists a program $\{a_i\}_{i=0}^{q-p}$ such that $a_0 = a_{q-p} = x_0^*$,

$$\sum_{i=0}^{q-p-1} (\phi + \theta)(a_i, a_{i+1}) \leq \ell(\phi, q - p) + 6C_1(\delta_0). \tag{4.25}$$

Consider a program $\{z_i\}_{i=0}^N$, where $z_i = y_i \, (i \in \{0, \ldots, p-1\} \cup \{q+1, \ldots, N\})$, $z_i = a_{i-p} \, (i = p, \ldots, q)$. The relation $\{y_i\}_{i=0}^N \in A(u, N, R_0)$ and our choice of p and q imply that

$$\begin{aligned} R_0 &\geq \sum_{i=0}^{N-1} u(y_i, y_{i+1}) - \sum_{i=0}^{N-1} u(z_i, z_{i+1}) \\ &= \sum_{i=p-1}^{q} [(\theta + \phi)(y_i, y_{i+1}) - (\theta + \phi)(z_i, z_{i+1})] \\ &\geq \sum_{i=p}^{q-1} (\theta + \phi)(y_i, y_{i+1}) - \sum_{i=0}^{q-p-1} (\theta + \phi)(a_i, a_{i+1}) - 4R_0 - 2C_1(\delta_0). \end{aligned}$$

It follows from this relation, relation (4.25), and the choice of R_0 and δ_0 that

$$\sum_{i=p}^{q-1} (\theta + \phi)(y_i, y_{i+1}) \leq \ell(\phi, q - p) + 5R_0 + 8C_1(\delta_0) \leq R + \ell(\phi, q - p).$$

By Lemma 2, which holds for $\varepsilon = \delta$ and R, we have $d(y_{p+i}, x_0^*) \leq \delta_0 \, (i = 0, \ldots, q - p)$. The lemma is proved.

5. Proof of theorems. We consider the set E of all functions $v \in \mathbf{C}(K \times K)$ for which there exist an integer $m(v) \geq 1$, a sequence $\{x_i^*(v)\}_{i=0}^{m(v)} \subset K$, continuous functions $\pi_v : K \to \mathbf{R}^1$, $\theta_v : K \times K \to \mathbf{R}^1$, and a number μ_v such that the following conditions hold:

1. For $i, j \in \{0, \ldots, m(v)\}$ satisfying $i < j$ the equality $x_i^*(v) = x_j^*(v)$ holds if and only if $i = 0$, $j = m(v)$;
2. $v(x, y) = \mu_v + \pi_v(x) - \pi_v(y) + \theta_v(x, y)(x, y \in K)$;
3. $\|\theta_v\| > 0$, the function θ_v is nonnegative, $\theta_v(x, y) = 0$ if and only if

$$(x, y) \in \{(x_i^*(v), x_{i+1}^*(v)) : i = 0, \ldots, m(v) - 1\}.$$

At the beginning of §2 we already noted that E is dense everywhere in $\mathbf{C}(K \times K)$. It is easy to see that $\mu(v) = \mu_v (v \in E)$, and for every $v \in E$ we can apply Lemmas 1–14. For every $v \in E$ and every $i \in \{0, \pm 1, \ldots\} \backslash \{0, \ldots, m(v)\}$ define $x_i^*(v) \in K$ such that $x_{i+m(v)}^*(v) = x_i^*(v) \, (i = 0, \pm 1, \ldots)$. Let $v \in E$. We set

$$D(v) = 8^{-1} \inf\{d(x_i^*(v), x_j^*(v)) : i, j \in \{0, \ldots, m(v) - 1\}, \, i \neq j\}. \tag{5.1}$$

If $m(v) = 1$, then $D(v) = +\infty$.

Let $p \in \{1, 2, \ldots\}$. We define

$$\delta(v, p) = \inf\{2^{-1} D(v), p^{-1}\}. \tag{5.2}$$

It is easy to see that there exist numbers $\Gamma(v,p) \in (0,\delta(v,p))$, $d(v,p) \in (0,\Gamma(v,p))$, positive numbers $Q_1(v,p)$, $Q_2(v,p)$, and an integer $Q_3(v,p) \geq 1$ such that the following conditions hold:

a) If $m(v) = 1$, then Lemma 14 holds for $\delta = \delta(v,p)$, $R_0 = 4\Gamma(v,p)$, $Q = Q_3(v,p)$, $x_0^* = x_0^*(v)$;

b) Lemma 12 holds for $\Delta = \delta(v,p)$, $R(\Delta) = 4\Gamma(v,p)$, $m = m(v)$, $x_i^* = x_i^*(v)$ $(i = 0, \pm 1, \ldots)$;

c) Lemma 13 holds for $\Delta = \delta(v,p)$, $R = 4\Gamma(v,p)$, $Q = Q_3(v,p)$, $x_i^* = x_i^*(v)$ $(i = 0, \pm 1, \ldots)$;

d) Lemma 4 holds for $\varepsilon_0 = \Gamma(v,p)$, $r = d(v,p)$, $x_i^* = x_i^*(v)$ $(i = 0, \pm 1, \ldots)$;

e) Lemma 6 holds for $\Delta = \Gamma(v,p)$, $r(\Delta) = d(v,p)$, $Q_1(\Delta) = Q_1(v,p)$, $Q_2(\Delta) = Q_2(v,p)$, $m = m(v)$, $x_i^* = x_i^*(v)$ $(i = 0, \pm 1, \ldots)$;

f) Lemma 10 holds for $\Delta = \Gamma(vp)$, $h = d(vp)$.

Now, we define $F = \cap_{p=1}^{\infty} \cup_{v \in E} B(v, d(v,p))$. It follows from the construction of F that Theorems 3 and 4 of [13] are valid for F. For the set F we will prove Theorems 1 – 4 of this paper.

It is easy to see that Theorem 1 follows from Lemma 10 and the definition of F (see condition f).

Proof of Theorem 2. Let $u \in F$, $\delta > 0$. Choose an integer $p \geq 1$ satisfying $4p^{-1} < \delta$. There exists $v \in E$ for which $u \in B(v, d(v,p))$. We define

$$W(u) = B(v, d(v,p)), \quad Q_1 = Q_1(v,p), \quad Q_2 = Q_2(v,p). \tag{5.3}$$

By condition d) we have

$$\operatorname{dist}(H(u), \{(x_i^*(v), x_{i+1}^*(v)) : i = 0, \ldots, m(v) - 1\}) \leq \Gamma(v,p). \tag{5.4}$$

By condition e) Lemma 6 holds for $\Delta = \Gamma(v,p)$, $r(\Delta) = d(v,p)$, $Q_i(\Delta) = Q_i$ $(i = 1,2), m = m(v), x_i^* = x_i^*(v) (i = 0, \pm 1, \ldots)$.

Let $w \in W(u)$, N be an integer, $N \geq 1$, $M > 0$, and $\{y_i\}_{i=0}^N \in A(w, N, M)$. Then, by Lemma 6

(5.5)
$$\operatorname{Card}\left\{ i \in \{0, \ldots, N-1\} : (y_i, y_{i+1}) \notin \bigcup_{j=0}^{m(v)-1} \bar{B}((x_j^*(v), x_{j+1}^*(v)), \Gamma(v,p)) \right\} \leq Q_1 + MQ_2.$$

This relation and relation (5.4) imply

$$\operatorname{Card}\{i \in \{0, \ldots, N-1\} : d_1((y_i, y_{i+1}), H(u)) > 4\Gamma(v,p)\} \leq Q_1 + MQ_2.$$

This completes the proof of the theorem.

We have the following result.

PROPOSITION 2. *Let $u \in F$, $(x_0, x_1) \in H(u)$. Then there exists a sequence $\{x_i\}_{i=-\infty}^{\infty} \subset K$ such that $(x_i, x_{i+1}) \in H(u)$, $\theta^u(x_i, x_{i+1}) = 0$ $(i = 0, \pm 1, \ldots)$.*

Proof. Let $\{z_i\}_{i=0}^{\infty}$ be a (u)-good program. For an integer $p \geq 0$ denote by Z^p a sequence $\{z_i^p\}_{i=-p}^{\infty}$, where $z_i^p = z_{i+p}$ $(i = -p, -p+1, \ldots)$. It is easy to see that there exists a subsequence $\{Z^{p_k}\}_{k=1}^{\infty}$ such that

$$z_i^{p_k} \xrightarrow[k \to \infty]{} x_i \in K \qquad (i = 0, \pm 1, \ldots).$$

Clearly, $\{x_i\}_{i=-\infty}^{\infty}$ is the required sequence.

Proof of Theorem 3. Let $u \in F$, $\varepsilon > 0$. We choose an integer $p \geq 1$ satisfying $4p^{-1} < \varepsilon$. There exists $v \in E$ such that $u \in B(v, d(v,p))$. By condition d) and Lemma 4 we have

$$\operatorname{dist}(H_0(u), \{x_i^*(v) : i = 0, \ldots, m(v) - 1\}) \leq \Gamma(v,p), \tag{5.6}$$

$$\operatorname{dist}(H(u), \{(x_i^*(v), x_{i+1}^*(v)) : i = 0, \ldots, m(v) - 1\}) \leq \Gamma(v,p). \tag{5.7}$$

We define

$$W(u) = B(v, d(v,p)), \qquad \delta = \Gamma(v,p). \tag{5.8}$$

Let $w \in W(u)$ and $\{x_i\}_{i=0}^{\infty}$ be a program such that $\theta^w(x_i, x_{i+1}) = 0\ (i = 0, 1, \ldots)$, $d(x_0, H_0(u)) \leq \delta$. Then, by (5.6), (5.8)

$$d(x_0, \{x_i^*(v) : i = 0, \ldots, m(v) - 1\}) \leq 2\Gamma(v,p),$$

and by condition b) and Lemma 12

$$d_1((x_i, x_{i+1}), \{(x_i^*(v), x_{i+1}^*(v)) : i = 0, \ldots, m(v) - 1\}) \leq \delta(v,p)$$

for $i = 0, 1, \ldots$. By (5.7) and by our choice of p

$$d_1((x_i, x_{i+1}), H(u)) \leq 2\delta(v,p) \leq 2p^{-1} < \varepsilon \qquad (i = 0, 1, \ldots).$$

Assertion 1 is proved.

Let us prove assertion 2. Let $u \in F$, $\varepsilon > 0$. We choose an integer $p \geq 1$ satisfying $4 < p\varepsilon$. There exists $v \in E$ such that $u \in B(v, d(v,p))$. By condition d) and Lemma 4 we have

$$\operatorname{dist}(H(u), \{(x_i^*(v), x_{i+1}^*(v)) : i = 0, \ldots, m(v) - 1\}) \leq \Gamma(v,p). \tag{5.9}$$

We define

$$W(u) = B(v, d(v,p)), \qquad N = Q_3(v,p). \tag{5.10}$$

Let $w \in W(u)$ and $\{y_i\}_{i=0}^{\infty}$ be a program satisfying $\theta^w(y_i, y_{i+1}) = 0\,(i = 0, 1, \ldots)$. Then, by condition c), Lemma 13, and relation (5.9)

$$d_1((y_{i+N}, y_{i+N+1}), \{(x_i^*(v), x_{i+1}^*(v)) : i = 0, \ldots, m(v) - 1\}) \leq \delta(v,p)\,(i = 0, 1, \ldots),$$
$$d_1((y_i, y_{i+1}), H(u)) \leq 2\delta(v,p) < \varepsilon \qquad (i \geq N).$$

Hence, Assertion 2 is proved.

Proof of Theorem 4. Let $u \in F$, $\varepsilon > 0$. We choose an integer $p \geq 1$ satisfying $4p^{-1} < \varepsilon$ There exists $v \in E$ such that $u \in B(v, d(v,p))$. By condition d) and Lemma 4 we have

$$\operatorname{dist}(H(u), \{(x_i^*(v), x_{i+1}^*(v)) : i = 0, \ldots, m(v) - 1\}) \leq \Gamma(v,p). \tag{5.11}$$

By corollary 2 of Theorem 3 there exists a neighborhood $W(u)$ of u in $\mathbf{C}(K \times K)$ such that $W(u) \subset B(v, d(v,p))$, and for every $w \in W(u)$, for every program $\{x_i\}_{i=-\infty}^{\infty}$ satisfying $\theta^w(x_i, x_{i+1}) = 0\,(i = 0, \pm 1, \ldots)$ the relation $d_1((x_i, x_{i+1}), H(u)) \leq \Gamma(v,p)\,(i = 0, \pm 1, \ldots)$ holds.

Let $w \in W(u)$ and $\{y_i\}_{i=-\infty}^{\infty}$ be a program satisfying $\theta^w(y_i, y_{i+1}) = 0$ $(i = 0, \pm 1, \ldots)$. By (5.11)

$$(y_i, y_{i+1}) \in \bigcup_{j=0}^{m(v)-1} \bar{B}((x_j^*(v), x_{j+1}^*(v)), 2\Gamma(v,p)) \qquad (i = 0, \pm 1, \ldots),$$

and there is an integer k such that

$$d_1((y_i, y_{i+1}), (x_{i+k}^*(v), x_{i+k+1}^*(v))) \leq 2\Gamma(v,p) \qquad (i = 0, \pm 1, \ldots).$$

Now it is easy to see that for all integers i, j the relation $d(y_i, y_{i+jm(v)}) \leq 4\Gamma(v,p) \leq \varepsilon$ holds. Theorem 4 is proved.

We will prove the following result.

THEOREM 5. 1. *Let $u \in F$, $\{x_i\}_{i=0}^{\infty}$ be a (u)-good program. Then there exists a program $\{y_i\}_{i=-\infty}^{\infty}$ such that*

$$\theta^u(y_i, y_{i+1}) = 0 \quad (i = 0, \pm 1, \ldots), \quad \lim_{i\to\infty} d(x_i, y_i) = 0.$$

2. *Let $u \in F$, $\{x_i\}_{i=-\infty}^{\infty}$ be a (u)- minimal energy configuration. Then there exist programs $\{y_i\}_{i=-\infty}^{\infty}$, $\{z_i\}_{i=-\infty}^{\infty}$ such that*

$$\theta^u(y_i, y_{i+1}) = 0, \quad \theta^u(z_i, z_{i+1}) = 0 \quad (i = 0, \pm 1, \ldots),$$
$$\lim_{i\to+\infty} d(x_i, y_i) = 0, \quad \lim_{i\to-\infty} d(x_i, z_i) = 0.$$

Proof. First we will prove assertion 1. Let $u \in F$, $\{x_i\}_{i=0}^{\infty}$ be a (u)-good program. By [13, Thm. 3] the program $\boldsymbol{x}$ is asymptotic almost periodic. Therefore for every integer $p \geq 1$ there exist integers $k(p) \geq 2$, $m(p) \geq 2$ such that

$$d(x_i, x_{i+m(p)j}) \leq p^{-1} \quad (i \geq k(p),\ j \geq 1). \tag{5.12}$$

For every integer $p \geq 0$ let us denote by X^p a sequence $\{x_i^p\}_{i=-p}^{\infty}$, where $x_i^p = x_{i+p}$ $(i = -p, -p+1, \ldots)$. We define

$$M(p) = \prod_{i=1}^{p} m(i) \qquad (p = 1, 2, \ldots).$$

Let Q be an integer, $Q \geq 1$, and let us consider a program $X^{M(Q)} = \{x_i^{M(Q)}\}_{i=-M(Q)}^{\infty}$. Let $p \in \{1, \ldots, Q\}$ and i be an integer satisfying $i \geq k(p)$. By (5.12) and the definition of $M(Q)$ we have $d(x_i, x_i^{M(Q)}) = d(x_i, x_{i+M(Q)}) \leq p^{-1}$. We have proved that for every integer $Q \geq 1$ the following relation holds:

$$d(x_i, x_i^{M(Q)}) \leq p^{-1} \qquad (i \text{ is an integer}, i \geq k(p)),\ (p \in \{1, \ldots, Q\}). \tag{5.13}$$

There exists a subsequence $\{X^{M(Q_j)}\}_{j=1}^{\infty}$ such that

$$x_i^{M(Q_j)} \xrightarrow[j\to\infty]{} y_i \in K \quad \text{for every integer } i.$$

Assertion 1 now follows from (5.13).

Now we will prove assertion 2. Let $u \in F$ and $\{x_i\}_{i=-\infty}^{\infty}$ be a (u)-minimal energy configuration. The existence of configuration $\{y_i\}_{i=-\infty}^{\infty}$ follows from Assertion 1, which holds for the (u)-good program $\{x_i\}_{i=0}^{\infty}$. Let us prove the existence of the program $\{z_i\}_{i=-\infty}^{\infty}$. Fix an integer $p \geq 1$. There exists $v \in E$ for which $u \in B(v, d(v,p))$. By condition d) and Lemma 4 we have

$$\text{dist}(H(u), \{(x_i^*(v), x_{i+1}^*(v)) : i = 0, \dots, m(v)-1\}) \leq \Gamma(v,p). \tag{5.14}$$

By Theorem 2 there exists an integer $Q(p) \geq 1$ such that for all integers N_1 and N_2 satisfying $N_1 < N_2$ the following relation holds:

$$\text{Card}\{i \in \{N_1, \dots, N_2 - 1\} : d_1((x_i, x_{i+1}), H(u)) > 2^{-1}(\delta(v,p) - \Gamma(v,p))\} \leq Q(p).$$

By this relation and by (5.14) there exists an integer $k(p) \geq 1$ such that

$$\begin{aligned} &d_1((x_i, x_{i+1}), H(u)) \leq 2^{-1}(\delta(v,p) - \Gamma(v,p)) \qquad (i \leq -k(p)),\\ &d_1((x_i, x_{i+1}), \{(x_i^*(v), x_{i+1}^*(v)) : i = 0, \dots, m(v)-1\})\\ &\quad \leq 2^{-1}(\delta(v,p) + \Gamma(v,p)) \qquad (i \leq -k(p)). \end{aligned} \tag{5.15}$$

We have $\delta(v,p) \leq 16^{-1} \inf\{d(x_i^*(v), (x_j^*(v)) : i, j \in \{0, \dots, m(v)-1, i < j\}$; hence, there exists an integer G such that

$$\begin{aligned} &d_1((x_{-k(p)+i}, x_{-k(p)+i+1}), (x_{G+i}^*(v), x_{G+i+1}^*(v))) \leq 2^{-1}(\delta(v,p) + \Gamma(v,p))\\ &(i = 0, -1, -2, \dots). \end{aligned} \tag{5.16}$$

Set $N(p) = m(v)$. Relation (5.16) implies

$$d(x_i, x_{i+jN(p)}) \leq 2p^{-1} \qquad (i \leq -k(p), j \leq -1). \tag{5.17}$$

For an integer $p \geq 1$ we define $M(p) = 2^p \prod_{i=1}^{p} N(i)$ and denote by X^p a sequence $\{x_i^p\}_{i=-\infty}^{\infty}$, where $x_i^p = x_{i-p} (i = 0, \pm 1, \dots)$.

Let $Q \geq 1$ be an integer, $p \in \{1, \dots Q\}$, and i be an integer satisfying $i \leq -k(p)$. It follows from (5.17) and the definition of $M(Q)$ that

$$d(x_i, x_i^{M(Q)}) = d(x_i, x_{i-M(Q)}) \leq 2p^{-1} \qquad (i \leq -k(p)).$$

We have proved that for every integer $Q \geq 1$ the following relation holds:

$$d(x_i, x_i^{M(Q)}) \leq 2p^{-1} \qquad (i \leq -k(p)), (p \in \{1, \dots Q\}). \tag{5.18}$$

There exists a subsequence $\{X^{M(Q_j)}\}_{j=1}^{\infty}$ such that

$$x_i^{M(Q_j)} \xrightarrow[j \to \infty]{} z_i \in K \text{ for every integer } i.$$

Assertion 2 then follows from (5.15) and (5.18). Theorem 5 is proved.

We define $\tilde{\mathbf{C}}(K \times K) = \{v \in \mathbf{C}(K \times K) : \mu(v) = \max\{v(x,x) : x \in K\}\}$. It is easy to see that $\tilde{\mathbf{C}}(K \times K)$ is a closed subspace of $\mathbf{C}(K \times K)$. The space $\tilde{\mathbf{C}}(K \times K)$ has also the topology of the uniform convergence. We reenforce the previous theorems for $u \in \tilde{\mathbf{C}}(K \times K)$ and prove the existence of a set $F_0 \subset F \cap \tilde{\mathbf{C}}(K \times K)$ which is a countable intersection of subsets of $\tilde{\mathbf{C}}(K \times K)$ open everywhere dense in $\tilde{\mathbf{C}}(K \times K)$ and for which Theorem 6 holds. This theorem shows that for every $u \in F_0$ and every $x \in K$ there is a (u)-overtaking optimal program $\{x_i\}_{i=0}^{\infty}$ satisfying $x_0 = x$ and establishes an analogue of the strong turnpike theorem for $u \in F_0$.

THEOREM 6. 1. Card $(H_0(u)) = 1(u \in F_0)$.

2. *Let $u \in F_0$ and δ be a positive number. Then there exists a neighborhood $W(u)$ of u in $\mathbf{C}(K \times K)$ such that for every $w \in W(u)$, for every (w)-good program $\boldsymbol{x}$ the following relation holds*: dist $(\Omega(\mathbf{x}), (H_0(u) \times H_0(u))) \leq \delta$.

3. *Let $u \in F_0$ and $\{x_i\}_{i=0}^{\infty}$ be a program for which $\theta^u(x_i, x_{i+1}) = 0\,(i = 0, 1, \dots)$. Then $\{x_i\}_{i=0}^{\infty}$ is a (u)-overtaking optimal program, and moreover, if $\{y_i\}_{i=0}^{\infty}$ is a program such that $y_0 = x_0$,*

$$\liminf \sum_{i=0}^{N-1} [u(y_i, y_{i+1}) - u(x_i, x_{i+1})] = 0$$

then $\theta^u(y_i, y_{i+1}) = 0(i = 0, 1, \dots)$.

4. *Let $u \in F_0$, $\varepsilon > 0$. Then there exist a neighborhood $W(u)$ of u in $\mathbf{C}(K \times K)$, an integer $Q \geq 1$, and $\varepsilon_0 \in (0, \varepsilon)$ such that for every $w \in W(u)$, every integer $N \geq 2Q$, and every program $\{y_i\}_{i=0}^{N} \in A(w, N, \varepsilon_0)$ the following relation holds*: $d(y_i, H_0(u)) \leq \varepsilon(i = Q, \dots, N-Q)$ *and if $d(y_0, H_0(u)) \leq \varepsilon_0$, then $d(y_i, H_0(u)) \leq \varepsilon(i \in \{0, \dots, N-Q\})$.*

Proof. We define

$$E_0 = \{v \in E : m(v) = 1\}, \qquad F_0 = \bigcap_{p=1}^{\infty} \bigcup_{v \in E_0} B(v, d(v, p)).$$

It is easy to see that E_0 is dense everywhere in $\tilde{\mathbf{C}}(K \times K)$, $F_0 \subset F \cap \tilde{\mathbf{C}}(K \times K)$. Condition d) and Lemma 4 imply the validity of assertion 1. Assertion 2 follows from Assertion 3 of [13, Thm. 3] and the relation Card $\{H_0(u)\} = 1(u \in F_0)$. This relation also implies assertion 3.

Let us prove assertion 4. Let $u \in F_0$, $\varepsilon > 0$. Choose an integer $p \geq 1$ such that $4p^{-1} < \varepsilon$. There is $v \in E_0$ for which $u \in B(v, d(v, p))$. We define $W(u) = B(v, d(v, p))$, $Q = Q_3(v, p)$, $\varepsilon_0 = \Gamma(v, p)$. By condition d) and Lemma 4

$$d(H_0(u), x_0^*(v)) \leq \Gamma(v, p). \tag{5.19}$$

Let $w \in W(u)$, $N \geq 2Q$, and $\{y_i\}_{i=0}^{N} \in A(w, N, \varepsilon_0)$. Then, by condition a) and by Lemma 14 the following conditions hold:

$$d(y_i, x_0^*(v)) \leq \delta(v, p) \qquad (i \in \{Q, \dots, N-Q\});$$

if $d(y_0, x_0^*(v)) \leq 4\Gamma(v, p)$, then $d(y_i, x_0^*(v)) \leq \delta(v, p)(i \in \{0, \dots, N-Q\})$. By these conditions and (5.19) we have

$$d(y_i, H_0(u)) \leq 2\delta(v, p) \leq 2p^{-1} \leq \varepsilon \qquad (i \in \{Q, \dots, N-Q\}),$$

and if $d(y_0, H_0(u)) \leq \varepsilon_0 = \Gamma(v, p)$, then $d(y_0, x_0^*(v)) \leq 2\Gamma(v, p)$,

$$d(y_i, x_0^*(v)) \leq \delta(v, p) \qquad (i \in \{0, \dots, N-Q\}),$$
$$d(y_i, H_0(u)) \leq 2\delta(v, p) \leq 2p^{-1} \leq \varepsilon \qquad (i = 0, \dots, N-Q).$$

Theorem 6 is proved.

Acknowledgments. The author thanks Dr. Arie Leizarowitz for fruitful discussions.

REFERENCES

[1] A. Leizarowitz, *Infinite horizon autonomous systems with unbounded cost*, Appl. Math. Optim., 13 (1985), pp. 19–43.

[2] ———, *Optimal trajectories of infinite-horizon deterministic control systems*, Appl. Math. Optim., 19 (1989), pp. 11–32.

[3] A. Leizarowitz and V. J. Mizel, *One-dimensional infinite-horizon variational problems arising in continuum mechanics*, Arch. Rational Mech. Anal., 106 (1989), pp. 161–194.

[4] S. Aubry and P. Y. Le Daeron, *The discrete Frenkel–Kontorova model and its extensions* I, Phys. D, 8 (1983), pp. 381–442. (In English).

[5] A. J. Zaslavski, *Ground states in Frenkel–Kontorova models*, Izv. Akad. Nauk SSSR Ser. Mat., 50 (1986), pp. 969–999.

[6] D. Gale, *On optimal development in multisector economy*, Rev. Econom. Stud., 34 (1967), pp. 1–19.

[7] C. C. von Weizsäcker, *Existence of optimal programs of accumulation for an infinite horizon*, Rev. Econom. Stud., 32 (1965), pp. 85–104.

[8] W. A. Brock and A. Haurie, *On existence of overtaking optimal trajectories over an infinite time horizon*, Math. Oper. Res., 1 (1976), pp. 337–346.

[9] D. Carlson, A. Haurie, and A. Leizarowitz, *Infinite Horizon Optimal Control*, Springer-Verlag, Berlin, 1991.

[10] Z. Artstein and A. Leizarowitz, *Tracking periodic signals with the overtaking criterion*, IEEE Trans. Automat. Control, AC-30, (1985), pp. 1123–1126.

[11] V. L. Makarov and A. M. Rubinov, *Mathematical Theory of Economic Dynamics and Equilibria*, Nauka, Moscow, 1973. (English translation: Springer-Verlag, New York, 1977.)

[12] A. M. Rubinov, *Superlinear Multivalued Mappings and Their Applications to Economic Mathematical Problems*, Leningrad, 1980.

[13] A. J. Zaslavski, *Optimal programs on infinite horizon* 1, SIAM J. Control Optim. 33 (1995), pp. 1643–1660.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 33, No. 6, pp. 1687-1715, November 1995

© 1995 Society for Industrial and Applied Mathematics
004

COCOLOG: A CONDITIONAL OBSERVER AND CONTROLLER LOGIC FOR FINITE MACHINES*

PETER E. CAINES[†] AND SUNING WANG[‡]

Abstract. The problem of observation and control for partially observed input–state–output machines is formulated in terms of a tree of first-order logical theories. A set of first-order languages for the description of the controlled evolution and state estimation of any given machine $\mathcal{M}$ is specified; further, extralogical conditional control rules are formulated so that closed loop control actions occur when extralogically specified past observation dependent conditions are fulfilled. In particular, conditional control rules may include commands that steer the system state from a current partially observed state (estimate) to a target state if such a sequence of controls can be proven to exist. Starting from a general theory of $\mathcal{M}$ at the initial instant, observations on the input–output behaviour of the system at each later instant are accepted by the system as new axioms; these are then used together with the previously generated theory to generate the current theory. The acronym COCOLOG is used to denote the family of first-order *conditional observer and controller logics* for any given input–state–output system. A semantics is supplied for each COCOLOG system in terms of interpretations of controlled transitions on a tree indexed by the possible sequences of input–output observations. Extralogical rules, including the conditional control rules, relating members of the family of theories of a COCOLOG system are presented in the form of a set of metalevel rules. Following the complete definition of a COCOLOG system, the consistency and completeness of the first-order theories in a COCOLOG system are established, decidability is obtained using a unique model property, and examples of the operation of a COCOLOG logic control system are given.

Key words. discrete event systems, finite machines, logic control

AMS subject classifications. 93, 68, 03

1. Introduction. In this paper we introduce certain partially ordered sets of first-order logical theories which we call *conditional observer and controller logics*, or COCOLOG systems for short. A COCOLOG system provides a logical system for (i) describing and reasoning about the state estimation and control of a given finite input–state–output machine $\mathcal{M}$ and (ii) acting upon $\mathcal{M}$ via a closed-loop logic regulator $\mathcal{R}$ carrying the corresponding COCOLOG system.

A particular subset of the formulas in each of the constituent logical languages of a COCOLOG system is called the set of *conditional control rules* (CCRs); these are formulated so that a certain control action is specified at an instant k when a certain set of past measurable (i.e., past observations-dependent) conditions C_k are fulfilled. In COCOLOG this translates precisely into the existence of a proof, in the corresponding first-order logical theory, of a formula describing the conditions C_k. Conditional control statements may include, for example, control commands that will steer the current system state or its estimate to a given target state $\mathbf{x}^{\mathbf{T}}$; such commands would be implemented whenever a sequence of controls achieving this objective can be proven to exist, with the uniqueness of the selected control ensured via a prescribed arrangement of the CCRs.

The conceptualization of a feedback regulator system adopted in this paper is qualitatively different from the usual notion of a feedback system. In the customary formulation, a feedback regulator, which shall be denoted by $\mathcal{R}$, is an input–output dynamical system whose inputs are typically the measured outputs of the controlled system $\mathcal{M}$ and whose outputs are the controlled inputs to $\mathcal{M}$. Hence the system and the regulator are objects of the same type, namely input(–state)–output dynamical systems. However, in our formulation, when $\mathcal{M}$ is in

* Received by the editors February 24, 1992; accepted for publication (in revised form) July 12, 1994.

[†] Department of Electrical Engineering, McGill University, Montreal, Québec, H3A 2A7, Canada. The research of this author was supported by Natural Sciences and Engineering Research Council of Canada grant A1329 and Institut National de Recherche en Informatique et en Automatique, Sophia Antipolis, France.

[‡] SPAR Aerospace Limited, 21025 Trans-Canada Highway, Ste. Anne de Bellevue, Québec H9X 1C0, Canada. The research of this author was performed at McGill University.

a feedback loop with a logic regulator $\mathcal{R}$, the situation is quite different and we now give a sketch of the operation of the system $(\mathcal{M}, \mathcal{R})$.

At each discrete time instant k, the previous input $\mathbf{u}_{k-1}$ and output $\mathbf{y}_k$ of the system $\mathcal{M}$, taking the values $\mathbf{u}^{\mathbf{i}}$ and $\mathbf{y}^{\mathbf{j}}$ in the finite sets $\mathbf{U}$ and $\mathbf{Y}$, are mapped extralogically into formulas called *observation axioms*, which are accepted by the regulator $\mathcal{R}$. In the present case, $\mathcal{R}$ is conceived of as a *dynamical logical system* mapping theories to theories (see the papers of Caines, Greiner, and Wang [1988], [1991]) and emitting outputs via a second extralogical map. Let the theory carried by $\mathcal{R}$ at the instant $k-1$ be denoted $Th(o_1^{k-1})$, where o_1^{k-1} denotes the sequence of observations over $[1, k-1]$. At the instant k the equality predicates relating the constant symbols $U(k-1)$ and $Y(k)$ at k to the observed quantities $\mathbf{u}^{\mathbf{i}}$ and $\mathbf{y}^{\mathbf{j}}$ are accepted as new information into the theory $Th(o_1^{k-1})$. By this we mean that these equality predicates are taken as new axioms to be added to $Th(o_1^{k-1})$. In addition, the state estimation axioms indexed by k are also accepted as new axioms. The theory carried by $\mathcal{R}$ is then transformed into the deductive closure of (i) $Th(o_1^{k-1})$, (ii) the observation axioms, and (iii) the state estimation axioms, and this is relabeled as $Th(o_1^k)$. By their design, the conditional control rules yield a unique, deducible, constant symbol for the input, $U(k)$, to $\mathcal{M}$ at k. The predicate defining this value is then mapped by the second extralogical transformation referred to above into the quantity which forms the input, lying in the finite set $\mathbf{U}$, to $\mathcal{M}$ at the instant k, and this predicate is also handed on to the subsequent axiom set as an observation axiom. The system $\mathcal{M}$ now performs another dynamical evolution step to generate the observed output $\mathbf{y}_{k+1} \epsilon \mathbf{Y}$ at the instant $k+1$, and this completes the dynamics-to-logical theory cycle $\mathcal{M} \rightarrow \mathcal{R} \rightarrow \mathcal{M}$.

The process above is initiated with the system $\mathcal{M}$ in its initial state $\mathbf{x}_0$ and the regulator $\mathcal{R}$ carrying only $Th_o \underline{\Delta} Th(o_1^o)$ (where $\underline{\Delta}$ denotes "denotes"), where $Th(o_1^o)$ consists of the deductive closure of the dynamical axioms of $\mathcal{M}$ (i.e., those describing the state transition and output maps) together with the logical axioms, the axioms for equality, the axioms for the reachability predicates, and the axioms for a simple arithmetic.

(It is evident that the $\mathcal{M} \rightarrow \mathcal{R} \rightarrow \mathcal{M}$ feedback loop may be generalized to a loop $\mathcal{L} \rightarrow \mathcal{R} \rightarrow \mathcal{L}$, where $\mathcal{L}$ is itself a dynamical logical system, but the investigation of this is left for future work.)

The exposition in this paper is in terms of finite state machines solely to establish the theory of COCOLOG systems in its simplest context. There is no obstruction, in principle, to extending the theory to machines in continuous time, extended state machines (Ostroff [1989]), and the automata of Ramadge–Wonham discrete events systems theory (Ramadge and Wonham [1987], [1989]).

The development of COCOLOG for dynamical control systems has a twofold motivation: first, the hierarchical nature of contemporary computer controlled systems may be better understood and enhanced by a study of regulator systems conceptualized at the logico-linguistic level. A notable example in this context is the capacity of reasoning systems to accept and operate on existential assertions, something a classical dynamical regulator is incapable of doing. Second, a control objective in COCOLOG, such as steering the system state to some state $\mathbf{x}^{\mathbf{T}}$, may be modified at any instant in the controlled machine's operation by conditions which are expressed via conjunctions and disjunctions of predicates; such conditions will be accepted by a COCOLOG regulator as new CCRs on the basis of which new control laws will be deduced. By their nature, conventional dynamical regulators cannot easily accept significantly modified objectives but must be redesigned to fit a new task. (We note that this is in contradistinction to the ability of conventional regulators with fixed control objectives to adapt to changing system dynamics.) It would also appear that information concerning system dynamics and objectives involving combinations of rules, necessary conditions, and sets of alternatives is best expressed logically, and hence a logic-based controller is most suitable for operating in this domain. (See Dyck and Caines [1995].)

Previous work on the formulation of the theory presented here and its ramifications has appeared in papers by Caines, Greiner, and Wang, [1991]; Caines, Wang, and Greiner [1988]; Caines and Wang [1989a], [1989b]; Wang and Caines [1991]; and, in particular, by Caines and Wang [1990] and in the thesis by Wang [1991], on which the exposition in this paper is based in part.

Other works which are analogous to, but different from, that presented here are the situated automaton work of Rosenschein and Kaebling [1987] and the situation calculus work of Reiter [1991]; these fall within the long-standing program of research in artificial intelligence to create logical decision-making systems which can predict and analyse the properties of formally specified systems (see, e.g., McCarthy and Hayes [1969] and Green [1969]) and which could, in principle, interact with them as they evolve in time.

Further, there is the line of research of Thistle and Wonham [1986], Ostroff [1987], [1989], and Ostroff and Wonham [1985]. In this work a fully elaborated temporal logic framework is presented to verify the correctness of feedback control algorithms for extended state machines. More recently, Kohn [1988], [1991] has devised a formulation of the logic control problem in which equational axiom systems describe the dynamical properties of continuous time systems and the declarative language of the system expresses optimization goals and constraints. Automatic automata-based procedures then create what is called a declarative control architecture. A more distantly related line of work is the use of temporal logic to study the evolution of programmed computational processes (see Harrel [1979] and Goldblatt [1987]).

We conclude this introduction with a brief remark about computational implementation. In its most direct implementation, a COCOLOG controller requires the real-time implementation of automatic theorem-proving (ATP) programs. The status of automatic theorem proving might be thought to indicate that this would be a formidable task. However, the restrictive nature of the system dynamical axioms and the option of defining restricted but nontrivial sets of control rules suggest the possibility of an efficient implementation via some type of ATP program. Current ATP software development in the context of COCOLOG applications (see, e.g., Wang and Caines [1991], [1992]; Dyck and Caines [1995]; and Caines, Mackling, and Wei [1992]) appears to confirm this conjecture.

2. Finite state machines. In this section we formally introduce our finite machine setup and define the notions of observation and control which will be required in subsequent sections of this paper.

DEFINITION 2.1. *A* (partially observed) finite (input–state–output) machine *is a quintuple* $\mathcal{M} = (\mathbf{U},\mathbf{X},\mathbf{Y},\mathbf{\Phi},\boldsymbol{\eta})$, *where* $\mathbf{U}$ *is a (finite) set of* inputs, $\mathbf{X}$ *is a (finite) set of* states, $\mathbf{Y}$ *is a (finite) set of* outputs, $\mathbf{\Phi} : \mathbf{X} \times \mathbf{U} \rightarrow \mathbf{X}$ *is a total* transition function, *and* $\boldsymbol{\eta} : \mathbf{X} \rightarrow \mathbf{Y}$ *is a total* output function.

Concerning notation, we shall sometimes write $\mathbf{u}_i^n$ for the $(n-i+1)$–element sequence $[\mathbf{u}_i, \mathbf{u}_{i+1}, \mathbf{u}_{i+2}, \ldots, \mathbf{u}_n]$, where $\mathbf{u}_j \epsilon \mathbf{U}$ denotes the input at the time instance $j \epsilon \mathbb{N}_+$ (where $\mathbf{u}_j$ is identified with $[\mathbf{u}_j]$) and ϵ will denote the empty input string. We shall use the notion that for all sequences $\mathbf{u}_i^n = \epsilon$ whenever $i > n$.

The dynamical evolution of a finite machine $\mathcal{M} = (\mathbf{U},\mathbf{X},\mathbf{Y},\mathbf{\Phi},\boldsymbol{\eta})$ can be displayed by taking $\mathbf{U}^*$ to be the set of all finite sequence of inputs and by extending $\mathbf{\Phi} : \mathbf{X} \times \mathbf{U} \rightarrow \mathbf{X}$ to $\mathbf{\Phi} : \mathbf{X} \times \mathbf{U}^* \rightarrow \mathbf{X}$, where for all i, $n \epsilon \mathbb{N}_+$, for all $\mathbf{u}_i^n \epsilon \mathbf{U}^*$, and for all $\mathbf{x} \epsilon \mathbf{X}$, $\mathbf{\Phi}$ is recursively defined by

$$\begin{aligned} &\mathbf{\Phi}(\mathbf{x}, \epsilon) = \mathbf{x}, \\ &\mathbf{\Phi}(\mathbf{x}, \mathbf{u}_i^n) = \mathbf{\Phi}(\mathbf{\Phi}(\mathbf{x}, \mathbf{u}_i), \mathbf{u}_{i+1}^n). \end{aligned} \tag{2.1}$$

The *initial* (respectively, *current*) *state dynamical observer problem* for a finite machine $\mathcal{M}$ is to estimate $\mathcal{M}$'s *initial* (respectively, *current*) state from observations on its inputs and outputs over a finite time period. An *initial* (respectively, *current*) *state dynamical observer* takes as input the observed behaviour of a system, i.e., a sequence of input–output pairs, and generates as output a sequence of estimates of the initial (respectively, *current*) state of the system. The notion of estimate is made precise in set theoretic terms in the following definitions.

DEFINITION 2.2. *The N-element state sequence $\mathbf{x}_1^N \in \mathbf{X}^N$ is an N*-consistent state sequence with respect to the input–output sequence $\mathbf{o}_1^N = [\langle \mathbf{y}_1 \rangle, \langle \mathbf{u}_1, \mathbf{y}_2 \rangle, \dots, \langle \mathbf{u}_{N-1}, \mathbf{y}_N \rangle] \in \mathbf{O}_1^N \underline{\Delta} \mathbf{Y} \times (\mathbf{U} \times \mathbf{Y})^{N-1}$ *if $\mathbf{x}_1^N$ satisfies*

$$\mathbf{x}_k = \mathbf{\Phi}(\mathbf{x}_1, \mathbf{u}_1^{k-1}) \quad \textit{and} \quad \mathbf{y}_k = \boldsymbol{\eta}(\mathbf{x}_k) \quad \textit{for all } k \in [1, \dots, N]. \tag{2.2}$$

The set of all N-element state sequences with respect to $\mathbf{o}_1^N$ is denoted $CSS(\mathbf{o}_1^N)$.

We shall denote the projection operator from $\mathbf{o}_1^N$ onto the mth u-coordinate, $1 \le m \le N-1$, by $\mathbf{u}_m(\cdot)$ and that which projects onto the nth y-coordinate, $1 \le n \le N$, by $\mathbf{y}_n(\cdot)$. (Note that $\mathbf{u}_m(\mathbf{o}_1^N) = \mathbf{u}_m(\mathbf{o}_1^{N'})$ for $1 \le m \le N-1 \le N'-1$ and $\mathbf{y}_n(\mathbf{o}_1^N) = \mathbf{y}_n(\mathbf{o}_1^{N'})$ for $1 \le n \le N \le N'$.)

DEFINITION 2.3. *An* initial state estimate *set, with respect to the N-element observation sequence, $\mathbf{o}_1^N$, written $\{\widehat{\mathbf{x}}_1\}(\mathbf{o}_1^N)$, is the set of initial elements of consistent state sequences corresponding to $\mathbf{o}_1^N$, i.e.,*

$$\{\widehat{\mathbf{x}}_1\}(\mathbf{o}_1^N) = \{\mathbf{x} \in \mathbf{X};\ \mathbf{x} = P_1(\mathbf{x}_1^N) \textit{for some} \, \mathbf{x}_1^N \in CSS(\mathbf{o}_1^N)\}, \tag{2.3}$$

where $P_1(\cdot)$ denotes projection on the first component of the argument. Analogously, a current state estimate set, *with respect to the N-element observation sequence $\mathbf{o}_1^N$, written $\{\widehat{\mathbf{x}}_N\}(\mathbf{o}_1^N)$, is the set of final elements of consistent state sequences with respect to $\mathbf{o}_1^N$, i.e.,*

$$\{\widehat{\mathbf{x}}_N\}(\mathbf{o}_1^N) = \{\mathbf{x} \in \mathbf{X};\ \mathbf{x} = P_N(\mathbf{x}_1^N) \textit{for some} \, \mathbf{x}_1^N \in CSS(\mathbf{o}_1^N)\}, \tag{2.4}$$

where $P_N(\cdot)$ denotes projection on the Nth component of the argument.

DEFINITION 2.4. *A finite machine $\mathcal{M} = (\mathbf{U}, \mathbf{X}, \mathbf{Y}, \mathbf{\Phi}, \boldsymbol{\eta})$ is said to be* initial *(respectively,* current) state observable *if there exists a $K \in \mathrm{N}_+$ such that for all $N \ge K$ and all $\mathbf{u}_1^N \in \mathbf{U}^N$, the initial (respectively, current) state estimate $\{\widehat{\mathbf{x}}_1\}(\mathbf{o}_1^N)$ (respectively, $\{\widehat{\mathbf{x}}_N\}(\mathbf{o}_1^N)$) is a singleton.*

Consider any finite machine $\mathcal{M} = (\mathbf{U}, \mathbf{X}, \mathbf{Y}, \mathbf{\Phi}, \boldsymbol{\eta})$. Then for any observation sequence, $\mathbf{o}_1^N \in \mathbf{O}^N$, the following equations hold:

$$\begin{aligned} \{\widehat{\mathbf{x}}_1\}\,(\mathbf{o}_1^{N+1}) &= \{\widehat{\mathbf{x}}_1\}(\mathbf{o}_1^N) \cap \mathbf{\Phi}^{-1}(\boldsymbol{\eta}^{-1}(\mathbf{y}_{N+1}), \mathbf{u}_1^N) \\ &= \bigcap_{k=1}^{N+1} \mathbf{\Phi}^{-1}(\boldsymbol{\eta}^{-1}(\mathbf{y}_k), \mathbf{u}_1^{k-1}), \end{aligned} \tag{2.5}$$

$$\begin{aligned} \{\widehat{\mathbf{x}}_{N+1}\}(\mathbf{o}_1^{N+1}) &= \mathbf{\Phi}(\{\widehat{\mathbf{x}}_N\}(\mathbf{o}_1^N), \mathbf{u}_N) \cap \boldsymbol{\eta}^{-1}(\mathbf{y}_{N+1}) \\ &\subseteq \bigcap_{k=1}^{N} \mathbf{\Phi}(\boldsymbol{\eta}^{-1}(\mathbf{y}_k), \mathbf{u}_k^N) \cap \boldsymbol{\eta}^{-1}(\mathbf{y}_{N+1}), \end{aligned} \tag{2.6}$$

with equality in (2.6) if $\mathbf{\Phi}(\cdot, \mathbf{u})$ is one to one for each $\mathbf{u} \in \mathbf{U}$. In (2.5) and (2.6), $\mathbf{\Phi}$ has been extended to take sets of states in its first argument: $\mathbf{\Phi} : \mathcal{P}(\mathbf{X}) \times \mathbf{U}^* \mapsto \mathcal{P}(\mathbf{X})$, where $\mathcal{P}(\mathbf{S})$ denotes the power set of the set $\mathbf{S}$; this is done by setting $\mathbf{\Phi}(\mathbf{A}, \mathbf{u}_1^*) = \{\mathbf{x} \in \mathbf{X}; \mathbf{x} = \mathbf{\Phi}(\mathbf{x}', \mathbf{u}_1^*)$

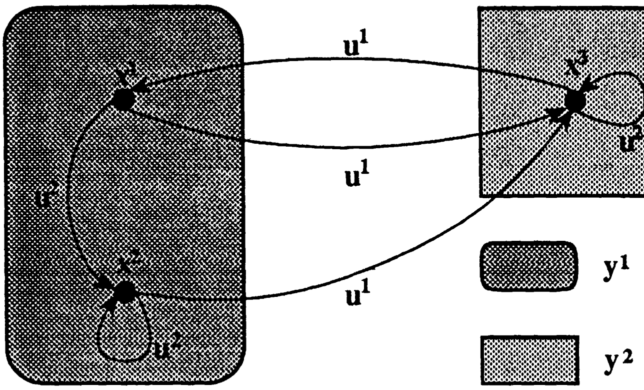


FIG. 1. *A three-state machine.*

for some $\mathbf{x}'\epsilon\mathbf{A}\}$. $\mathbf{\Phi}^{-1}$ denotes the inverse $\mathbf{\Phi}^{-1} : \mathcal{P}(\mathbf{X}) \times \mathbf{U}^* \mapsto \mathcal{P}(\mathbf{X})$, given by $\mathbf{\Phi}^{-1}(\mathbf{A},\mathbf{u}_1^*) = \{\mathbf{x}\epsilon\mathbf{X}; \mathbf{\Phi}(\mathbf{x},\mathbf{u}_1^*)\epsilon\mathbf{A}\}$ and similarly for η^{-1}. Finally, $\{\widehat{\mathbf{x}}_0\}(\mathbf{o}_1^0)$ is defined to be $\mathbf{X}$. (See the papers by Caines, Greiner, and Wang [1988], [1991].) We note that these equations possess the *predictor–corrector* form of many recursive algorithms in systems and control theory.

The corresponding partially ordered sets of initial and current state estimate sets will be referred to as the *initial* and *current observer trees*, respectively (for the given machine). Observe that although the state estimate sets may be identical for distinct input–output sequences, such distinct sequences uniquely define a directed acyclic graph with no confluences of edges. So at the cost of some redundancy, we shall label current state observation processes by the branches of the tree of input–output sequences and shall do the same for a COCOLOG family of theories.

Example 2.1. The following is an elementary example illustrating the notions introduced above. The finite machine $\mathcal{M}^3 = (\mathbf{U},\mathbf{X},\mathbf{Y},\mathbf{\Phi},\eta)$ is given in Fig. 1, where the input, state, and output sets are, respectively, $\mathbf{U} = \{\mathbf{u}^1, \mathbf{u}^2\}, \mathbf{X} = \{\mathbf{x}^1, \mathbf{x}^2, \mathbf{x}^3\}$, and $\mathbf{Y} = \{\mathbf{y}^1, \mathbf{y}^2\}$; the output function is given by $\eta(\mathbf{x}^1) = \eta(\mathbf{x}^2) = \mathbf{y}^1, \eta(\mathbf{x}^3) = \mathbf{y}^2$, and $\mathbf{\Phi}$ is given explicitly by the graph in the figure.

The notion of an N-consistent state sequence with respect to an input–output sequence is illustrated by the sequence of observations

$$\langle \mathbf{y}^1 \rangle, \langle \mathbf{u}^2, \mathbf{y}^1 \rangle, \langle \mathbf{u}^1, \mathbf{y}^2 \rangle,$$

which gives the corresponding 1-, 2-, and 3-consistent state sequences

$$\{\langle \mathbf{x}^1 \rangle, \langle \mathbf{x}^2 \rangle\}, \{\langle \mathbf{x}^1, \mathbf{x}^2 \rangle, \langle \mathbf{x}^2, \mathbf{x}^2 \rangle\}, \{\langle \mathbf{x}^1, \mathbf{x}^2, \mathbf{x}^3 \rangle, \langle \mathbf{x}^2, \mathbf{x}^2, \mathbf{x}^3 \rangle\}$$

and hence the sequence of initial state estimate sets

$$\{\mathbf{x}^1, \mathbf{x}^2\}, \{\mathbf{x}^1, \mathbf{x}^2\}, \{\mathbf{x}^1, \mathbf{x}^2\}$$

and the sequence of current state estimate sets

$$\{\mathbf{x}^1, \mathbf{x}^2\}, \{\mathbf{x}^2\}, \{\mathbf{x}^3\}.$$

We denote that the machine $\mathcal{M}^3$ is current state observable but not initial state observable.

DEFINITION 2.5. *A finite machine $\mathcal{M}$ is said to be* controllable *if for all* $\mathbf{x},\mathbf{x}'\epsilon\mathbf{X}$ *there exists a sequence* $\mathbf{u}_1^{n(\mathbf{x},\mathbf{x}')}$ *such that* $\mathbf{\Phi}(\mathbf{x}, \mathbf{u}_1^{n(\mathbf{x},\mathbf{x}')}) = \mathbf{x}'$ *or, equivalently, if for all* $\mathbf{x},\mathbf{x}'\epsilon\mathbf{X}$ *there exists* $n(\mathbf{x},\mathbf{x}')$ *such that* $\mathbf{x}'$ *is reachable from* $\mathbf{x}$ *in* $n(\mathbf{x},\mathbf{x}')$ *steps.*

By inspection, the machine in Example 2.1 is controllable.

The papers by Caines and Wang [1989] and Caines, Greiner, and Wang [1988], [1991] contain results concerning the combinatoric properties of initial and current state observer trees and give the following elementary dynamic programming theorem for the partially observed control problem: consider the problem of steering any unknown initial condition to an arbitrary target state x^T; then for current state observable and controllable finite state machines there exists a controller whose feedback control law is a function only of x^T and the sequence of current state estimate sets and which steers any initial unknown state to x^T.

3. COCOLOG: Syntax and semantics.

3.1. Syntax of COCOLOG L. The COCOLOG language consists of a set of symbols $S(L)$ and specified formation rules (or syntax). The subject of a COCOLOG language L is some given finite machine $\mathcal{M} = (\mathbf{U},\mathbf{X},\mathbf{Y},\mathbf{\Phi},\boldsymbol{\eta})$, where $\mathbf{U}$ is the set of inputs, $\mathbf{X}$ is the set of states, $\mathbf{Y}$ is the set of outputs, $\mathbf{\Phi}$ is a state transition function $\mathbf{\Phi} : \mathbf{X} \times \mathbf{U} \to \mathbf{X}$, and $\boldsymbol{\eta}$ is a state output function $\boldsymbol{\eta} : \mathbf{X} \to \mathbf{Y}$. It will be evident that we could construct a language L without reference to a finite machine, but this case will not be of interest to us in this paper.

We first define $S(L)$ *for the machine* $\mathcal{M}$ as

$$S(L) = \mathrm{Cons}_L \cup \mathrm{Var}_L \cup \mathrm{Fun}_L \cup \mathrm{Apr}_L \cup \mathrm{Qua}_L \cup \mathrm{Lco}_L \cup \{\perp\}.$$

The component sets of $S(L)$ are defined as follows.

Constant symbols. The constant symbols fall into the following typed subsets:

$$\mathrm{Cons}_L = \{u^1,\dots,u^m\} \cup \{x^1,\dots,x^N\} \cup \{y^1,\dots,y^p\} \cup \{0,1,\dots,k(N),k(N)+1\},$$

where $k(N)$ is an integer symbol and where the symbols $u^1,\dots,u^m; x^1,\dots,x^N$; and $y^1,\dots,y^p$ denote the individual elements of the input set $\mathbf{U}$, the state set $\mathbf{X}$, and the output set $\mathbf{Y}$, respectively.

Variable symbols. The variable symbols fall into the following typed subsets:

$$\mathrm{Var}_L = \{u,u',u'',\dots\} \cup \{x,x',x'',\dots\} \cup \{y,y',y'',\dots\} \cup \{l,l',l'',\dots\},$$

where the variables will be taken to be varying over different domains. $u,u',u'',\dots$ will be interpreted to represent elements in the set of inputs $\mathbf{U}$; the variables $x,x',x'',\dots,$ elements in the set of states $\mathbf{X}$; the variables $y,y',y'',\dots,$ elements in the set of outputs $\mathbf{Y}$; and the variables $l,l',l'',\dots,$ elements in the set of numbers $I_{k(N)}$; see §3.2 below.

Function symbols.

$$\mathrm{Fun}_L = \{\bar{\Phi}(\cdot,\cdot),\bar{\eta}(\cdot),+_L(\cdot,\cdot),-_L(\cdot,\cdot)\}.$$

Terms. The elements of the set Term_L are defined by the following:

(i) Each constant and variable symbol is a term; i.e., $\mathrm{Cons}_L \cup$ $\mathrm{Var}_L \subseteq \mathrm{Term}_L$.

(ii) If $t_1,\dots,t_n$ are terms and f is a function symbol of arity n, then $f(t_1,\dots,t_n)$ is a term.

(iii) The elements of Term_L are constructed only by steps (i) and (ii) above.

Atomic predicate symbols. $\mathrm{Apr}_L = \{\mathrm{E}q(\cdot,\cdot), Rbl(\cdot,\cdot,\cdot)\}$.

Sorting constraints. The predicate and function symbols are taken to satisfy the following sorting constraints on their arguments, where all admissible symbol strings are of finite length:

$\mathrm{E}q(\alpha,\beta)$: α and β are terms of the same type.

$\mathrm{E}q(\bar{\Phi}(a,b),c)$: a and c are symbols in $\{x^1,\dots,x^N\}$ or in $\{x,x',\dots\}$ or are symbols of the form $\bar{\Phi}(a,b)$, and b is a symbol in $\{u^1,\dots,u^m\}$ or in $\{u,u',\dots\}$.

$Eq(\bar{\eta}(a), d)$: a is a symbol as described above, and d is a symbol in $\{y^1, y^2, \ldots, y^p\}$ or in $\{y, y', y'', \ldots\}$ or is a symbol of the form $\bar{\eta}(a)$.

$Eq(+_L(i,j),k)$ and $Eq(-_L(i,j),k)$: i, j, and k are symbols in $\{0, 1, 2, \ldots, k(N), k(N)+1\}$ or in $\{l, l', \ldots\}$ or are of the form $+_L(i,j)$ or $-_L(i,j)$, where j is of the same type as i.

$Rbl(a, a', i)$: a and a' are symbols of the same type as the symbol a described above, and i is a symbol as described above.

Quantifiers. $\mathrm{Qua}_L = \{\forall\}$.

Logical connectives. $\mathrm{Lco}_L = \{\rightarrow\}$.

Logical constants. $\{\perp\}$.

Any *well-formed formula* of L is given by the *Backus–Naur* syntactic rule (see Goldblatt [1987]):

$$A ::= \varphi(t_1, \ldots, t_n) \mid A_1 \rightarrow A_2 \mid \perp \mid \forall v A_1,$$

where $\varphi(\cdot, \cdot, \ldots, \cdot) \in \mathrm{Apr}_L$; A_1 and A_2 are well-formed formulas; and $t_1, \ldots, t_n \in \mathrm{Term}_L$, in the sense that a well-formed formula is an expression that parses according to these rules until after a finite sequence of steps one halts at a set consisting only of elements of $S(L)$. The set of such formulas will be denoted Fma_L or L.

The other logical connectives $\neg$, $\vee$, $\wedge$, and $\longleftrightarrow$ and the quantifier $\exists$ are defined as follows, where the parentheses (and) are used whenever they clarify the meaning of a formula:

$$\begin{aligned}
\neg A &\triangleq A \rightarrow \perp, \\
A_1 \wedge A_2 &\triangleq \neg(A_1 \rightarrow \neg A_2), \\
A_1 \longleftrightarrow A_2 &\triangleq (A_1 \rightarrow A_2) \wedge (A_2 \rightarrow A_1), \\
A_1 \vee A_2 &\triangleq \neg A_1 \rightarrow A_2, \\
\exists v A &\triangleq \neg(\forall v \neg A).
\end{aligned}$$

3.2. Semantics of COCOLOG L. In the following discussion we shall distinguish symbols used in the specification of a (set-theoretic) finite machine $\mathcal{M}$ and those used in a COCOLOG language L; this is achieved by the convention that COCOLOG symbols will be the lightface versions of the boldface symbols denoting the constants, variables, and functions (in this case also with an overbar) of the machine $\mathcal{M}$. Following standard terminology (see, e.g., Goldblatt [1987]), an *L-structure $\mathcal{U}_L = (\mathbf{D}, I)$*, or an *interpretation I* (with *domain* $\mathbf{D}$), is a pair where, first, $\mathbf{D} = \mathbf{U} \cup \mathbf{X} \cup \mathbf{Y} \cup \mathbf{I}_{\mathbf{k(N)}}$, where $\mathbf{U}$, $\mathbf{X}$, and $\mathbf{Y}$ are the sets appearing in the specification of some *arbitrary* finite machine $\mathcal{M}$ as specified in §2 (which is not necessarily the machine which defined the language L) and $\mathbf{I}_{\mathbf{k(N)}} = \{\mathbf{0}, \mathbf{1}, \mathbf{2}, \ldots, \mathbf{k(N)}, \mathbf{k(N)}+\mathbf{1}\}$ is the set of integers between 0 and $k(N)+1$, inclusive, and, second, I is an interpretation function which respects the typing of L and which in COCOLOG is defined as follows:

$$\begin{aligned}
I(\bar{\Phi}) &: \mathbf{X} \times \mathbf{U} \rightarrow \mathbf{X}, \\
I(\bar{\eta}) &: \mathbf{X} \rightarrow \mathbf{Y}, \\
I(+_L) &= +_{\mathbf{k(N)}} : \mathbf{I}_{\mathbf{k(N)}} \times \mathbf{I}_{\mathbf{k(N)}} \rightarrow \mathbf{I}_{\mathbf{k(N)}}, \\
I(-_L) &= -_{\mathbf{k(N)}} : \mathbf{I}_{\mathbf{k(N)}} \times \mathbf{I}_{\mathbf{k(N)}} \rightarrow \mathbf{I}_{\mathbf{k(N)}}, \\
I(c) &\in \mathbf{D}, \quad \text{where } c \in \mathrm{Cons}_L, \\
I(Eq) &= \{(\mathbf{t}, \mathbf{t}'); \mathbf{t}, \mathbf{t}' \in \mathbf{D}, \mathbf{t} = \mathbf{t}'\} \subseteq \mathbf{D}^2, \\
I(Rbl) &\subseteq \mathbf{X}^2 \times \mathbf{I}_{\mathbf{k(N)}}.
\end{aligned}$$

Note that without further conditions being imposed, $I(\cdot)$ may be a many-to-one function on Cons_L.

Here $\mathbf{k(N)}$ is taken to be an arbitrary number greater than $|\mathbf{X}|^2$; this is motivated by the fact that (see Caines, Greiner, and Wang [1991]), counting from the root, a current state observer tree can have at most $|\mathbf{X}|^2$ layers at which there is a reduction in size of some nonsingleton node.

Addition $+_{\mathbf{k(N)}}$ and subtraction $-_{\mathbf{k(N)}}$ in the arithmetic $\mathbf{A_{k(N)}}$ over the finite set of integers $\{\mathbf{0,1,2,\ldots,k(N),k(N)+1}\}$, where $\mathbf{k(N)+1}$ plays the role of infinity, are defined by the following expressions, where we follow the convention that $+_{\mathbf{k(N)}}$ and $-_{\mathbf{k(N)}}$ denote the addition and subtraction in the L-structure $\mathcal{U}_L$ and $+$ and $-$ denote the standard integer arithmetical operations:

$$\mathbf{a} +_{\mathbf{k(N)}} \mathbf{b} = \begin{cases} \mathbf{a+b} & \text{if}\, \mathbf{a+b} \leq \mathbf{k(N)}, \\ \mathbf{k(N)+1} & \text{if}\, \mathbf{a+b} > \mathbf{k(N)}, \end{cases}$$

$$\mathbf{a} -_{\mathbf{k(N)}} \mathbf{b} = \begin{cases} \mathbf{a-b} & \text{if}\, \mathbf{a-b} \geq \mathbf{0}\, \text{and}\, \mathbf{a} \neq \mathbf{k(N)+1}\, \text{and}\, \mathbf{b} \neq \mathbf{k(N)+1}, \\ \mathbf{k(N)+1} & \text{if}\, \mathbf{a-b} < \mathbf{0}\, \text{or if}\, \mathbf{a} = \mathbf{k(N)+1}. \end{cases}$$

These finite-integer arithmetical operations are chosen to express the dynamical properties of $\mathcal{M}$ over a bounded integral number of steps or discrete time instants.

In COCOLOG, a $\mathcal{U}_L$*-valuation* (i.e., I*-valuation*) is a function V :$\text{Var}_L \longrightarrow \mathbf{D}$ satisfying

$$V(v) \,\epsilon \begin{cases} \mathbf{X} & \text{if}\, v \,\epsilon \{x, x', x'', \ldots\}, \\ \mathbf{Y} & \text{if}\, v \,\epsilon \{y, y', y'', \ldots\}, \\ \mathbf{U} & \text{if}\, v \,\epsilon \{u, u', u'', \ldots\}, \\ \mathbf{I_{k(N)}} & \text{if}\, v \,\epsilon \{l, l', l'', \ldots\}, \end{cases}$$

which can be extended to V : $\text{Term}_L \longrightarrow \mathbf{D}$ by

$$V(t) = \begin{cases} V(t) & \text{if}\, t \,\epsilon\, \text{Var}_L, \\ I(t) & \text{if}\, t \,\epsilon\, \text{Cons}_L, \\ I(f)(V(t_1), V(t_2), \ldots, V(t_n)) & \text{if}\, t = f(t_1, t_2, \ldots, t_n)\, \text{and}\, f \,\epsilon\, \text{Fun}_L \text{ is a} \\ & \qquad \text{function of arity}\, n. \end{cases}$$

We take $V \sim_v V'$ to mean that V and V' are identical except in the value they assign to v and

$$V(v/\mathbf{d}) = V' \quad \text{iff}\; V \sim_v V' \,\text{and}\, V'(v) = \mathbf{d}.$$

$\mathcal{U}_L \models A[V]$ stands for the property that a structure $\mathcal{U}_L$ (or interpretation I) satisfies a formula A under the valuation V; this is defined recursively by

$$\begin{aligned} &\mathcal{U}_L \models \text{E}q(t,t')[V] \quad \text{iff}\; V(t) = V(t'), \\ &\mathcal{U}_L \models Rbl(x,x',k)[V] \quad \text{iff}\, (V(x), V(x'), V(k)) \,\epsilon\, I(Rbl), \\ &\mathcal{U}_L \models (A_1 \rightarrow A_2)[V] \quad \text{iff}\, \mathcal{U}_L \models A_1[V]\, \text{implies}\, \mathcal{U}_L \models A_2[V], \\ &\mathcal{U}_L \models \perp [V] \quad \text{iff not}\, \mathcal{U}_L \models \perp [V], \\ &\mathcal{U}_L \models \forall v A[V] \quad \text{iff for all}\, \mathbf{d} \,\epsilon\, \mathbf{D},\, \text{it is the case that}\, \mathcal{U}_L \models A[V(v/\mathbf{d})]. \end{aligned}$$

The property that the formula A is *true* in the structure $\mathcal{U}_L$, written $\mathcal{U}_L \models A$, or equivalently that I (on the domain $\mathbf{D}$) is a *model* for A, is defined by

$$\mathcal{U}_L \models A \quad \text{iff for all}\; V \;\text{it is the case that}\; \mathcal{U}_L \models A[V];$$

and A is *false* in $\mathcal{U}_L$, written $\mathcal{U}_L \not\models A$, or I (on the domain $\mathbf{D}$) is *not a model* for A, is defined by

$$\mathcal{U}_L \not\models A \quad \text{iff for some}\; V \;\text{it is the case that}\; \mathcal{U}_L \not\models A[V],\; \text{i.e.,}\, \mathcal{U}_L \models \neg A[V].$$

In standard terminology, a formula A is called *valid* if it is true in all structures $\mathcal{U}_L$ (i.e., all interpretations I), which is the case if and only if for all $\mathcal{U}_L$, $\mathcal{U}_L \models A$. A formula A is *satisfiable* if there exists some structure $\mathcal{U}_L$ and some valuation V such that the satisfaction relation $\mathcal{U}_L \models A[V]$ holds. Obviously a formula A is valid if and only if $\neg A$ is unsatisfiable. Unless we relativize to a set of interpretations, the only valid formulas in a theory are those given by the logical axiom schemata given below. This is because these must hold for any set theoretic interpretation. Other formulas, in particular the special axioms of a particular theory, may not be true under some interpretation.

The constant symbols, variable symbols, and function symbols, and hence the terms of a COCOLOG language, are typed because the constant and variable symbols are sorted as indicated in §3.1; furthermore, all interpretation functions I together with their associated valuations respect this sorting as specified above. As a result, the term model used in this paper is the same as that used in conventional typed logic (see, e.g., Goldblatt [1987]).

3.3. Axiomatic theory of Th_0. The formal logical theory Th_0 of the language L for $\mathcal{M}$ consists of a set of axioms, that is to say, a set of formulas from Fma_L which shall be required to hold in the intended models (in particular for $\mathcal{M}$), and the set of inference rules operating on Fma_L; these are taken together with the concepts of proof and theoremhood.

In this subsection we first present the axiomatic COCOLOG theory Th_0 corresponding to the information at the root node of the observer tree for a given finite machine $\mathcal{M}$. Further specializations of this theory to $Th(o_1^k)$ are obtained as observations are collected (as time proceeds) on the input–output behaviour of $\mathcal{M}$. This development is presented in the next subsection. Note that since the dynamics of $\mathcal{M}$ are known, the incomplete information aspect of $Th(o_1^k)$ is due solely to the partial observation nature of the problem.

Th_0 has a set of *logical axioms*, a set of *equality axioms*, a set of *arithmetic axioms*, and a set of *special axioms* which specify facts concerning the subject that the logic describes (in at least one of its interpretations). Correspondingly, $Th(o_1^k)$ is a logical theory that has the *observation axioms* and the *state estimation axioms* (all defined below) added to the logical theory Th_0.

The first two sets of axiom schemata below are relatively standard (see, e.g., Mendelson [1964]).

Logical axiom schemata. For all $A, B, C \in Fma_L$, $t \in \mathrm{Term}_L$, and $v \in \mathrm{Var}_L$,

$$(\mathrm{AXM}^{\mathrm{log}})\qquad \begin{array}{rl} \text{(i)} & A \rightarrow (B \rightarrow A) \\ \text{(ii)} & (A \rightarrow (B \rightarrow C)) \rightarrow ((A \rightarrow B) \rightarrow (A \rightarrow C)); \\ \text{(iii)} & (\neg B \rightarrow \neg A) \rightarrow ((\neg B \rightarrow A) \rightarrow B); \\ \text{(iv)} & \forall v\, A(v) \rightarrow A(t); \\ \text{(v)} & \forall v (A \rightarrow B) \rightarrow (A \rightarrow \forall v B), \quad v \text{ not free in } A. \end{array}$$

Any formula having the same form as one of these logical axiom schemata shall be called a logical axiom.

Equality axiom schemata. In the following equality axiom schemata, $t, t', t'' \in \mathrm{Term}_L$, $f \in \mathrm{Fun}_L$, and $P \in \mathrm{Apr}_L$:

$$(\mathrm{AXM}^{\mathrm{eq}}(L))\qquad \begin{array}{rl} \text{(i)} & \mathrm{E}q(t,t); \\ \text{(ii)} & \mathrm{E}q(t,t') \rightarrow \mathrm{E}q(t',t); \\ \text{(iii)} & \mathrm{E}q(t,t'') \wedge \mathrm{E}q(t',t'') \rightarrow \mathrm{E}q(t,t'); \\ \text{(iv)} & \mathrm{E}q(t,t') \rightarrow \mathrm{E}q(f(t),f(t')); \\ \text{(v)} & \mathrm{E}q(t,t') \rightarrow (P(t) \rightarrow P(t')). \end{array}$$

Arithmetic axiom schemata. For any constant symbols l, l', l'' with corresponding elements $\mathbf{l}, \mathbf{l}', \mathbf{l}'' \in \mathbf{I}_{\mathbf{k}(\mathbf{N})}$, if $\mathbf{l} +_{\mathbf{k}(\mathbf{N})} \mathbf{l}' = \mathbf{l}''$, or $\mathbf{l} -_{\mathbf{k}(\mathbf{N})} \mathbf{l}' = \mathbf{l}''$, respectively, then the following axioms for the arithmetic $A_{k(N)}$ hold respectively:

$$(\mathrm{AXM}^{\mathrm{arith}}(L)) \qquad \mathrm{E}q(l +_L l', l''), \qquad \mathrm{E}q(l -_L l', l'').$$

Finite machine axiom schemata. The special axioms for a given finite machine $\mathcal{M}$ are as follows: for any pair of constant symbols x^i, x^j and constant symbol u^l, if $\mathbf{x}^\mathbf{j} = \mathbf{\Phi}(\mathbf{x}^\mathbf{i}, \mathbf{u}^\mathbf{l})$ holds for $\mathcal{M}$, then the following *dynamic axiom* holds:

$$(\mathrm{AXM}^{\mathrm{dyn}}(L)) \qquad \mathrm{E}q(\bar{\Phi}(x^i, u^l), x^j).$$

Further, for any pair of constant symbols x^i, y^j such that $\boldsymbol{\eta}(\mathbf{x}^\mathbf{i}) = \mathbf{y}^\mathbf{j}$ holds for $\mathcal{M}$, the following *output axiom* holds:

$$(\mathrm{AXM}^{\mathrm{out}}(L)) \qquad \mathrm{E}q(\bar{\eta}(x^i), y^i).$$

Reachability axioms. We recursively define the *reachability predicate* $Rbl(\cdot, \cdot, \cdot)$ by the following axioms:

$$(\mathrm{AXM}^{Rbl}(L)) \quad \begin{array}{l} 0.\ \forall x \forall x', \mathrm{E}q(x, x') \longleftrightarrow Rbl(x, x', 0), \\ 1.\ \forall x \forall x', (\exists u, \mathrm{E}q(\bar{\Phi}(x, u), x')) \longleftrightarrow Rbl(x, x', 1), \\ 2.\ \forall x \forall x'' \forall l, \mathrm{E}q(l, k(N) + 1) \vee [\{\exists x' \exists u, Rbl(x', x'', l) \\ \quad \wedge \mathrm{E}q(\bar{\Phi}(x, u) x')\} \longleftrightarrow Rbl(x, x'', l +_L 1)]. \end{array}$$

The reachability axioms specify the l-step reachability relation $Rbl(x, x', l)$ between any pair of states x, x', with axioms 0 and 1 having obvious interpretations. Axiom 2 first excludes consideration of the infinity case and then characterizes reachability on the finite numbers in the arithmetic; specifically it states that either l equals $k(N) + 1$ or x'' is reachable from x in $l + 1$ steps if and only if there is an intermediate state x' such that state x'' is reachable in l steps and x' is reachable from x in one step. We note that if $l = k(N)$, i.e., the predecessor of $k(N) + 1$, then x'' is reachable from x in $k(N) + 1$ steps if and only if x'' is reachable from some x' in $k(N)$ steps, where x' is reachable from x in one step. We further note this does not necessarily make all states mutually reachable in $k(N) + 1$ steps.

Rules of inference. R1. Modus ponens (MP):

$$\frac{A, A \rightarrow B}{B} \quad \text{where } A, B \in Fma_L.$$

R2. Generalization:

$$\frac{A}{\forall v A} \quad \text{where } v \in Var_L, A \in Fma_L.$$

DEFINITION 3.1. *Let Σ denote the set of special axioms of $\mathcal{M}$ expressed in L, i.e., $\Sigma = \{\mathrm{AXM}^{\mathrm{arith}}(L), \mathrm{AXM}^{\mathrm{dyn}}(L), \mathrm{AXM}^{\mathrm{out}}(L), \mathrm{AXM}^{Rbl}(L)\}$. Then Σ is said to be the* axiom set for the finite machine $\mathcal{M}$.

A *proof* in L is a sequence of formulas $A_1, \ldots, A_k$ in Fma_L where A_i, $1 \leq i \leq k$, is either an axiom or a direct consequence of previous formulas via R1 or R2. The last formula A_k in the sequence is called a *theorem*, and $A_1, \ldots, A_{k-1}$ is a proof of the theorem A_k. A formula A is a theorem of a first-order theory with equality, written $\vdash A$, if, in a proof of A, only logical axioms and equality axioms have been involved; A is called a *consequence* (or *theorem*) of Σ, written $\Sigma \vdash A$, if, in a proof of A, only logical axioms, equality axioms, and axioms in Σ are involved. For brevity we write $Th_0 \equiv Th_0(L)$ for the set of theorems of Σ; hence we have $Th_0 = \{A : \Sigma \vdash A\}$ and we use the standard notation $Th_0 \vdash A$, which is customarily read as A is a *theorem* of, or is *provable* (*derivable*) in, the theory Th_0.

A structure $\mathcal{U}_L$ (i.e., an interpretation I on a domain $\mathbf{D}$) is called a *model* of the theory Th_0 if and only if all the axioms Σ of Th_0 are interpreted *true* in $\mathcal{U}_L$, i.e., if and only if I is a model for each axiom of Th_0.

Example 3.1. An illustration of the construction of a model of the COCOLOG theory Th_0 may be given in terms of the simple machine $\mathcal{M}^3$.

Let L_3 denote the first-order COCOLOG language whose constant symbols are taken to be

$$\mathrm{Cons}_{L_3} = \{u^1, u^2\} \cup \{x^1, x^2, x^3\} \cup \{y^1, y^2\} \cup \{0, 1, \ldots, 9, 9+1\}.$$

For L_3 let the variable symbols, function symbols, atomic predicate symbols, sorting constraints, quantifiers, logical connectives, and logical constants be taken to be exactly as specified in §3.1 above. The first-order COCOLOG theory Th_0 of the first-order language L_3 is taken to consist of the general logical axioms, the equality axioms, the axioms of reachability, and the rules of inference as given in §3.3 above. The special machine axioms are taken to be as follows:

$$(\mathrm{AXM}^{\mathrm{dyn}}(L_3)) \quad \begin{matrix} \mathrm{E}q(\bar{\Phi}(x^1, u^1), x^3) & \mathrm{E}q(\bar{\Phi}(x^2, u^2), x^3) & \mathrm{E}q(\bar{\Phi}(x^3, u^1), x^1), \\ \mathrm{E}q(\bar{\Phi}(x^1, u^2), x^2) & \mathrm{E}q(\bar{\Phi}(x^2, u^2), x^2) & \mathrm{E}q(\bar{\Phi}(x^3, u^2), x^3), \end{matrix}$$

$$(\mathrm{AXM}^{\mathrm{out}}(L_3)) \quad \mathrm{E}q(\bar{\eta}(x^1), y^1) \qquad \mathrm{E}q(\bar{\eta}(x^2), y^1) \qquad \mathrm{E}q(\bar{\eta}(x^3), y^2).$$

The axioms $\mathrm{AXM}^{\mathrm{arith}}(L_3)$ of the finite arithmetic A_9 of the theory Th_0 shall be taken to be those satisfied by the following finite arithmetic model $\mathbf{A_9}$, where here $\mathbf{k(N)} = 9$. Let $\mathbf{I_9} = \{\mathbf{0}, \mathbf{1}, \mathbf{2}, \ldots, \mathbf{9} + \mathbf{1}\}$; then $\mathbf{A_9} = (\mathbf{I_9}, +_\mathbf{9}, -_\mathbf{9})$ is given by the following addition and subtraction operations:

$$\mathbf{a} +_\mathbf{9} \mathbf{b} = \begin{cases} \mathbf{a}+\mathbf{b} & \text{if}\, \mathbf{a}+\mathbf{b} \le \mathbf{9}, \\ \mathbf{9}+\mathbf{1} & \text{if}\, \mathbf{a}+\mathbf{b} > \mathbf{9}, \end{cases}$$

$$\mathbf{a} -_\mathbf{9} \mathbf{b} = \begin{cases} \mathbf{a}-\mathbf{b} & \text{if}\, \mathbf{a}-\mathbf{b} \ge \mathbf{0} \,\text{and}\, \mathbf{a} \ne \mathbf{9}+\mathbf{1} \,\text{and}\, \mathbf{b} \ne \mathbf{9}+\mathbf{1}, \\ \mathbf{9}+\mathbf{1} & \text{if}\, \mathbf{a}-\mathbf{b} < \mathbf{0} \,\text{or if}\, \mathbf{a} = \mathbf{9}+\mathbf{1}. \end{cases}$$

We denote the special axioms $\mathrm{AXM}^{\mathrm{arith}}(L_3)$, $\mathrm{AXM}^{\mathrm{dyn}}(L_3)$, $\mathrm{AXM}^{\mathrm{out}}(L_3)$, and $\mathrm{AXM}^{Rbl}(L_3)$ collectively by $\Sigma(L_3)$.

Next consider the input–state–output machine $\mathcal{M}^3$, with (two-element) input set $\mathbf{U} = \{\mathbf{u}^1, \mathbf{u}^2\}$; (three-element) state set $\mathbf{X} = \{\mathbf{x}^1, \mathbf{x}^2, \mathbf{x}^3\}$; (two-element) output set $\mathbf{Y} = \{\mathbf{y}^1, \mathbf{y}^2\}$; state transition function $\boldsymbol{\Phi} : \mathbf{X} \times \mathbf{U} \to \mathbf{X}$ given by

$$\begin{matrix} \boldsymbol{\Phi}(\mathbf{x}^1, \mathbf{u}^1) = \mathbf{x}^3, & \boldsymbol{\Phi}(\mathbf{x}^2, \mathbf{u}^1) = \mathbf{x}^3, & \boldsymbol{\Phi}(\mathbf{x}^3, \mathbf{u}^1) = \mathbf{x}^1, \\ \boldsymbol{\Phi}(\mathbf{x}^1, \mathbf{u}^2) = \mathbf{x}^2, & \boldsymbol{\Phi}(\mathbf{x}^2, \mathbf{u}^2) = \mathbf{x}^2, & \boldsymbol{\Phi}(\mathbf{x}^3, \mathbf{u}^2) = \mathbf{x}^3; \end{matrix}$$

and state output function $\boldsymbol{\eta} : \mathbf{X} \to \mathbf{Y}$ given by

$$\boldsymbol{\eta}(\mathbf{x}^1) = \mathbf{y}^1, \quad \boldsymbol{\eta}(\mathbf{x}^2) = \mathbf{y}^1, \quad \boldsymbol{\eta}(\mathbf{y}^3) = \mathbf{x}^2.$$

We claim that a model (in the sense of §3.2) for Th_0 is given by $(\mathbf{D}, I)$, where $\mathbf{D} \,\underline{\triangle}\, \mathbf{U} \cup \mathbf{X} \cup \mathbf{Y} \cup \mathbf{I_9}$ and where I, the interpretation function defined in §3.2, is specified as follows: I maps from the language symbols $c \,\epsilon\, \mathrm{Cons}_L$ in L_3 to the corresponding elements of $\mathbf{D}$. The function symbols $\bar{\Phi}(\cdot, \cdot)$ and $\bar{\eta}(\cdot)$ map, respectively, to $I(\bar{\Phi}) = \boldsymbol{\Phi}$, the function from $\mathbf{X} \times \mathbf{U}$ to $\mathbf{X}$ denoted by the equations above, and $I(\bar{\eta}) = \boldsymbol{\eta}$, the function from $\mathbf{X}$ to $\mathbf{Y}$ denoted by the equations above. The binary arithmetic function symbols $+_L$ and $-_L$ map to $I(+_L) = +_{\mathbf{9}}$ and $I(-_L) = -_{\mathbf{9}}$, whose values on $I_9 \times I_9$ are given above.

Depending upon the I-valuation $V(\cdot)$ given by the particular interpretation I, $v \,\epsilon\, \{u, u', u'', \ldots\}$ maps to some $V(v) \,\epsilon\, \mathbf{U} \subset \mathbf{D}$, a variable $v \,\epsilon \{x, x', x'', \ldots\}$ maps to some $V(v) \,\epsilon\, \mathbf{X} \subset \mathbf{D}$, and, similarly, $v \,\epsilon \{y, y', y'', \ldots\}$ to some $V(v) \,\epsilon\, \mathbf{Y} \subset \mathbf{D}$, and $v \,\epsilon \{l, l', l'', \ldots\}$ to some $V(v) \in \mathbf{I}_9 \epsilon \mathbf{D}$.

The predicate symbols $\mathrm{E}q(\cdot\,, \cdot)$ and $Rbl(\cdot\,, \cdot\,, \cdot)$ are mapped to the relations on $\mathbf{D} \times \mathbf{D}$ and on $\mathbf{X} \times \mathbf{X} \times \mathbf{I}_9 \subset \mathbf{D} \times \mathbf{D} \times \mathbf{D}$ as specified in §3.2.

The claim that this definition of $(\mathbf{D}, I)$ provides a model for $\mathrm{AXM}^{\mathrm{log}} \cup \mathrm{AXM}^{\mathrm{eq}}(L_3) \cup \Sigma(L_3)$ is established by inspection, namely by the explicit verification that the function relations above for $\mathbf{\Phi}$ and $\boldsymbol{\eta}$ satisfy the machine axioms and that the interpretation of the equality and reachability predicates satisfy the axioms given in $\Sigma(L_3)$. The corresponding verification of $\mathrm{AXM}^{\mathrm{arith}}(L_3)$ for A_9 is immediate by virtue of its construction in terms of $\mathbf{A_9}$.

The set of theorems of $Th_0(L_3)$ is exactly the set of formulas which are true in all models satisfying the axioms $\Sigma(L_3)$; this is guaranteed by the completeness result proved below in §4.

To illustrate logical deduction in COCOLOG we shall give a proof of the theorem $Rbl(x^1, x^3, 2)$ in theory Th_0; this theorem asserts that the state x^1 is controllable to the state x^3 in two steps. The logical truth of (i.e., satisfaction of) $Rbl(x^1, x^3, 2)$ in the model $\mathcal{U}_{L_3} \triangleq (\mathbf{D}, I)$ of the axioms $\Sigma(L_3)$ can be verified from the model in Fig. 1. In standard notation we have $\mathcal{U}_{L_3} \models Rbl(x^1, x^3, 2)$, which reads that the relation corresponding to the formula $Rbl(x^1, x^3, 2)$ holds in the model $\mathcal{U}_{L_3}$.

Proof of $Rbl(x^1, x^3, 2)$. 1. $\mathrm{E}q(\bar{\Phi}(x^2, u^1), x^3)$: $\mathrm{AXM}^{\mathrm{dyn}}(L_3)$.

2. $\exists u, \mathrm{E}q(\bar{\Phi}(x^2, u), x^3)$: 1, $\mathrm{AXM}^{\mathrm{log}}$(iv), and definition of $\exists$.

3. $Rbl(x^2, x^3, 1)$: 2, $\mathrm{AXM}^{\mathrm{log}}$(iv), $\mathrm{AXM}^{Rbl}(L_3)(1)$, and MP.

4. $\mathrm{E}q(\bar{\Phi}(x^1, u^2), x^2)$: $\mathrm{AXM}^{\mathrm{dyn}}(L_3)$.

5. $\exists u, \mathrm{E}q(\bar{\Phi}(x^1, u), x^2)$: 4, $\mathrm{AXM}^{\mathrm{log}}$(iv), and definition of $\exists$.

6. $\neg \mathrm{E}q(1, 9 + 1)$: $\mathrm{AXM}^{\mathrm{arith}}(L_3)$.

7. $\exists x' \exists u, Rbl(x, x^3, 1) \wedge \mathrm{E}q(\bar{\Phi}(x^1, u), x)$: $\Leftrightarrow Rbl(x^1, x^3, 1 +_L 1)$, 6, MP, and instantiation of $\mathrm{AXM}^{Rbl}(2)$.

8. $Rbl(x^2, x^3, 1) \wedge \exists u, \mathrm{E}q(\bar{\Phi}(x^1, u), x^2)$: 3 and 5.

9. $\exists x' \exists u, Rbl(x', x^3, 1) \wedge \ \mathrm{E}q(\bar{\Phi}(x^1, u), x')$: 8 and definition of $\exists$.

10. $Rbl(x^1, x^3, 1 +_L 1)$: 7, 9, and MP.

11. $\mathrm{E}q(2, 1 +_L 1)$: Arithmetic axiom.

12. $Rbl(x^1, x^3, 2)$: 10, 11, and $\mathrm{AXM}^{\mathrm{eq}}(L_3)$(v).

3.4. The COCOLOG language $L(o_1^k)$: Syntax and semantics. The language $L^k \triangleq L(o_1^k)$ is an extension of the language L obtained by adding new constant symbols and atomic predicates in the following way:

$$S(L^k) = S(L) \bigcup_{j=1}^{k} \mathrm{Cons}_{L^j} \bigcup_{j=1}^{k} \mathrm{Apr}_{L^j},$$

where $\mathrm{Apr}_j = \{CSE_j(\cdot)\}$, $j \geq 1$; $\mathrm{Cons}_{L^1} = \{Y(1)\}$; and $\mathrm{Cons}_{L_j} = \{U(j-1), Y(j)\}, j \geq 2$, where $U(j-1), 2 \leq j \leq k, Y(j), 1 \leq j \leq k$, are constant names equal, respectively, to a sequence of names of input set elements $\mathbf{u}_1^{k-1}$ and output set elements $\mathbf{y}_1^k$ in $\mathbf{o}_1^k$; this sequence names the set $\mathbf{o}_1^k = (\mathbf{y}_1, \langle \mathbf{u}_1, \mathbf{y}_2 \rangle, \ldots, \langle \mathbf{u}_{k-1}, \mathbf{y}_k \rangle)$ indexing the language $L^k = L(o_1^k)$. Here we maintain the convention that lightface subscripts and superscripts on boldface symbols denote time indices. The sort of each new constant symbol in L^k is given by

$U(j)$ is a symbol of type $\{u^1, \ldots, u^m\}$ and $Y(j)$ is a symbol of type $\{y^1, \ldots, y^p\}$.

Set $L \equiv Fma_{L^0} \equiv Fma_L$; then the set of well-formed formulas $L^j \equiv Fma_{L^j}$, $j \geq 1$, is defined as the set formulas which parse according to

$$A ::= \varphi(t_1, \ldots, t_n) | CSE_k(x) | B | A' \rightarrow A'' | \forall v A',$$

where $\varphi \in \mathrm{Apr}_{L^j}, t_1, \ldots, t_n \in \mathrm{Term}_{L^j}$, $x \in \mathrm{Term}_j$ is of state type, $B \in Fma_{L^{j-1}}$, and A', $A'' \in Fma_{L^j}$.

Let I be the mapping associated with the L-structure $\mathcal{U}_L = (\mathbf{D}, I)$. An L^k-structure $\mathcal{U}_{L^k} = (\mathbf{D}, I_k)$ is a pair where the interpretation function I_k is an extension of the mapping I into an *arbitrary* machine $\mathcal{M}$ as given by

$I_k(U(j-1)) \in \mathbf{U}$, where the image is the input string element $\mathbf{u}_{j-1}(\mathbf{o}_1^j)$, $1 \leq j \leq k$, lying in $\mathbf{U}$;

$I_k(Y(j)) \in \mathbf{Y}$, where the image is the output string element $\mathbf{y}_j(\mathbf{o}_1^j)$, $1 \leq j \leq k$, lying in $\mathbf{Y}$;

$I_k(CSE_j) \subseteq \mathbf{X}$, $1 \leq j \leq k$.

The satisfaction relation $\mathcal{U}_{L^k} \models A[V]$ is the extension of $\mathcal{U}_L \models A[V]$ obtained by adding the following definitions:

$\mathcal{U}_{L^k} \models CSE_k(x)[V]$ iff $V(x) \in I_k(CSE_k)$,

$\mathcal{U}_{L^k} \models B[V]$ iff $\mathcal{U}_{L^{k-1}} \models B[V]$ for any $B \in Fma_{L^{k-1}}$,

$\mathcal{U}_{L^k} \models \forall v A[V]$ iff for all $\mathbf{x} \in \mathbf{D}$, it is the case that $\mathcal{U}_{L^k} \models A[V](v/\mathbf{x})$ for any $A \in Fma_{L^k}$,

$\mathcal{U}_{L^k} \models (A_1 \rightarrow A_2)[V]$ iff $\mathcal{U}_{L^k} \models A_1[V]$ implies $\mathcal{U}_{L^k} \models A_2[V]$.

The properties *true* and *false* for a formula A in a structure $\mathcal{U}_{L^k}$, or, equivalently, the existence of a *model* for a formula A, are defined as for a structure $\mathcal{U}_L$ in §3.2.

3.5. Axiomatic theory of $Th(o_1^k)$. We shall refer to a sequence $\mathbf{o}_1^k$, $k \geq 1$, which satisfies the input–output relation of a machine $\mathcal{M}$ for some initial condition, as *a sequence generated by the machine* $\mathcal{M}$, without reference to any initial condition for the machine $\mathcal{M}$.

At the instant $k = 1$, the system generates some observed system output $\mathbf{y}_1$ and this is recorded in theory $Th(o_1^1)$ via the axiom $\mathrm{E}q(Y(1), *)$, where $* = y_1(o_1^1)$, the constant symbol in the language L denoting $\mathbf{y}_1$. Similarly, at each successive instant $k \geq 2$, the axiom set of the theory $Th(o_1^k)$ is augmented with the axioms $\mathrm{E}q(U(k-1), *)$, where $* = u_{k-1}(o_1^k)$, and $\mathrm{E}q(Y(k), *)$, where $* = y_k(o_1^k)$, to record the fact that the input $\mathbf{u}_{k-1}$ and output $\mathbf{y}_k$ generated by the system $\mathcal{M}$ have been received as observations. We express this formally by the following set of axioms, which we shall term the observation axioms corresponding to $\mathbf{o}_1^k$.

Observation axiom schemata $(\mathbf{o}_1^k)$.

$$(\mathrm{AXM}^{\mathrm{obs}}(L^k)) \quad \begin{array}{lll} 1.\ \mathrm{E}q(Y(1), *), & \text{where } * = y_1(o_1^k) \in \mathrm{Cons}_L, & k = 1, \\ 2.\ \mathrm{E}q(U(k -_L 1), *), & \text{where } * = u_{k-1}(o_1^k) \in \mathrm{Cons}_L, & k \geq 2, \\ 3.\ \mathrm{E}q(Y(k), *), & \text{where } * = y_k(o_1^k) \in \mathrm{Cons}_L, & k \geq 2. \end{array}$$

State estimation axioms. The following axioms express the recursive formulas (2.6) for the current state estimate sets.

In case $k = 1$:

$$(\mathrm{AXM}^{\mathrm{est}}(L^1)) \quad \begin{array}{c} \mathrm{E}q(\bar{\eta}(x^1), Y(1)) \longleftrightarrow CSE_1(x^1), \\ \vdots \qquad \vdots \qquad \vdots \\ \mathrm{E}q(\bar{\eta}(x^N), Y(1)) \longleftrightarrow CSE_1(x^N). \end{array}$$

In case $k \geq 2$:

$(\mathrm{AXM}^{\mathrm{est}}(L^k))$

$$\begin{array}{c}
\exists x, CSE_{k-1}(x) \wedge \mathrm{E}q(\bar{\Phi}(x, U(k -_L 1)), x^1) \wedge \mathrm{E}q(\bar{\eta}(x^1), Y(k)) \longleftrightarrow CSE_k(x^1), \\
\vdots \\
\exists x, CSE_{k-1}(x) \wedge \mathrm{E}q(\bar{\Phi}(x, U(k -_L 1)), x^N) \wedge \mathrm{E}q(\bar{\eta}(x^N), Y(k)) \longleftrightarrow CSE_k(x^N).
\end{array}$$

We extend Definition 3.1 in the following way.

DEFINITION 3.2. *The* axiom set $\Sigma_k, k \geq 0$, for the finite machine $\mathcal{M}$ and the (input–output) sequence $\mathbf{o}_1^k, k \geq 0$, *is defined as the union of the axiom set* Σ *for the machine* $\mathcal{M}$, *the observation axiom sets indexed by* j, $1 \leq j \leq k$, *corresponding to the sequence* $\mathbf{o}_1^k$, *and the estimation axioms, i.e.,*

$$\Sigma_k = \Sigma \cup \bigcup_{j=1}^{k} \{\mathrm{AMX}^{\mathrm{obs}}(L^j), \mathrm{AMX}^{\mathrm{est}}(L^j)\}, \qquad k \geq 1. \tag{3.1}$$

A structure $\mathcal{U}_{L^k} = (\mathbf{D}, I_k)$ is called a *model* of the theory Th_k if and only if all the axioms Σ_k of Th_k are interpreted true in $\mathcal{U}_{L^k}$, i.e., if and only if I_k is a model for each axiom of Th_k.

We observe that if a language L and a set of axioms Σ_L are given a priori, one may obviously construct a set theoretic machine $\mathcal{M}$ which satisfies Σ_L and for which $|\mathbf{U}| = m, |\mathbf{X}| = N$, and $|\mathbf{Y}| = p$. However, in general, given a priori some axiom system Σ_k corresponding to a symbol sequence $o_1^k, k \geq 1$, it is not the case that such a set theoretic machine $\mathcal{M}$ will form part of a model $\mathcal{U}_{L^k}$ for Σ_k. In fact, it may happen that the only machine in a model $\mathcal{U}_{L^k}$ is the trivial machine with singleton sets of inputs, states, and outputs, respectively.

3.6. Extralogical conditional control rules and theory transitions.

Conditional Control Rules. The following is the general form of a set of CCRs at the instant $k \geq 1$, where C_j is a *conditional control formula* expressed in Fma_{L^k}, which we note contains no appearance of $U(k)$:

$$(\mathrm{CCR}(L^k)) \qquad \begin{array}{lll}
\mathbf{IF}\; C_1 & \mathbf{THEN} & \mathrm{E}q(U(k), u^1), \\
\mathbf{IF}\; \neg C_1 \wedge C_2 & \mathbf{THEN} & \mathrm{E}q(U(k), u^2), \\
\vdots & \vdots & \\
\mathbf{IF}\; \bigwedge_{j=1}^{m-1} \neg C_j \wedge C_m & \mathbf{THEN} & \mathrm{E}q(U(k), u^m), \\
\mathbf{IF}\; \bigwedge_{j=1}^{m} \neg C_j & \mathbf{THEN} & \mathrm{E}q(U(k), u_k^*),
\end{array}$$

where u_k^* is an arbitrary element of $\{u^1, u^2, \ldots, u^p\}$.

The sets of rules $\mathrm{CCR}(L^k), k \geq 1$, are central to the construction of COCOLOG. The function of any given set $\mathrm{CCR}(L^k)$ in the feedback control of $\mathcal{M}$ has the following specification: if the condition C_1 is true (and hence provable) in the decidable theory $Th(o_1^k)$ generated by the axiom set $\Sigma_{k,\mathcal{M}}$ specified in §4, then, invoking the first rule, we obtain the defined constant value u^1 as the value of the control constant $U(k)$; if not, but if C_2 is true (and hence can be proved), then the second axiom gives the defined value u^2 to the control constant $U(k)$; and so on. If none of the conditions $C_1, C_2, \ldots, C_m$ hold at the instant k, then the last rule sets the control constant $U(k)$ equal to the arbitrary constant $u_k^* \in \{u^1, \ldots, u^p\}$.

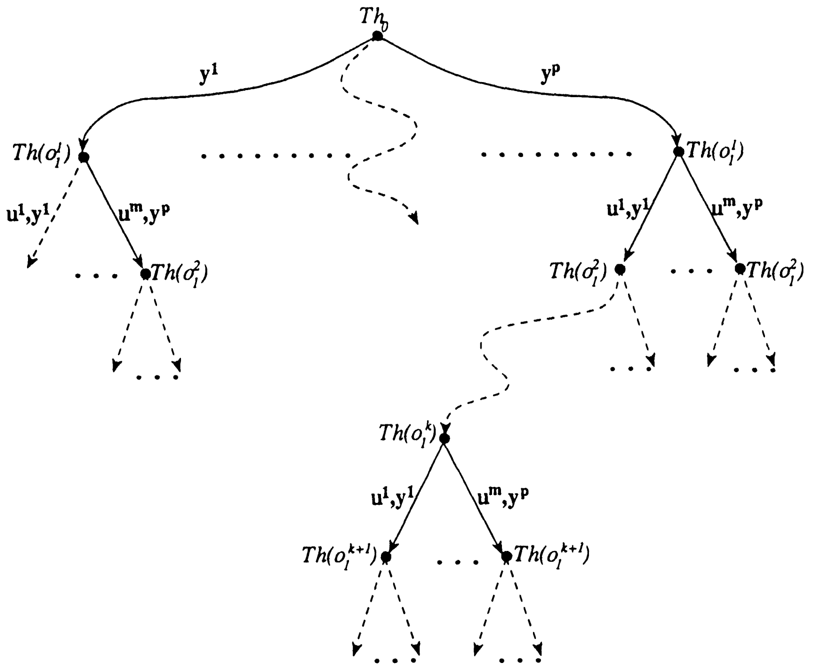


FIG. 2. *A COCOLOG tree of logical theories.*

This procedure uniquely determines the value of $U(k)$. It is proven in §4 that each theory $Th(o_1^k), k \geq 1$, is decidable. Hence an effective procedure to establish the value of $U(k)$ is to run in parallel a set of procedures each of which will effectively decide the truth of a distinct conjunction of control formulas associated with a distinct **IF** statement.

Extralogical theory transitions. When $k \rightarrow k+1$, we make the extralogical step of passing to the theory $Th(o_1^{k+1})$, carrying along all the previous axioms and adding axioms recording the observation of the input $\mathbf{u}_k = \mathbf{u}^\mathbf{i}$ and the resulting output $\mathbf{y}_{k+1} = \mathbf{y}^\mathbf{j}$ for some $\mathbf{u}^\mathbf{i} \in \mathbf{U}, \mathbf{y}^\mathbf{j} \in \mathbf{Y}$ (see Fig. 2). This is formally enforced by the definition of the axiom set Σ_{k+1} generating $Th(o_1^{k+1})$. Hence, in the new theory $Th(o_1^{k+1})$, the observed control action symbol $U(k)$ and the constant symbol u^i determined via $Th(o_1^k)$ satisfy an equality predicate.

Example 3.2. Consider a machine $\mathcal{M} = (\mathbf{U},\mathbf{X},\mathbf{Y},\mathbf{\Phi},\boldsymbol{\eta})$ for which $\mathbf{Y} = \mathbf{X}$ and $\boldsymbol{\eta}$ is the identity map, together with the control objective: if $\mathbf{x}^T$ is reachable from all the current state estimates (i.e., all states satisfying the $CSE_k(\cdot)$ predicate at the current instant k) starting with the application of some common control, then steer towards $\mathbf{x}^T$ and apply some specified control when the current state estimate equals $\mathbf{x}^T$, or else apply $\mathbf{u}^*$.

An example of the operation of the state estimation axioms, $\mathrm{AXM}^{\mathrm{est}}(L^k)$, in this case is the following:

$k = 1$: Let the initial state of the system $\mathcal{M}$ be $\mathbf{x}^\mathbf{j}$; then $\mathbf{y}_1 = \mathbf{y}^\mathbf{j} = \mathbf{x}^\mathbf{j}$. Hence $\mathrm{E}q(Y(1), y^j)$ and it must be the case that $\mathrm{E}q(\bar{\eta}(x^j), y^j)$, hence, by use of $\mathrm{AXM}^{\mathrm{est}}(L^1)$, $CSE_1(x^j)$ is derivable in Th_1.

$k = 2$: Let $\mathbf{u}_1 = \mathbf{u}^\mathbf{s}$ and assume $\mathbf{\Phi}(\mathbf{x}^\mathbf{j}, \mathbf{u}^\mathbf{s}) = \mathbf{x}^l$. Then $\mathbf{y}_2 = \mathbf{y}^l = \mathbf{x}^l$. Hence, $\mathrm{E}q(\bar{\Phi}(x^j, u^s), x^l)$ and $\mathrm{E}q(\bar{\eta}(x^l), y^l)$ are formulas in Th_o. But the axiom set of Th_2 includes the observation axioms

$$\mathrm{E}q(U(1), u^s) \quad \text{and} \quad \mathrm{E}q(Y(2), y^l)$$

in $\mathrm{AXM}^{\mathrm{obs}}(L^2)$. So taking u as u^s, y as y^l, and x as x^j, it follows that the following formula is deducible in Th_2:

$$CSE_1(x^j) \wedge \mathrm{E}q(\bar{\Phi}(x^j, u^s), x^l) \wedge \mathrm{E}q(U(1), u^s) \wedge \mathrm{E}q(\bar{\eta}(x^l), y^l) \wedge \mathrm{E}q(Y(2), y^l),$$

and hence, by use of $\mathrm{AXM}^{\mathrm{eq}}(L^1)(v)$,

$$\exists x \; CSE_1(x) \wedge \mathrm{E}q(\bar{\Phi}(x, U(1)), x^l) \wedge \mathrm{E}q(\bar{\eta}(x^l), Y(2))$$

is derivable in Th_2. So by $\mathrm{AXM}^{\mathrm{est}}(L^1)$

$$CSE_2(x^l) \quad \text{and} \quad \neg CSE_2(x^m), \qquad m \neq l,\ 1 \leq m \leq N,$$

are derivable in Th_2, where the negated predicates are obtained by deducing a negation of a disjunction of the negations of the literals on the left-hand sides of the corresponding lines of $\mathrm{AXM}^{\mathrm{est}}(L^1)$.

For the control problem in this example we may set

$$\begin{aligned}
C_1 &= \forall\, x(\neg CSE_k(x) \vee \mathrm{E}q(x, x^T) \vee [\exists x' \exists l \mathrm{E}q(\bar{\Phi}(x, u^1), x') \wedge \mathit{Rbl}(x', x^T, l)]),\\
C_2 &= \forall\, x(\neg CSE_k(x) \vee \mathrm{E}q(x, x^T) \vee [\exists x' \exists l \mathrm{E}q(\bar{\Phi}(x, u^2), x') \wedge \mathit{Rbl}(x', x^T, l)]),\\
&\vdots\\
C_m &= \forall\, x(\neg CSE_k(x) \vee \mathrm{E}q(x, x^T) \vee [\exists x' \exists l \mathrm{E}q(\bar{\Phi}(x, u^m), x') \wedge \mathit{Rbl}(x' x^T, l)])
\end{aligned}$$

and arrange these conditional control rules into the schemata $\mathrm{CCR}(L^k)$.

If the jth condition C_j is derivable in Th_k, then for any instantiation x^+ of the variable x, either

first, the current state estimate predicate CSE_k does not hold at x^+, or,

second, x^+ is equal to the target state x^T, or,

third, there exists a path of length $l^* + 1$ (where l^* instantiates l, and is greater than or equal to zero) from the current state x^+, via the intermediate state x^* (where x^* instantiates x'), to the target state x^T (since the reachability predicate $\mathit{Rbl}(x^+, x^T, l^*)$ holds), and the control u^j either steers the state x^+ to x^T in one step or is an initial control of a sequence (of length greater than one) which steers the state x^+ to x^T. (Note that the case of a path of length one corresponds to x^* equal to x^T and l equal to 0.)

The corresponding CCR states that C_k holds for no control index k strictly less than j but C_j itself holds. We observe that if the $m + 1$–th control option is used at the instant k by this COCOLOG controller, this indicates that for every control u there is some x^+ that satisfies $CSE_k(\cdot)$ and which is such that, first, x^+ is not equal to x^T and, second, x^T is not reachable from x^+ by the application of a sequence of controls starting with u.

Further inspection of the CCRs shows that they allow the possibility, for certain machines, that the system would maintain itself in a sequence of states from which x^T was reachable in a fixed number of steps l (l greater than or equal to 1) without ever actually converging to x^T. Such apparently perverse behaviour may be prevented by the following elaboration of C_j^* for $1 \leq j \leq m$:

$$\begin{aligned}
C_j^* = \forall\, x\{&\neg CSE_k(x) \vee \mathrm{E}q(x, x^T) \vee [\exists x' \exists\, l \,\forall\, s, \mathrm{E}q(\bar{\Phi}(x, u^j), x') \wedge \mathit{Rbl}(x', x^T, l)\\
&\wedge \{\mathrm{E}q(s, k(N) + 1) \vee \neg \mathrm{E}q(s -_L (l +_L 1), k(N) + 1) \vee \neg \mathit{Rbl}(x, x^T, s)\}]\}.
\end{aligned}$$

This conditional control formula C_j^* states that for all instantiations of the variable x, the conditions stated in C_j above hold and, in addition, x^T is not reachable from each such instantiation of x in strictly less than $l + 1$ steps, where l depends upon x.

4. Consistency, completeness, and decidability of COCOLOG theories. The consistency and completeness of the axiomatic theories presented in §3 can be established by a simple application of classical results on the consistency and completeness of first-order theories with equality; these results depend upon demonstrations of the existence of a model for any given theory.

THEOREM 4.1 (model existence and soundness for Σ). *Consider a finite machine $\mathcal{M}$; then there exists a model $\mathcal{U}_L$ for the axiom set Σ for $\mathcal{M}$, and for any model $\mathcal{U}_L$ of Σ and for any formula $A \epsilon Fma_L$ it is the case that*

$$\Sigma \vdash A \Rightarrow \mathcal{U}_L \models A.$$

Proof. By definition, the axioms in Σ are true formulas in any model $\mathcal{U}_L$ of Σ. It is immediate from the definition of a valuation V of a model $\mathcal{U}_L$ that the rules of inference preserve satisfaction under V and hence truthfulness. It follows that all theorems deducible from the axioms will be true in any model of the axioms.

To establish the theorem it remains to show that there exists at least one model $\mathcal{U}_L$ for the axiom set Σ. We do this by the completely natural choice of the machine $\mathcal{M}$ for $\mathcal{U}_L$ that defined $S(L)$ and the axioms Σ. That is to say, we now choose the particular domain and interpretation function I (and hence structure $\mathcal{U}_L$) given by $\mathbf{D}, I(\bar{\Phi}) = \mathbf{\Phi}, I(\bar{\eta}) = \boldsymbol{\eta}$, where $\mathbf{D}$,$\mathbf{\Phi}$, and $\boldsymbol{\eta}$ are defined in the specification of $\mathcal{M}$. I shall be chosen to map constant names to the constants themselves in $\mathbf{D}$ and to map the reachability predicate to the reachability set relation.

We observe that any structure $\mathcal{U}_L$ of the form specified in §3.2 is a structure for L subject to the restriction that it is *typed*. Hence, if it is verified that $\mathcal{U}_L \models \Sigma$, then a model of Σ has been shown to exist. This is the case for each of the axioms as is shown in the following way:

$\mathrm{AXM}^{\mathrm{dyn}}(L)$: Each of the axioms has the form $\mathrm{E}q(\bar{\Phi}(x^i, u^l), x^j)$, where x^i and x^j lie in $\{x^1, \ldots, x^N\}$ and u^l lies in $\{u^1, \ldots, u^m\}$ and where any such axiom falls in this set if and only if it is the case for the machine $\mathcal{M}$ that $\mathbf{\Phi}(\mathbf{x^i}, \mathbf{u}^l) = \mathbf{x^j}$.

$\mathrm{AXM}^{\mathrm{out}}(L)$: The second subset has the form $\mathrm{E}q(\bar{\eta}(x^i), y^l)$, where x^i lies in $\{x^1, \ldots, x^n\}$ and y^l lies in $\{y^1, \ldots, y^p\}$ and where any such axiom falls in this set if and only if it is the case for $\mathcal{M}$ that $\boldsymbol{\eta}(\mathbf{x^i}) = \mathbf{y}^l$.

Now $\mathcal{U}_L \models \mathrm{E}q(\bar{\Phi}(x^i, u^l), x^j)$ if and only if the equality relation $I(\bar{\Phi}(x^i, u^l)) = I(x^j)$ holds, i.e., if and only if $I(\bar{\Phi})(I(x^i), I(u^l)) = I(x^j)$, i.e., if and only if $\mathbf{\Phi}(\mathbf{x^i}, \mathbf{u}^l) = \mathbf{x^j}$, which is the case since this equality holds for $\mathcal{M}$.

A similar evaluation of $I(\mathrm{E}q(\bar{\eta}(x^i), y^l))$ as the equality relation $I(\bar{\eta}(x^i)) = y^l)$ shows that $\mathcal{U}_L \models \mathrm{E}q(\bar{\eta}(x^i), y^l)$ if and only if $I(\bar{\eta}((x^i)) = I(y^l)$, i.e., if and only if $\boldsymbol{\eta}(\mathbf{x^i}) = \mathbf{y}^l$, which is the case since this equality holds for $\mathcal{M}$. The verification that $\mathcal{U}_L$ is a model for Σ is straightforward for $\mathrm{AXM}^{\mathrm{arith}}(L)$; here we shall only carry out the verification of $\mathcal{U}_L \models \Sigma$ for the reachability axioms.

$\mathrm{AXM}^{Rbl}(L)$: Let us denote the three lines of the reachability axioms by $\mathrm{AXM}^{Rbl}(L)(0)$, $\mathrm{AXM}^{Rbl}(L)(1)$, and $\mathrm{AXM}^{Rbl}(L)(2)$. Then we may proceed case by case.

(i) $\mathcal{U}_L \models \mathrm{AXM}^{Rbl}(L)(0)$.

This holds if and only if for each V

$$\mathcal{U}_L \models \mathrm{E}q(x, x')[V] \text{ implies and is implied by (denoted } \Longleftrightarrow) \\ \mathcal{U}_L \models Rbl(x, x', 0)[V],$$

i.e., if and only if

$$V(x) = V(x') \Longleftrightarrow \mathbf{\Phi}(V(x), \mathbf{u}_1^0) = V(x'), \tag{4.1}$$

but by definition, $\mathbf{\Phi}(\mathbf{x},\mathbf{u}_1^0) = \mathbf{\Phi}(\mathbf{x},\epsilon) = \mathbf{x}$, and so the right-hand equality is true if and only if $V(x) = V(x')$. Consequently the left-hand side of equation implies and is implied by the right-hand side, demonstrating that $\mathrm{AXM}^{Rbl}(L)(0)$ is satisfied by $\mathcal{U}_L$.

(ii) $\mathcal{U}_L \models \mathrm{AXM}^{Rbl}(L)(1)$.

This holds if and only if for each V

$$\mathcal{U}_L \models \exists u, \mathrm{E}q(\bar{\Phi}(x,u),x'))[V] \Longleftrightarrow \mathcal{U}_L \models Rbl(x,x',1)[V],$$

i.e., if and only if

$$\mathcal{U}_L \not\models \forall u \neg \mathrm{E}q(\bar{\Phi}(x,u)x')[V] \Longleftrightarrow (V(x),V(x'),V(1)) \in I(Rbl),$$

i.e., if and only if

$$\begin{aligned}&\{\text{it does not hold that for all } \mathbf{u}' \in \{\mathbf{u}^1,\dots,\mathbf{u}^m\}\\ &\ \mathcal{U}_L \models \neg\mathrm{E}q(\bar{\Phi}(x,u),x')[V(u/\mathbf{u}')]\}\\ &\qquad \Longleftrightarrow \exists \mathbf{u} \in \{\mathbf{u}^1,\dots,\mathbf{u}^m\} \text{ such that } \mathbf{\Phi}(V(x),\mathbf{u}) = V(x').\end{aligned}$$

i.e., if and only if

$$\begin{aligned}&\{\text{it does not hold that for all } \mathbf{u} \in \{\mathbf{u}^1,\dots,\mathbf{u}^m\}\\ &\ \text{it is not the case that } \mathbf{\Phi}(V(x),\mathbf{u}) = V(x')\}\\ &\qquad \Longleftrightarrow \exists \mathbf{u} \in \{\mathbf{u}^1,\dots,\mathbf{u}^m\} \text{ such that } \mathbf{\Phi}(V(x),\mathbf{u}) = V(x'),\end{aligned}$$

i.e., if and only if

$$\begin{aligned}&\{\text{it is the case that for some } \mathbf{u} \in \{\mathbf{u}^1,\dots,\mathbf{u}^m\}\\ &\ \mathbf{\Phi}(V(x),\mathbf{u}) = V(x')\}\\ &\qquad \Longleftrightarrow \exists \mathbf{u} \in \{\mathbf{u}^1,\dots,\mathbf{u}^m\} \text{ such that } \mathbf{\Phi}(V(x),\mathbf{u}) = V(x').\end{aligned}$$

But this is the case, demonstrating that $\mathcal{U}_L \models \mathrm{AXM}^{Rbl}(L)(1)$.

(iii) $\mathcal{U}_L \models \mathrm{AXM}^{Rbl}(L)(2)$.

This holds if and only if $\mathcal{U}_L \models \neg\mathrm{E}q(l, k(N)+1)$ implies

$$\begin{aligned}&\mathcal{U}_L \models \exists x', \exists u, Rbl(x',x'',l) \wedge \mathrm{E}q(\bar{\Phi}(x,u),x')\\ &\qquad \Longleftrightarrow \mathcal{U}_L \models Rbl(x,x'',l +_L 1) \qquad (\text{rule for } \mathcal{U}_L \models A \vee B)\end{aligned}$$

i.e., if and only if for each V, if $V(l) \neq V(k(N)+1)$, then

$$\begin{aligned}&\{\text{for some } \mathbf{d} \in \mathbf{D}, \text{ it is the case that } \mathcal{U}_L \models Rbl(x',x'',l)[V(x'/\mathbf{d})]\\ &\ \text{and for some } \mathbf{c} \in \mathbf{D}, \text{ it is the case that } \mathcal{U}_L \models \mathrm{E}q(\bar{\Phi}(x,u),x')[V(u/\mathbf{c},x'/\mathbf{d})]\}\\ &\qquad \Longleftrightarrow (V(x),V(x''),V(l +_L 1)) \in I(Rbl),\end{aligned}$$

i.e., if and only if for each V, if $\mathbf{l} \neq \mathbf{k}(\mathbf{N}) + \mathbf{1}$, then

$$\begin{aligned}&\{\text{for some } \mathbf{d} \in \mathbf{D},\ (\mathbf{d}, V(x''), V(l)) \in I(Rbl) \text{ and for some } \mathbf{c} \in \mathbf{D},\ \Phi(V(x),\mathbf{c}) = \mathbf{d}\}\\ &\qquad \Longleftrightarrow \{\exists \mathbf{u}_1^{l+1} \in \mathbf{D}^{l+1} \text{ for } 1 \le l \le k(N) \text{ such that } \mathbf{\Phi}(V(x),\mathbf{u}_1^{l+1}) = V(x'')\},\end{aligned}$$

i.e., if and only if for each V, if $\mathbf{l} \neq \mathbf{k}(\mathbf{N}) + \mathbf{1}$, then

$$\begin{aligned}&\{\text{for some } \mathbf{d} \in \mathbf{D}, \exists \mathbf{u}_1^l \in \mathbf{D}^l,\ 1 \le l \le k(N), \text{ such that } \mathbf{\Phi}(\mathbf{d},\mathbf{u}_1^l) = V(x'')\\ &\ \text{and for some } \mathbf{c} \in \mathbf{D},\ \mathbf{\Phi}(V(x),\mathbf{c}) = \mathbf{d}\}\\ &\qquad \Longleftrightarrow \{\exists \mathbf{u}_1^{l+1} \in \mathbf{D}^{l+1} \text{ for } 1 \le l \le k(N) \text{ such that } \mathbf{\Phi}(V(x),\mathbf{u}_1^{l+1}) = V(x'')\}.\end{aligned}$$

But this is the case, demonstrating that $\mathcal{U}_L \models \mathrm{AXM}^{Rbl}(2)$. This completes all three cases for AXM^{Rbl} and hence completes the proof of the theorem. □

THEOREM 4.2 (model existence and soundness for Σ_k). *Consider any input–output sequence $\mathbf{o}_1^k, k \geq 1$, generated by a finite machine $\mathcal{M}$; then there exists a model $\mathcal{U}_{L^k}$ for the axiom set $\Sigma_k, k \geq 1$, for $\mathcal{M}$ and the sequence $\mathbf{o}_1^k$, and for any model $\mathcal{U}_{L^k}$ for Σ_k and any formula $A \,\epsilon Fma_{L^k}$ it is the case that*

$$\Sigma_k \vdash A \Rightarrow \mathcal{U}_{L^k} \models A.$$

Proof. A structure $\mathcal{U}_{L^k}$, defined as an extension of a structure $\mathcal{U}_L$ in §3.4, is specified by the interpretation function I_k mapping (1) the constant symbols $U(j-1), 2 \leq j \leq k$, and $Y(j), 1 \leq j \leq k$, into elements of $\mathbf{U}$ and $\mathbf{Y}$ so as to satisfy the observation axioms (3.5.1); and (2) the consistent state estimate predicates $CSE_j(\cdot), 1 \leq j \leq k$, into subsets of $\mathbf{X}$ so as to satisfy the state estimation axioms (3.5.2) and (3.5.3). The elements of $\mathbf{U}$ and $\mathbf{Y}$ we select so as to satisfy the first condition are those given by the sequence $\mathbf{o}_1^k$, and the subsets of $\mathbf{X}$ we select to satisfy the second condition are those generated by the formula (2.6).

In order to establish the soundness of the first-order rules of inference in $Th(o_1^k)$ we must verify that they preserve satisfaction under any valuation V of any model $\mathcal{U}_{L^k}$ of Σ_k. Since it is evident that satisfaction is preserved by MP and generalization for any given V for any model $\mathcal{U}_{L^k}$, it is sufficient to demonstrate that there exists at least one model $\mathcal{U}_{L^k}$ for the axiom set Σ_k.

Because we have verified the existence of a model for the axioms $\Sigma \subset \Sigma_k$, and because the structure $\mathcal{U}_{L^k}$ contains $\mathcal{U}_L$ as a subset, it is sufficient to verify that a model $\mathcal{U}_{L^k}$ satisfies the axioms $\mathrm{AMX}^{\mathrm{obs}}(L^j), 1 \leq j \leq k$, and $\mathrm{AMX}^{\mathrm{est}}(L^j), 1 \leq j \leq k$. As in Theorem 4.1, we do this by using the natural and obvious choice of the machine $\mathcal{M}$ that defined the axioms Σ and generated the input–output sequence $\mathbf{o}_1^k$.

We begin with the observation axiom schemata for the observation sequences $\mathbf{o}_1^1, \mathbf{o}_1^2, \ldots, \mathbf{o}_1^k$.

$\mathrm{AXM}^{\mathrm{obs}}(L^1)(1)$: For $k = 1, \mathcal{U}_{L^1} \models \mathrm{E}q(Y(1), \star)$, where $\star = y_1(o_1^1)$, if and only if the equality $I_1(Y(1)) = I_1(\star)$ holds, where $\star = y_1(o_1^1)$. But this is the case since $I_1(Y(1)) = \mathbf{y}_1 = I_1(*)$ by the definition of I_1 in §3.4 and the definition of the observation axiom in the axiom set Σ_1 in §3.5.

$\mathrm{AXM}^{\mathrm{obs}}(L^k)(2)$: For $k \geq 2$ we need to verify that, for $2 \leq j \leq k, \mathcal{U}_{L^k} \models \mathrm{E}q(U(j -_L 1), \star)$, where $\star = u_{j-1}(o_1^j)$, if and only if $I_k(U(j -_L 1)) = I_k(\star)$, where $\star = u_{j-1}(o_1^j)$. But this is the case since $I_k(U(j -_L 1)) = I_k(\star)$ by the definition of I_k in §3.4 and the definition of the observation axiom schemata in the axiom system $\Sigma_j, 2 \leq j \leq k$, in §3.5.

$\mathrm{AXM}^{\mathrm{obs}}(L^k)(3)$: Similarly for $2 \leq j \leq k$, we may verify $\mathcal{U}_{L^k} \models \mathrm{E}q(Y(j), \star)$, where $\star = y_j(o_1^j)$, and this shows that the third set of observation axiom schemata are satisfied by $\mathcal{U}_{L^k}$.

Next, we consider the state estimation axioms.

$\mathrm{AXM}^{\mathrm{est}}(L^1)$: Taking $k = 1$ first, we have $\mathcal{U}_{L^1} \models \mathrm{AXM}^{\mathrm{est}}(L^1)(1)$ if and only if, for each valuation

$$\begin{array}{ccc}
\mathcal{U}_{L^1} \models \mathrm{E}q(\bar{\eta}(x^1),\ Y(1)) & \Longleftrightarrow & \mathcal{U}_{L^1} \models CSE_1(x^1), \\
\vdots & \vdots & \vdots \\
\mathcal{U}_{L^1} \models \mathrm{E}q(\bar{\eta}(x^N), Y(1)) & \Longleftrightarrow & \mathcal{U}_{L^1} \models CSE_1(x^N),
\end{array}$$

i.e., if and only if

$$\begin{aligned} I_1(\bar{\eta}(x^1)) &= I_1(Y(1)) \\ &\Longleftrightarrow \mathcal{U}_{L^1} \models CSE_1(x^1), \\ \vdots \qquad & \vdots \qquad \vdots \\ I_1(\bar{\eta}(x^N)) &= I_1(Y(1)) \Longleftrightarrow \\ \mathcal{U}_{L^1} &\models CSE_1(x^N), \end{aligned}$$

i.e., if and only if

$$\begin{aligned} \{\boldsymbol{\eta}(\mathbf{x}^1) = \mathbf{y}_1(\mathbf{o}_1^1)\} &\Longleftrightarrow \mathbf{x}^1 \in \{\mathbf{x}; \mathbf{x} \in \boldsymbol{\eta}^{-1}(\mathbf{y}_1(\mathbf{o}_1^1))\}, \\ \vdots \qquad \vdots \qquad & \qquad\qquad\qquad \vdots \\ \{\boldsymbol{\eta}(\mathbf{x}^{\mathrm{N}}) = \mathbf{y}_1(\mathbf{o}_1^1)\} &\Longleftrightarrow \mathbf{x}^N \in \{\mathbf{x}; \mathbf{x} \in \boldsymbol{\eta}^{-1}(\mathbf{y}_1(\mathbf{o}_1^1))\}, \end{aligned}$$

which is the case.

AXM$^{\mathrm{est}}(L^2)$: We now verify that the model $\mathcal{U}_{L^k}$ satisfies the estimation axioms in the case $k \geq 2$. For $k = 2$, we have $\mathcal{U}_{L^2} \models$ AXM$^{\mathrm{est}}(L^2)$ if and only if for each $\mathcal{U}_{L^2}$ valuation V

$$\begin{aligned} \mathcal{U}_{L^2} &\models (\exists x, CSE_1(x) \wedge \mathrm{E}q(\bar{\Phi}(x, U(1)), x^1) \wedge \mathrm{E}q(\bar{\eta}(x^1), Y(2)))[V] \\ &\Longleftrightarrow \mathcal{U}_{L^2} \models CSE_2(x^1)[V], \\ &\vdots \qquad \vdots \\ \mathcal{U}_{L^2} &\models (\exists x, CSE_1(x) \wedge \mathrm{E}q(\bar{\Phi}(x, U(1)), x^N) \wedge \mathrm{E}q(\bar{\eta}(x^N), Y(2)))[V] \\ &\Longleftrightarrow \mathcal{U}_{L^2} \models CSE_2(x^N)[V], \end{aligned}$$

i.e., if and only if for each V

$$\begin{aligned} \mathcal{U}_{L^2} &\models (\exists x, CSE_1(x) \wedge \mathrm{E}q(\bar{\Phi}(x, U(1)), x^1))[V] \\ &\quad \text{and } \mathcal{U}_{L^2} \models \mathrm{E}q(\bar{\eta}(x^1), Y(2))[V] \\ &\quad \Longleftrightarrow \mathcal{U}_{L^2} \models CSE_2(x^1)[V], \\ &\quad \vdots \quad \vdots \qquad \vdots \quad \vdots \\ \mathcal{U}_{L^2} &\models (\exists x CSE_1(x) \wedge \mathrm{E}q(\bar{\Phi}(x, U(1)), x^N))[V] \\ &\text{and } \mathcal{U}_{L^2} \models \mathrm{E}q(\bar{\eta}(x^N), Y(2))[V] \\ &\Longleftrightarrow \mathcal{U}_{L^2} \models CSE_2(x^N)[V], \end{aligned}$$

i.e., if and only if for each V

$$\begin{aligned} \text{for some } \mathbf{d} \in \mathbf{D},\ \mathcal{U}_{L^2} &\models CSE_1(x) \wedge \mathrm{E}q(\bar{\Phi}(x, U(1)), x^1)[V(x/\mathbf{d})] \\ &\text{and } \mathcal{U}_{L^2} \models \mathrm{E}q(\bar{\eta}(x^1), Y(2))[V] \\ &\Longleftrightarrow \mathcal{U}_{L^2} \models CSE_2(x^1)[V] \\ &\vdots \quad \vdots \quad \vdots \quad \vdots \\ \text{for some } \mathbf{d} \in \mathbf{D},\ \mathcal{U}_{L^2} &\models CSE_1(x) \wedge \mathrm{E}q(\bar{\Phi}(x, U(1)), x^N)[V(x/\mathbf{d})] \\ &\text{and } \mathcal{U}_{L^2} \models \mathrm{E}q(\bar{\eta}(x^N), Y(2))[V] \\ &\Longleftrightarrow \mathcal{U}_{L^2} \models CSE_2(x^N)[V] \end{aligned}$$

if and only if for each V

$$\begin{aligned}
&\text{for some } \mathbf{d} \in \mathbf{D},\ \mathbf{d} \in I_2(CSE_1) \text{ and } \boldsymbol{\Phi}(\mathbf{d},\mathbf{u}_1) = \mathbf{x}^1\\
&\qquad \text{and } \boldsymbol{\eta}(\mathbf{d}) = \mathbf{y}_1\\
&\qquad \Longleftrightarrow \mathbf{x}^1 \in I_2(CSE_2)\\
&\qquad \vdots \quad \vdots \quad \vdots \quad \vdots\\
&\text{for some } \mathbf{d} \in \mathbf{D},\ \mathbf{d} \in I_2(CSE_1) \text{ and } \boldsymbol{\Phi}(\mathbf{d},\mathbf{u}_1) = \mathbf{x}^N\\
&\qquad \text{and } \boldsymbol{\eta}(\mathbf{d}) = \mathbf{y}_1\\
&\qquad \Longleftrightarrow \mathbf{x}^N \in I_2(CSE_2),
\end{aligned}$$

i.e., if and only if

$$\begin{aligned}
&\text{for some } \mathbf{d} \in \mathbf{D}, \boldsymbol{\eta}(\mathbf{d}) = \mathbf{y}_1\\
&\text{and } \boldsymbol{\Phi}(\mathbf{d},\mathbf{u}_1) = \mathbf{x}^1\\
&\text{and } \boldsymbol{\eta}(\mathbf{x}^1) = \mathbf{y}_2\\
&\Longleftrightarrow \mathbf{x}^1 \in \{\hat{\mathbf{x}}_2\}(\mathbf{o}_1^2)\\
&\vdots \quad \vdots \quad \vdots \quad \vdots\\
&\text{for some } \mathbf{d} \in \mathbf{D},\ \boldsymbol{\eta}(\mathbf{d}) = \mathbf{y}_1\\
&\text{and } \boldsymbol{\Phi}(\mathbf{d},\mathbf{u}_1) = \mathbf{x}^N\\
&\text{and } \boldsymbol{\eta}(\mathbf{x}^N) = \mathbf{y}_2\\
&\Longleftrightarrow \mathbf{x}^N \in \{\hat{\mathbf{x}}_2\}(\mathbf{o}_1^2),
\end{aligned}$$

i.e., if and only if

$$\begin{aligned}
&\text{for some } \mathbf{d} \in \mathbf{D},\\
&\boldsymbol{\eta}(\mathbf{d}) = \mathbf{y}_1 \text{ and } \boldsymbol{\Phi}(\mathbf{d},\mathbf{u}_1) = \mathbf{x}^1 \text{ and } \boldsymbol{\eta}(\mathbf{x}^1) = \mathbf{y}_2\\
&\Longleftrightarrow \mathbf{x}^1 \in \{\mathbf{x} \in \mathbf{X}; \boldsymbol{\eta}(\mathbf{x}) = \mathbf{y}_2 \text{ and there exists}\\
&\qquad\qquad \mathbf{x}' \in \mathbf{X} \text{ such that } \mathbf{x} = \boldsymbol{\Phi}(\mathbf{x}', \mathbf{u}_1) \text{ and } \boldsymbol{\eta}(\mathbf{x}') = \mathbf{y}_1\},\\
&\quad \vdots \quad \vdots \quad \vdots \quad \vdots\\
&\text{for some } \mathbf{d} \in \mathbf{D},\\
&\boldsymbol{\eta}(\mathbf{d}) = \mathbf{y}_1 \text{ and } \boldsymbol{\Phi}(\mathbf{d},\mathbf{u}_1) = \mathbf{x}^N \text{ and } \boldsymbol{\eta}(\mathbf{x}^N) = \mathbf{y}_2\\
&\Longleftrightarrow \mathbf{x}^N \in \{\mathbf{x} \in \mathbf{X}; \boldsymbol{\eta}(\mathbf{x}) = \mathbf{y}_2 \text{ and there exists}\\
&\qquad\qquad \mathbf{x}' \in \mathbf{X} \text{ such that } \mathbf{x} = \boldsymbol{\Phi}(\mathbf{x}', \mathbf{u}_1) \text{ and } \boldsymbol{\eta}(\mathbf{x}') = \mathbf{y}_1\},
\end{aligned}$$

which is evidently the case.

To verify that $\mathcal{U}_{L^k} \models \mathrm{AXM}^{\mathrm{est}}(L^k)$ for any $k > 2$ we need to establish that for any $\mathcal{U}_{L^k}$ and valuation V

$$\begin{aligned}
&\mathcal{U}_{L^k} \models (\exists x, CSE_{k-1}(x) \wedge \mathrm{E}q(\bar{\Phi}(x, U(k-1)), x^1) \wedge \mathrm{E}q(\bar{\eta}(x^1), Y(k)))[V]\\
&\qquad \Longleftrightarrow \mathcal{U}_{L^k} \models CSE_k(x^1)[V],\\
&\quad \vdots \quad \vdots \quad \vdots \quad \vdots\\
&\mathcal{U}_{L^k} \models (\exists x, CSE_{k-1}(x) \wedge \mathrm{E}q(\bar{\Phi}(x, U(k-1)), x^N) \wedge \mathrm{E}q(\bar{\eta}(x^N), Y(k)))[V]\\
&\qquad \Longleftrightarrow \mathcal{U}_{L^k} \models CSE_k(x^N)[V].
\end{aligned}$$

Arguing in a parallel manner to the case $k = 2$, the implications above will hold if and only if

$$\begin{aligned}
\{\text{for some } \mathbf{d} \epsilon \mathbf{D} \text{ and some } \mathbf{x}_1^{k-2} &= (\mathbf{x}_1, \dots, \mathbf{x}_{k-2}) \, \epsilon \, \mathbf{D}^{k-2}, \\
\mathbf{u}_1^{k-2} &= (\mathbf{u}_1, \dots, \mathbf{u}_{k-2}) \, \epsilon \, \mathbf{D}^{k-2}, \\
\text{it is the case that} & \\
\mathbf{x}_j &= \mathbf{\Phi}(\mathbf{x}_1, \mathbf{u}_1^{j-1}), j \, \epsilon \, [1, \dots, k-2] \\
\mathbf{y}_j &= \boldsymbol{\eta}(\mathbf{x}_j), j \, \epsilon \, [1, \dots, k-2], \\
\mathbf{d} &= \mathbf{\Phi}(\mathbf{x}_{k-2}, \mathbf{u}_{k-2}), \\
\text{and} & \\
\mathbf{x}^{\mathbf{1}} &= \mathbf{\Phi}(\mathbf{d}, \mathbf{u}_{k-1}) \text{ with } \mathbf{y}_k = \boldsymbol{\eta}(\mathbf{x}^{\mathbf{1}})\} \\
\Longleftrightarrow \mathbf{x}^{\mathbf{1}} &\; \epsilon \; \{\widehat{\mathbf{x}}_k\}(\mathbf{o}_1^k), \\
\vdots \quad \vdots \qquad \vdots &\qquad \vdots \quad \vdots \quad \vdots \\
\{\text{for some } \mathbf{d} \epsilon \mathbf{D} \text{ and some } \mathbf{x}_1^{k-2} &= (\mathbf{x}_1, \dots, \mathbf{x}_{k-2}) \, \epsilon \, \mathbf{D}^{k-2}, \\
\mathbf{u}_1^{k-2} &= (\mathbf{u}_1, \dots, \mathbf{u}_{k-2}) \epsilon \, \mathbf{D}^{k-2}, \\
\text{it is the case that} & \\
\mathbf{x}_j &= \mathbf{\Phi}(\mathbf{x}_1, \mathbf{u}_1^{j-1}), j \, \epsilon \, [1, \dots, k-2], \\
\mathbf{y}_j &= \boldsymbol{\eta}(\mathbf{x}_j), j \, \epsilon \, [1, \dots, k-2], \\
\mathbf{d} &= \mathbf{\Phi}(\mathbf{x}_{k-2}, \mathbf{u}_{k-2}) \\
\text{and} & \\
\mathbf{x}^N &= \mathbf{\Phi}(\mathbf{d}, \mathbf{u}_{k-1}) \text{ with } \mathbf{y}_k = \boldsymbol{\eta}(\mathbf{x}^{\mathbf{N}})\} \\
\Longleftrightarrow \mathbf{x}^{\mathbf{N}} &\; \epsilon \; \{\widehat{\mathbf{x}}_k\}(\mathbf{o}_1^k),
\end{aligned}$$

which is the case. This completes the proof of the theorem. □

THEOREM 4.3 (consistency). *Σ_k is consistent with respect to the first-order theory with equality, i.e., $\Sigma_k \nvdash \bot$ for all $k \geq 0$.*

Proof. This follows in the standard way from the existence of the model $\mathcal{U}_{L^k}$ for the set of axioms Σ_k. Take any formula $A \, \epsilon \, \Sigma_k$. Then we have $\mathcal{U}_{L^k} \not\models \neg A$. By Theorems 4.1 and 4.2, this implies $\Sigma_k \nvdash \neg A$. But by $\mathrm{AXM}^{\log}(1), \vdash \bot \rightarrow (A \rightarrow \bot)$, that is to say $\vdash \bot \rightarrow \neg A$, for any formula A; consequently $\Sigma_k \nvdash \bot$, and hence Σ_k is consistent. □

THEOREM 4.4 (completeness). *Let $\mathbf{o}_1^k, k \geq 0$, be any input–output sequence generated by a finite machine $\mathcal{M}$. Then any formula $A \, \epsilon \, Fma_{L^k}$ is true in every model $\mathcal{V}_{L^k}$ of the axiom set $\Sigma_k, k \geq 0$, for $\mathcal{M}$ and the input–output sequence $\mathbf{o}_1^k$, if and only if A is a consequence of Σ_k, i.e.,*

$$\textit{for all } \mathcal{V}_{L^k}, \; \mathcal{V}_{L^k} \models A \Longleftrightarrow \Sigma_k \vdash A.$$

Proof. We prove only the completeness part here, as soundness follows from Theorems 4.1 and 4.2.

Suppose $\Sigma_k \nvdash A$; then it is sufficient to show there exists a model $\mathcal{V}_{L^k}$ such that $\mathcal{V}_{L^k} \not\models A$. $\Sigma_k \cup \{\neg A\}$ is consistent since Σ_k is consistent and the assumption is that $\Sigma_k \nvdash A$. Now Henkin's theorem (see, e.g., Mendelson [1964]) states that every consistent set of first-order formulas has a model. The standard proof of this theorem applied to the present case gives a model $\mathcal{V}_{L^k}$ for $\Sigma_k \cup \{\neg A\}$. But the model $\mathcal{V}_{L^k}$ is also a model for Σ_k, and clearly $\mathcal{V}_{L^k} \not\models A$, as required. □

We note that in the case above the domain produced by the proof method in the general uninterpreted symbol case is collapsed by the replacements of the classes of cosets of interpretations of the equality and binary arithmetic symbols.

Next, we establish the unique model property for Σ_k, $k \geq 0$, and from this we shall obtain the decidability of COCOLOG theories.

As previously stated, we can get a unique model for a theory Th_o by adding additional axioms to specify the sizes of $\mathbf{U}, \mathbf{X}$, and $\mathbf{Y}$. Otherwise, there can be infinitely many different models. For example, any finite machine $\mathcal{M}' = (\mathbf{U}', \mathbf{X}', \mathbf{Y}', \mathbf{\Phi}', \boldsymbol{\eta}')$ satisfying $\mathbf{U} \subseteq \mathbf{U}'$, $\mathbf{X} \subseteq \mathbf{X}'$, and $\mathbf{Y} \subseteq \mathbf{Y}'$ and such that $\mathbf{\Phi}'$ and $\boldsymbol{\eta}'$ are compatible with $\mathbf{\Phi}$ and $\boldsymbol{\eta}$ on $\mathbf{U}, \mathbf{X}$, and $\mathbf{Y}$ can be used to construct a model of the given machine axioms. This property is analogous to that displayed by nonminimal realizations of input–state–output systems. We conclude that the machine axioms alone cannot uniquely characterize a given finite machine. In fact, even with the addition of further axioms, the most one can achieve by axiomatization is a set of equivalent models up to isomorphism, and consequently uniqueness is used only in this sense.

Suppose $|\mathbf{X}| = N$, $|\mathbf{U}| = m$, and $|Y| = p$. We first consider the state space *size axioms* for $\mathcal{M}$.

Size axioms $\mathbf{X}^{\mathcal{M}}$.

$$
\begin{aligned}
&(1)\quad X_N^{\mathcal{M}} : \neg\mathrm{E}q(x^1, x^2) \wedge \neg\mathrm{E}q(x^1, x^3) \wedge \neg\mathrm{E}q(x^1, x^4) \wedge \cdots \wedge \neg\mathrm{E}q(x^1, x^N) \\
&\qquad\qquad \wedge \neg\mathrm{E}q(x^2, x^3) \wedge \neg\mathrm{E}q(x^2, x^4) \wedge \cdots \wedge \neg\mathrm{E}q(x^2, x^N) \\
&\qquad\qquad \wedge \neg\mathrm{E}q(x^3, x^4) \wedge \cdots \wedge \neg\mathrm{E}q(x^3, x^N), \\
&(\mathrm{AXM}^{\mathrm{size}}(L)) \qquad\qquad \vdots \\
&\qquad\qquad \wedge \neg\mathrm{E}q(x^{N-1}, x^N), \\
&(2)\quad \neg X_{N+1}^{\mathcal{M}} : \forall x \left(\bigvee_{i=1}^{N} \mathrm{E}q(x, x^i) \right).
\end{aligned}
$$

The formula $X_N^{\mathcal{M}}$ expresses the statement that there are at least N distinct constant elements in the state space $\mathbf{X}$ of the finite machine $\mathcal{M}$, i.e., $|\mathbf{X}| \geq N$, and the statement that there are at most N elements in $\mathbf{X}$, i.e., $|\mathbf{X}| \leq N$, is expressed by the formula $\neg X_{N+1}^{\mathcal{M}}$.

Adding $X_N^{\mathcal{M}}$ and $\neg X_{N+1}^{\mathcal{M}}$ to the originally proposed machine axioms ensures that all models of the axioms will have exactly N distinct states.

Size axioms $\mathbf{U}^{\mathcal{M}}$. These are analogous to the size axioms $\mathbf{X}^{\mathcal{M}}$ and specify $|\mathbf{U}| = m$.

Size axioms $\mathbf{Y}^{\mathcal{M}}$. These are analogous to the size axioms $\mathbf{X}^{\mathcal{M}}$ and specify $|\mathbf{Y}| = p$.

In the following we let $\mathcal{M}$ and $\mathcal{M}'$ denote finite machines and $\mathbf{D}$ and $\mathbf{D}'$ denote the sets $\mathbf{D} = \mathbf{U} \cup \mathbf{X} \cup \mathbf{Y}$ and $\mathbf{D}' = \mathbf{U}' \cup \mathbf{X}' \cup \mathbf{Y}'$.

DEFINITION 4.1. *Let $\mathcal{M}$ and $\mathcal{M}'$ be two finite machines in the models $\mathcal{U}_L$ and $\mathcal{U}_L'$, respectively; then a map h from $\mathcal{M}, \mathbf{I}_{\mathbf{k(N)}}$ to $\mathcal{M}', \mathbf{I}'_{\mathbf{k(N)}}$ is a* homomorphism *if for all $\mathbf{u} \in \mathbf{U}, \mathbf{x} \in \mathbf{X}, \mathbf{l}, \mathbf{l}' \in \mathbf{I}_{k(\mathbf{N})}$, it is the case that*

$$
\begin{aligned}
h(\mathbf{\Phi}(\mathbf{x}, \mathbf{u})) &= h(\mathbf{\Phi})(h(\mathbf{x}), h(\mathbf{u})), \\
h(\boldsymbol{\eta}(\mathbf{x})) &= h(\boldsymbol{\eta})(h(\mathbf{x})), \\
h(+_{\mathbf{k(N)}}(\mathbf{l}, \mathbf{l}')) &= h(+_{\mathbf{k(N)}})(h(\mathbf{l}), h(\mathbf{l}')), \\
h(-_{\mathbf{k(N)}}(\mathbf{l}, \mathbf{l}')) &= h(-_{\mathbf{k(N)}})(h(\mathbf{l}), h(\mathbf{l}')).
\end{aligned}
$$

h is an isomorphism *if there exists a homomorphism h' from $\mathbf{D}'$ to $\mathbf{D}$ such that the composition $h' \circ h$ of h' and h is the identity on $\mathbf{D}$.*

If two L-structures $\mathcal{U}_L, \mathcal{U}_L'$ for Th_0 are such that $\mathcal{M}, \mathbf{I}_{\mathbf{k(N)}}$ and $\mathcal{M}', \mathbf{I}'_{\mathbf{k(N)}}$ are isomorphic, we say $\mathcal{U}_L, \mathcal{U}_L'$ have *isomorphic preinterpretations*. If the sets in the products of the domains corresponding respectively to (i) the predicates $Rbl(\cdot, \cdot, \cdot)$ and (ii) the predicates expressing the axiom schemata $\mathrm{AXM}^{\mathrm{obs}}(L^k)$ and $\mathrm{AXM}^{\mathrm{est}}(L^k)$ are also isomorphic, we say the *interpretations*, or *models*, are *isomorphic*. (We note that the standard interpretation for the equality predicate has been taken as a fixed interpretation in all models $\mathcal{U}_L$.) Define

$$\Sigma_{k,\mathcal{M}} = \Sigma_k \cup X_N^{\mathcal{M}} \cup \neg X_{N+1}^{\mathcal{M}} \cup U_m^{\mathcal{M}} \cup \neg U_{m+1}^{\mathcal{M}} \cup Y_p^{\mathcal{M}} \cup \neg Y_{p+1}^{\mathcal{M}}$$

as the set of axioms for the given finite machine $\mathcal{M}$, at the instant $k \geq 0$.

THEOREM 4.5 (unique model property). *The logical theory generated by the axiom set* $\Sigma_{k,\mathcal{M}}, k \geq 0$, *for the finite machine* $\mathcal{M}$ *and the input–output sequence* $\mathbf{o}_1^k, k \geq 1$, *generated by* $\mathcal{M}$ *has a unique model up to isomorphism.*

Remark. In the proof of the following theorem it should be noted that there is an a priori restriction on any model $\mathcal{U}'_{L^k}$ of Σ_k. Recall from §§3.2 and 3.4 that the constant symbol sets

$$\text{Cons}_{L^o} \text{ and } \text{Cons}_{L^k} \triangleq \{u^1, \dots u^m, U(1), U(2), \dots, U(k-1); x^1, \dots, x^N; y^1, \dots, y^p, \\ Y(1), Y(2), \dots, Y(k)\}, \qquad k \geq 1,$$

in the languages $L^k, k \geq 0$, map under any interpretation $I'_k, k \geq 0$, to the corresponding constants in a the set theoretic model domain $\mathbf{D}'$ for some machine. All the axiom systems $\Sigma_0, \Sigma_1, \Sigma_2, \dots$ are axiomatizations of the behaviour (at successive time instants along a system trajectory) of the given set theoretic machine $\mathcal{M}$ defining the language L and the axioms Σ; by definition, this machine $\mathcal{M}$ has the input, state, and output elements $\{\mathbf{u^1}, \dots, \mathbf{u^m}\} = \mathbf{U}$, $\{\mathbf{x^1}, \dots, \mathbf{x^N}\} = \mathbf{X}$, and $\{\mathbf{y^1}, \dots, \mathbf{y^p}\} = \mathbf{Y}$. Consequently, whether or not the size axioms are in force, the images of the elements of Cons_{L^k} under the mapping I'_k of a structure $\mathcal{U}'_{L^k}, k \geq 0$, must be isomorphic to a subset of the union of the sets **U**,**X**, and **Y** appearing in the definition of $\mathcal{M}$. In particular, we shall see that when the size axioms are in force, then under any I'_k the images of $U(1), U(2), \dots; Y(1), Y(2), \dots$ must be in one-to-one correspondence with the occurrences of the inputs and outputs $\mathbf{u}_k, \mathbf{y}_k, k \geq 1$, along the given trajectory of the original machine $\mathcal{M}$ generating these observations; this is the case since the constant names $U(1), U(2), \dots; Y(1), Y(2), \dots$ appear in the observation axiom schemata $\text{AXM}^{\text{obs}}(L^k)$, $k \geq 1$, as the left arguments of equality predicates whose corresponding right arguments have as images under I'_k the elements of the trajectory of the model $\mathcal{U}'_{L^k}, k \geq 1$.

Proof. First we establish that all preinterpretations are isomorphic by showing the existence of a homomorphic mapping between any given pair of models of $\Sigma_0 = \Sigma$.

Now consider the model $\mathcal{U}_L$ constructed in Theorems 4.1 and 4.2 and any other model $\mathcal{U}'_L$. By the size axioms we have $|\mathbf{X}| = |\mathbf{X}'| = N$, $|\mathbf{Y}| = |\mathbf{Y}'| = p$, and $|\mathbf{U}| = |\mathbf{U}'| = m$, and by the machine axioms we have $\mathbf{\Phi} : \mathbf{X} \times \mathbf{U} \to \mathbf{X}$ and $\mathbf{\Phi}' : \mathbf{X}' \times \mathbf{U}' \to \mathbf{X}'$; $\boldsymbol{\eta} : \mathbf{X} \to \mathbf{Y}$ and $\boldsymbol{\eta}' : \mathbf{X}' \to \mathbf{Y}'$.

Let L denote the set of symbols of the theory for the machine $\mathcal{M}$ generated by the axioms $\Sigma_{k,\mathcal{M}}$, and let $I : L \to \mathbf{D}$ and $I' : L \to \mathbf{D}'$ be the interpretation functions corresponding to the domains $\mathbf{D}$ and $\mathbf{D}'$, respectively. Each of the maps may be seen to be bijective by invoking the size axioms and the definition of an interpretation function. Define $h : D \to D'$ such that the following relation is satisfied:

$$h(\mathbf{m}) = I'(I^{-1}(\mathbf{m})) \quad \text{for any } \mathbf{m} \in \mathbf{D}.$$

The relations among L, **D**,**D**$'$ and the mappings of I, I', and h are shown as follows:

$$\begin{tikzcd} & L \arrow[dl, "I"'] \arrow[dr, "I'"] & \\ \mathbf{D} \arrow[rr, "h"] & & \mathbf{D}' \end{tikzcd}$$

It is straightforward to show that h is a bijective mapping: first, to establish h is onto, take any $\mathbf{m}' \epsilon \mathbf{D}'$, then we have $I'^{-1}(\mathbf{m}') = l$ for some $l \epsilon L$ and $I(l) = \mathbf{m}$ for some $\mathbf{m} \epsilon \mathbf{D}$. This $\mathbf{m}$ is obviously the preimage of $\mathbf{m}'$ under h because

$$h(\mathbf{m}) = I'(I^{-1}(\mathbf{m})) = I'(l) = I'(I'^{-1}(\mathbf{m}')) = \mathbf{m}'.$$

Similarly, the one-to-one property is immediately obtained from

$$h(\mathbf{m}_1) = h(\mathbf{m}_2) \quad \text{iff}\, I'(I^{-1}(\mathbf{m}_1)) = I'(I^{-1}(\mathbf{m}_2)) \quad \text{iff}\, \mathbf{m}_1 = \mathbf{m}_2.$$

Now let us write $h(\mathbf{m}) = \mathbf{m}'$ for any $\mathbf{m} \epsilon \mathbf{D}$ and corresponding $\mathbf{m}' \epsilon \mathbf{D}'$ and take some dynamical axiom formula $\mathrm{E}q(\bar{\Phi}(x^i, u^l), x^j)$ from the language L. Since $\mathcal{U}_L$ is a model, the interpretation I will map this formula to $I(\Phi)(I(x^i), I(u^l)) = I(x^j)$, which is the equality $\mathbf{\Phi}(\mathbf{x^i}, \mathbf{u}^l) = \mathbf{x^j}$, and the interpretation I' will map the formula to $I'(\Phi)(I'(x^i), I'(u^l)) = I'(x^j)$, which is $\mathbf{\Phi}'(\mathbf{x}'^{\mathbf{i}}, \mathbf{u}'^l) = \mathbf{x}'^{\mathbf{j}}$. Then since $h(\mathbf{m}) = I'(I^{-1}(\mathbf{m}))$ we have the following relationship between the two models:

$$\mathbf{\Phi}(\mathbf{x^i}, \mathbf{u}^l) = \mathbf{x}^j \quad \text{iff}\, \mathbf{\Phi}'(\mathbf{x}'^{\mathbf{i}}, \mathbf{u}'^l) = \mathbf{x}'^{\mathbf{j}}.$$

Similarly, reference to the output axioms $\mathrm{AXM}^{\mathrm{out}}(L)$ yields

$$\boldsymbol{\eta}(\mathbf{x^i}) = \mathbf{y^m} \quad \text{iff}\, \boldsymbol{\eta}'(\mathbf{x}'^{\mathbf{i}}) = \mathbf{y}'^{\mathbf{m}}.$$

For the finite arithmetic of Th_0 things are yet simpler since the properties of $+_L, -_L$ were defined in terms of the fixed interpretation of $+_{\mathbf{k(N)}}, -_{\mathbf{k(N)}}$ on the fixed interpretation of a fragment $\mathbf{I}_{\mathbf{k(N)}}$ of the integers. Hence

$$+_{\mathbf{k(N)}}(\mathbf{l},\mathbf{l}') = \mathbf{l}'' \quad \text{iff } +'_{\mathbf{k(N)}}(\mathbf{l},\mathbf{l}') = \mathbf{l}'',$$
$$-_{\mathbf{k(N)}}(\mathbf{l},\mathbf{l}') = \mathbf{l}'' \quad \text{iff } -'_{\mathbf{k(N)}}(\mathbf{l},\mathbf{l}') = \mathbf{l}''.$$

Consequently, defining $h(\Phi)$ to be the functional relation from $\mathbf{X}' \times \mathbf{U}'$ to $\mathbf{X}'$ given by $\mathbf{\Phi}'$ and $h(\eta)$ to be that from $\mathbf{X}'$ to $\mathbf{Y}'$ given by $\boldsymbol{\eta}'$, it follows that $\Sigma_{k,\mathcal{M}}$ has a unique preinterpretation up to isomorphism.

For the interpretations of the predicate $Rbl(x, x', k)$, the interpretation functions I and I' give, respectively, the relational sets $\{(\mathbf{x}, \mathbf{x}', \mathbf{k}); \mathbf{x}, \mathbf{x}' \epsilon, \mathbf{X}, \mathbf{k} \epsilon \mathbf{I}_{\mathbf{k(N)}}$, such that there exists $\mathbf{u}_1^k \epsilon \mathbf{U}^k$ and $k \epsilon \mathbf{I}_{\mathbf{k(N)}}$ such that $\mathbf{\Phi}(\mathbf{x}, \mathbf{u}_1^k) = \mathbf{x}'\}$, and the same expression with the substitutions $\mathbf{X}'$ for $\mathbf{X}, \mathbf{U}'$ for $\mathbf{U}$, and $\mathbf{I}'_{\mathbf{k(N)}}$ for $\mathbf{I}_{\mathbf{k(N)}}$, etc. But it follows from the isomorphism of the preinterpretations that the corresponding relational sets are isomorphic, as required.

To establish the isomorphic nature of the interpretations of the observation and estimation axioms we proceed as follows.

For the first observation axiom schemata, in the case $k = 1$, we have $\mathrm{E}q(Y(1), \star)$, where $\star = y_1(o_1^k) \epsilon \mathrm{Cons}_{L^k}$.

Let us consider the particular case where $\mathbf{y}_1(\mathbf{o}_1^1) = \mathbf{y}^1 \epsilon \mathbf{Y}$. In that case we must treat the particular axiom for Th_1 given by $\mathrm{E}q(Y(1), y^1)$.

Under I and I', respectively, this formula maps to the relations

$$I(\mathrm{E}q(Y(1), y^1)), \quad \text{i.e., } \{I(Y(1)) = \mathbf{y}^1\}$$

and

$$I'(\mathrm{E}q(Y(1), y^1)), \quad \text{i.e., } \{I'(Y(1)) = \mathbf{y}'^1\}$$

for some $\mathbf{y}'^1 \epsilon \mathbf{Y}'$. But then the isomorphism $h = I' \circ I^{-1}$ gives the isomorphism of the relational sets in $\mathcal{U}_{L^1}$ and $\mathcal{U}'_{L^1}$ corresponding to the single observation axiom schemata in this case.

Moreover, the same assertion holds in each of the alternative cases

$$\mathbf{y}_1(\mathbf{o}_1^1) = \mathbf{y}^{\mathbf{j}} \epsilon \mathbf{Y}, \qquad \mathbf{2} \leq \mathbf{j} \leq \mathbf{p}.$$

Similar arguments hold for the observation axiom schemata (2) and (3) in the case $k \geq 2$.

It follows that the isomorphism $I'(I^{-1}(\cdot))$ maps the images of the constant names $\{Y(1), Y(2), \ldots, Y(k); U(1), U(2), \ldots, U(k-1)\}$ in the model $\mathcal{U}_{L^k}$ bijectively to the images of these constant names in $\mathcal{U}'_{L^k}$ and that the relational sets corresponding to the observation axiom schemata are also isomorphic.

Finally, we see that the relations given by the images of the state estimation axioms $\mathrm{AXM}^{\mathrm{est}}(L^k), k \geq 1$, under an interpretation I are isomorphic to their images under I' by the isomorphism $I' \circ I^{-1}$. This is the case since all of the set elements, function mappings, and set relations given by the interpretation of these axioms have been shown to be isomorphic to the image of these axioms under I' via the isomorphism $I' \circ I^{-1}$. □

DEFINITION 4.2 (proper formula). *A closed formula P is a* proper formula *with respect to a set of formulas Γ if P contains neither any constant, variable, predicate symbols nor function symbols which do not appear in any formulas in Γ.*

DEFINITION 4.3 (complete axiomatization). *A* set of formulas Γ *is said to be* complete *if either P or $\neg P$ is a consequence of Γ for any proper formula P with respect to Γ, and an* axiomatization Γ *is said to be* complete *if the set of formulas Γ is complete. A* theory *Th is said to be* complete *if it is consistent and for every closed formula A either $Th \vdash A$ or $Th \vdash \neg A$.*

THEOREM 4.6. *The axiomatization defined by $\Sigma_{k,\mathcal{M}}, k \geq 0$, for the finite machine $\mathcal{M}$ and the input–output sequence $\mathbf{o}_1^k, k \geq 0$, generated by $\mathcal{M}$, is a complete axiomatization.*

Proof. To prove that $\Sigma_{k,\mathcal{M}}$ is a complete axiomatization, we need to show that for any formula $A \epsilon L^k$ either $\Sigma_{k,\mathcal{M}} \vdash A$ or $\Sigma_{k,\mathcal{M}} \vdash \neg A$. We know $\Sigma_{k,\mathcal{M}}$ is consistent by the existence of a model for $\Sigma_{k,\mathcal{M}}$. By Lindenbaum's lemma (see Mendelson [1964]) if $\Sigma_{k,\mathcal{M}}$ is a consistent first-order theory, then there is a consistent complete extension of $\Sigma_{k,\mathcal{M}}$. But since $\Sigma_{k,\mathcal{M}}$ has a unique model, the complete extension of $\Sigma_{k,\mathcal{M}}$ is $\Sigma_{k,\mathcal{M}}$. Hence $\Sigma_{k,\mathcal{M}}$ is complete. □

THEOREM 4.7 (decidable theoremhood). *The logical theory Th_k generated by $\Sigma_{k,\mathcal{M}}$ for the finite machine $\mathcal{M}$ and the input–output sequence $\mathbf{o}_1^k, k \geq 0$, generated by $\mathcal{M}$, is recursively decidable.*

Proof. It is known that any axiomatizable complete theory is recursively decidable (Bell and Machover [1977, Thm. 11.5, p. 355]). Now the theory Th_k generated by $\Sigma_{k,\mathcal{M}}$ is (finitely) axiomatizable since it is generated by the set of postulates $\Sigma_{k,\mathcal{M}}$. Furthermore, Theorem 4.6 gives the completeness of the theory Th_k generated by $\Sigma_{k,\mathcal{M}}$; hence the result follows. □

5. Extralogical transitions between logical theories. We recall that a COCOLOG control system operates via a cyclical interaction between a controlled dynamical system, represented by the mathematical system $\mathcal{M}$, and a logic controller carrying the sequence of COCOLOG theories $\{Th(o_1^k), k \geq 0\}$.

At any instant $k \geq 0$, $\Sigma_{k,\mathcal{M}}$ contains axioms specifying the observed system input $\mathbf{u}_{k-1}$ and output $\mathbf{y}_k$. The deductive closure $Th(o_1^k)$ of $\Sigma_{k,\mathcal{M}}$ is then generated and the controller selects the control input $\mathbf{u}_k$ by the metalogical step of finding the unique **IF**. . .**THEN**. . . element of the list CCR_k for which the antecedent formula lies in $Th(o_1^k)$.

The equality predicate specifying (in the precise sense given in §3.5) the input $\mathbf{u}_k$ to $\mathcal{M}$ and the equality predicate specifying the resulting observed output $\mathbf{y}_{k+1}$ of $\mathcal{M}$ are then adjoined as new axioms to $\Sigma_{k,\mathcal{M}}$, along with the estimation axioms corresponding to the instant $k+1$, to form the axiom set $\Sigma_{k+1,\mathcal{M}}$ in the extended language $L_{k+1,\mathcal{M}}$.

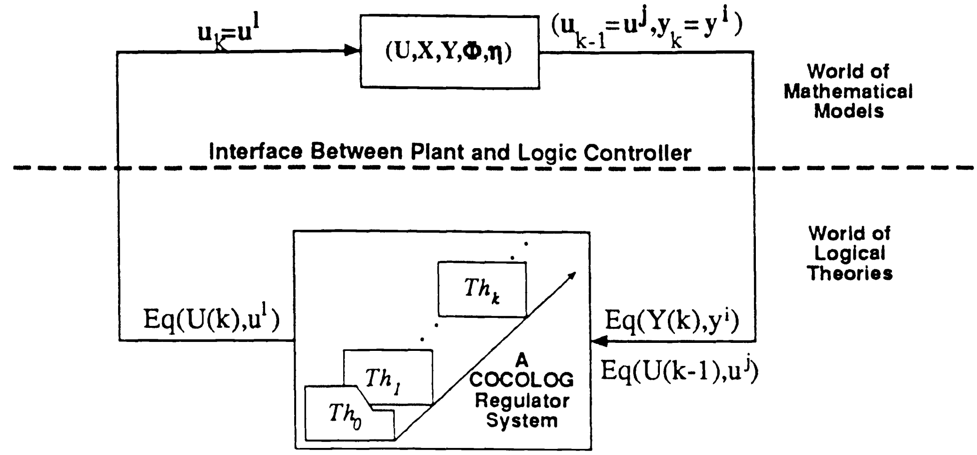

u_k=u^l
(U,X,Y,Φ,η)
(u_{k-1}=u^j,y_k=y^i)
World of Mathematical Models
Interface Between Plant and Logic Controller
World of Logical Theories
Eq(U(k),u^l)
Th_k
Th_l
Th_0
A COCOLOG Regulator System
Eq(Y(k),y^i)
Eq(U(k-1),u^j)

FIG. 3. *A closed-loop logic control system.*

With respect to the modelling of the relation between the evolution of the sequence of COCOLOG theories and the controlled system $\mathcal{M}$, it has been assumed that there are no errors in the observations or in the communication of control inputs and that the control inputs generated by the logic controller are implemented instantaneously with respect to the discrete time indices of the system $\mathcal{M}$. Hence there is no inconsistency between the values of system inputs and outputs as given by the axiom sets $\Sigma_{k,\mathcal{M}}$ and the evolution of the model $\mathcal{M}$. Furthermore, it has been assumed that the entire deductive closure $Th(o_1^k)$ of the axioms $\Sigma_{k,\mathcal{M}}$ is instantaneously generated in the logic regulator $\mathcal{R}$ at each discrete time instant k, thus permitting the instantaneous selection of the appropriate conditional control rule determining the subsequent control input.

These requirements constitute restrictions on the transitions between logical theories which cannot be represented within these theories themselves and which have consequently been expressed in the description of the overall system.

In this setting we evidently obtain the following statement concerning the evolution of a sequence of COCOLOG theories.

THEOREM 5.1 (nesting of theories). *Let* $\mathcal{M} = (\mathbf{U}, \mathbf{X}, \mathbf{Y}, \mathbf{\Phi}, \eta)$ *be a finite machine in an initial state* $\mathbf{x}_0$, *and let* $\{\mathrm{CCR}_k, k \geq 1\}$ *be a sequence of* CCRs *formulated in the sequence of languages* $\{L^k, k \geq 1\}$. *Assume that the input sequence to the machine* $\mathcal{M}$ *is such that at the instant k the input* $\mathbf{u}_k \in \mathbf{U}$ *is that determined by the associated* CCR_k. *Then* (*up to isomorphism*) *the observation axioms* $\{\mathrm{AXM}^{\mathrm{obs}}(L^k), k \geq 1\}$ *of the COCOLOG theories* $\{Th(o_1^k), k \geq 1\}$ *generated by* $\{\Sigma_{k,\mathcal{M}}, k \geq 1\}$ *satisfy the interpretation constraint expressed by*

$$I_k(U(k-1)) = \mathbf{u}_{k-1}(\mathbf{o}_1^k), \qquad I_k(Y(k)) = \mathbf{y}_k(\mathbf{o}_1^k),$$

and the sequence of COCOLOG theories $\{Th(o_1^k), k \geq 0\}$ *generated by* $\{\Sigma_{k,\mathcal{M}}, k \geq 0\}$ *satisfies*

$$Th_0 \subseteq Th(o_1^1) \subseteq \cdots \subseteq Th(o_1^k) \subseteq Th(o_1^{k+1}) \subseteq \cdots, \qquad k \geq 0.$$

A representation of a COCOLOG closed-loop control system is displayed in Fig. 3.

6. Conclusion. Several problems concerning COCOLOG systems may be posed at this point.

(i) There is the problem of the definition of a tractable, analysable fragment of COCOLOG obtained by suitably restricting the COCOLOG languages $\{L_k; k \geq 0\}$, the associated axioms, or the class of CCRs. This has been undertaken in Wei and Caines [1992], where the notion of the Markovian fragments of a COCOLOG system is defined and analysed.

(ii) The issue of implementability of COCOLOG for real-time systems leads to the question of automatic theorem proving in COCOLOG. As remarked in the introduction, current experiments using the FE-resolution extension of the GTP automatic theorem-proving software of Newborn [1987] (proposed by the authors and developed by Q.-X. Yu) are encouraging (see Wang and Caines [1991], [1992]), as are the results in Caines, Mackling, and Wei [1992], which employs the Blitzensturm theorem-proving software of Mackling [1993]. A complexity analysis of such algorithms would be of great value.

(iii) A realization of a COCOLOG system is a sequence of first-order theories generated by a given sequence of input–output observations corresponding to a path in a COCOLOG tree structure (see Fig. 2). The true formulas at the nodes of this tree can be captured by a possible-worlds interpretation of a model logic (see Goldblatt [1987]), and a study of the mathematical properties of such an overall modal logic formulation of COCOLOG merits attention.

Acknowledgments. The authors gratefully acknowledge conversations concerning this paper with Tom Mackling, Michael Makkai, and Yuan-Jun Wei and the valuable comments of anonymous referees; they also wish to thank David Delchamps for suggesting the name COCOLOG.

REFERENCES

J. Bell and M. Machover (1977), *A Course in Mathematical Logic*, North-Holland, Amsterdam.

P. E. Caines and S. Wang (1989a), *Classical and logic-based regulators for partially observed automata: Dynamic programming formulation*, in Proc. 1989 Conference on Information Sciences and Systems, Johns Hopkins University, Baltimore, MD, March.

——— (1989b), *Classical and logic-based regulator design and its complexity*, in Proc. 28th IEEE Conference on Decision and Control, Tampa, FL, December, pp. 132–137.

——— (1990), *COCOLOG: A conditional controller and observer logic for finite machines*, in Proc. 29th IEEE Conference on Decision and Control, Hawaii, pp. 2845–2850.

P. E. Caines, R. Greiner, and S. Wang (1991), *Classical and logic-based dynamic observers for finite automata*, IMA J. Math. Control Inform., 8, pp. 45–80.

P. E. Caines, T. Mackling, and Y.-J. Wei (1993), *Logical control via automatic theorem proving: COCOLOG fragments implemented in Blitzensturm* 5.0, in Proc. American Control Conference, San Francisco, pp. 1209–1213.

P. E. Caines, S. Wang, and R. Greiner (1988), *Dynamical logic observers for finite automata*, in Proc. 1988 Conference on Information Sciences and Systems, Princeton University, Princeton, NJ, March, pp. 50–56.

D. Dyck and P. E. Caines (1995), *The logical control of an elevator*, IEEE Trans. Automat. Control, 40, pp. 480–486.

R. Goldblatt (1987), *Logics of Time and Computation*, CSLI/Stanford, Stanford, CA.

C. Green (1969), *Application of theorem proving to problem solving*, in Proc. First IJCAI, Washington, D.C., Morgan Kaufmann, Los Altos, CA, pp. 219–239.

D. Harrel (1979), *First Order Dynamic Logic*, Springer-Verlag, Berlin.

W. Kohn (1988), *A declarative theory for rational controllers*, in Proc. 27th IEEE CDC, Vol. 1, Austin, TX, December 7–9, pp. 131–136.

——— (1991), *Declarative control*, Communications of the ACM, 34, pp. 65–79.

T. Mackling (1993), *The Equality Predicate and Subsumption in Resolution-Based Automatic Theorem Proving*, Research report, Dept. of Electrical Engineering, McGill University, Montreal, July.

J. McCarthy and P. Hayes (1969), *Some philosophical problems from the standpoint of artificial intelligence*, in Machine Intelligence 4, B. Meltzer and D. Michie, eds., pp. 463–502.

E. Mendelson (1964), *Introduction to Mathematical Logic*, Van Nostrand Reinhold, New York.

M. Newborn (1989), *The Great Theorem Prover*, Newborn Software, P.O. Box 429, Victoria Station, Westmount, PQ, Canada.

J. S. Ostroff (1987), *Real-Time Computer Control of Discrete Systems Modeled by Extended Machines: A Temporal Logic Approach*, Ph.D. thesis, University of Toronto, January.

——— (1989), *Real Time Temporal Logic*, John Wiley, New York.

J. Ostroff and M. Wonham (1985), *A temporal logic approach to real time control*, in Proc. 24th IEEE Conference on Decision and Control, Fort Lauderdale, FL, December, pp. 656–657.

P. Ramadge and W. M. Wonham (1987), *Supervisory control of a class of discrete-event processes*, SIAM J. Control Optim., 25, pp. 206-230.

——— (1989), *The control of discrete event systems*, in Proc. IEEE, 77, pp. 81–98.

R. Reiter (1991), *The frame problem in the situation calculus: A simple solution (sometimes) and a completeness result for goal regression*, in Artificial Intelligence and Mathematical Theory of Computation: Papers in Honor of John McCarthy, Vladimir Lifschitz, ed., Academic Press, San Diego, CA, pp. 359–380.

S. J. Rosenschein and L. P. Kaelbling (1987), *The Synthesis of Digital Machines with Provable Epistemic Properties*, Technical note 412, SRI International, Palo Alto, CA, April.

J. G. Thistle and W. M. Wonham (1986), *Control problems in a temporal logic framework*, Internat. J. Control, 44, pp. 943–976.

S. Wang (1991), *Classical and Logic Based Control Theory for Finite State Machines*, Ph.D. thesis, McGill University, Montreal, October.

S. Wang and P. E. Caines (1991), *Automated reasoning with function evaluation for COCOLOG*, in Control Theory, Stochastic Analysis and Applications, S. P. Chen and J. M. Yong, eds., World Scientific, Singapore, pp. 59–78.

——— (1992), *On a conditional observer and controller logic (COCOLOG) for machines and its automated reasoning methodology*, Recent Advances in Mathematical Theory of Systems, Control, Networks and Signal Processing II, Proceedings of the International Symposium on the Mathematical Theory of Networks and Systems MTNS-91, Kobe, H. Kimura and S. Kodama, eds., Kobe, Mita Press, Osaka, Japan, pp. 49–54.

Y.-J. Wei and P. E. Caines (1992), *On markovian fragments of COCOLOG for logic control systems*, Proc. 31st IEEE Conference on Decision and Control, Tucson, AZ, December, pp. 2967–2972, SIAM J. Control Optim.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 33, No. 6, pp. 1716–1730, November 1995

© 1995 Society for Industrial and Applied Mathematics

005

REMARKS ON NONLINEAR STOCHASTIC PARTIAL DIFFERENTIAL EQUATIONS: AN APPLICATION OF THE SPLITTING-UP METHOD*

NORIAKI NAGASE†

Abstract. The objective of this article is to apply the splitting-up method to the existence theorem for nonlinear stochastic partial differential equations. With use of this method an approximating sequence is constructed. By the compactness argument a convergent subsequence can be extracted, and this fact provides the solution.

Key words. splitting-up method, stochastic partial differential equation

AMS subject classifications. 60H15, 35R60

1. Introduction.

We consider in this article the existence of solutions for the following nonlinear stochastic partial differential equation derived by white noise:

$$dy(t) + A(t)y(t)\,dt = f(t, y(t))\,dt + g(t, y(t))\,dW(t), \tag{1.1}$$

where $A(t)$ is a second-order elliptic differential operator, $f(t,\cdot)$ and $g(t,\cdot)$ are continuous operators from $L^2(\mathbb{R}^d)$ to itself, and $W(t)$ is an m-dimensional Brownian motion. A solution $y(t)$ of the problem is sought in the space of Sobolev type $H^1(\mathbb{R}^d)$. (For the precise definition, see Definition 2.1.)

When $f(t,\cdot)$ and $g(t,\cdot)$ satisfy the Lipschitz condition, Pardoux [4] and Walsh [5] proved the existence and uniqueness of the solutions for (1.1) by Picard's method of successive approximation. But if $f(t,\cdot)$ and $g(t,\cdot)$ are merely continuous, Picard's method is not effective.

Recently Bensoussan [1] has given a new result for stochastic partial differential equations for Leray–Lions operators on a compact subset of $\mathbb{R}^d$. The main idea is to use the splitting-up method, considering $A(t)y(t)\,dt - f(t, y(t))\,dt - g(t, y(t))\,dW(t)$ as the sum of two operators. According to Bensoussan's idea, we shall show the existence of solutions for the nonlinear stochastic partial differential equation (1.1) on whole space $\mathbb{R}^d$. (See also [2].)

This paper is formulated as follows. In §2 we state our problem, and in §3 we state the main theorem. In §4 we construct an approximating sequence of the equation (1.1) by the splitting-up method and show some estimates for this sequence. In §5 we show the fundamental lemma which is the whole-space $\mathbb{R}^d$ version of Bensoussan's compactness results. By this lemma, we can apply an argument similar to that of Bensoussan [1]. Section 6 is devoted to the proof of the main theorem.

2. Setting of the problem.

2.1. Notations and assumptions.

We denote by L^2_γ, $\gamma \geq 0$, the space of real-valued Borel functions on $\mathbb{R}^d$ with the norm defined by

$$|u|_\gamma = \left(\int_{\mathbb{R}^d} |(1 + |x|^2)^{\gamma/2} u(x)|^2\,dx \right)^{1/2}.$$

Clearly L^2_γ becomes a Hilbert space with the inner product

$$(u, v)_\gamma = \int_{\mathbb{R}^d} (1 + |x|^2)^\gamma u(x)v(x)\,dx.$$

* Received by the editors March 26, 1993; accepted for publication (in revised form) July 14, 1994.

† Department of Mathematics, Faculty of Science, Hirosaki University, 3 Bunkyo-cho, Hirosaki, Aomori-ken 036, Japan.

Let us set $L^2_0 = L^2$, $|\cdot|_0 = |\cdot|$, $(\cdot\,,\cdot)_0 = (\cdot\,,\cdot)$ ($\gamma = 0$), for simplicity, if no confusion occurs. $H^1 = H^1(\mathbb{R}^d)$ denotes the Sobolev space $W^{1,2}(\mathbb{R}^d)$ with its norm $\|\cdot\|$.

We shall consider second-order uniformly elliptic operators of the form

$$\begin{aligned} A(t)u(x) = &-\sum_{i,j=1}^{d} \frac{\partial}{\partial x_i}\left(a_{ij}(t,x)\frac{\partial}{\partial x_j}u(x)\right) \\ &+\sum_{i=1}^{d} b_i(t,x)\frac{\partial}{\partial x_i}u(x) + c(t,x)u(x), \end{aligned} \tag{2.1}$$

where a_{ij}, $b_i (i,j = 1,\dots,d)$, and c are bounded Borel functions on $(0,T)\times\mathbb{R}^d$ satisfying the conditions

$$\begin{aligned} &\text{i)}\ \ a_{ij}(t,x) = a_{ji}(t,x), \qquad i,j = 1,\dots,d, \\ &\text{ii)}\ \ \sum_{i,j=1}^{d} a_{ij}(t,x)\xi_i\xi_j \ge \alpha|\xi|^2 \quad \forall\xi = (\xi_1,\dots,\xi_d)\in\mathbb{R}^d, \end{aligned} \tag{A.1}$$

where $\alpha > 0$.

It is easy to check that the operator $A(t)$ satisfies the property

$$2\langle A(t)u,u\rangle + \lambda|u|^2 \ge \alpha\|u\|^2 \quad \forall u\in H^1 \tag{2.2}$$

for some $\lambda\in\mathbb{R}$, where $\langle\cdot\,,\cdot\rangle$ denotes the duality pairing between H^{-1} (the dual space of H^1 under identifying L^2 with its dual) and H^1.

Next we shall consider maps $f(t,u)$, $g(t,u) = (g_1(t,u),\dots,g_m(t,u))$ such that

$$\begin{aligned} &\text{i)}\ \ f:(0,T)\times L^2\to L^2 \text{ measurable,} \\ &\text{ii)}\ \ \text{a.e. } t,\ f(t,\cdot): L^2\to L^2 \text{ continuous,} \\ &\text{iii)}\ \ |f(t,u)|^2 \le K(1+|u|^2); \end{aligned} \tag{A.2}$$

$$\begin{aligned} &\text{i)}\ \ g_j:(0,T)\times L^2\to L^2 \text{ continuous,} \\ &\text{ii)}\ \ |g(t,u)|^2 = \sum_{j=1}^{m}|g_j(t,u)|^2 \le K(1+|u|^2), \\ &\text{iii)}\ \ |g_j(t,u)-g_j(s,u)|^2 \le o(|t-s|)(1+|u|^2). \end{aligned} \tag{A.3}$$

With $o(h)$ monotonic increasing, $o(h)\to 0$ as $h\to 0$.

2.2. The problem. We consider the following equation:

$$\begin{cases} dy(t) + A(t)y(t)\,dt = f(t,y(t))\,dt + \displaystyle\sum_{j=1}^{m} g_j(t,y(t))\,dW_j(t), \\ y(0) = y_0, \end{cases} \tag{2.3}$$

where $W(t) = (W_1(t),\dots,W_m(t))$ is an m-dimensional Brownian motion.

DEFINITION 2.1. *By a solution of the equation* (2.3), *we mean an H^1-valued process $y = (y(t))$ defined on a probability space $(\Omega,\mathcal{F},P)$ with a reference family $(\mathcal{F}_t)$ such that*

(I) *there exists an m-dimensional $(\mathcal{F}_t)$-Brownian motion $W(t) = (W_1(t),\dots,W_m(t))$ with $W(0) = 0$;*

(II) $y = (y(t))$ *is adapted to* $(\mathcal{F}_t)$ *and*

$$E\left[\int_0^T \|y(t)\|^2 \, dt\right] < \infty;$$

(III) *for any* $\varphi \in C_0^\infty(\mathbb{R}^d)$ (C^∞*-function on* $\mathbb{R}^d$ *with compact support*) *and almost all* $t \in [0,T]$

$$\begin{aligned}(y(t),\varphi) &+ \int_0^t \langle A(s)y(s),\varphi\rangle \, ds \\ &= (y_0,\varphi) + \int_0^t (f(s,y(s)),\varphi)\, ds + \sum_{j=1}^m \int_0^t (g_j(s,y(s)),\varphi)\, dW_j(s)\end{aligned} \tag{2.4}$$

holds.

To emphasize the particular role of $(\mathcal{F}_t)$-Brownian motion $W(t)$, sometimes we call the pair (W,y) itself a solution of (2.3).

3. Existence of the solution. Besides (A.1)–(A.3), we assume the following conditions.

(A.4) For some $\gamma > 0$, the restrictions of $f(t,\cdot)$ and $g_j(t,\cdot)$ on L^2_γ operate to itself and satisfy the linear growth condition. Namely, there exists a constant K such that

$$|h(t,u)|^2_\gamma \le K(1+|u|^2_\gamma) \quad \forall u \in L^2_\gamma, \tag{3.1}$$

where $h = f$, g_j $(j = 1,\dots,m)$.

(A.5) $y_0 \in L^2_\gamma$, where γ is the same number as in (A.4).

THEOREM 3.1. *Under the assumptions* (A.1)–(A.5), *the equation* (2.3) *has a solution in the sense of Definition* 2.1.

4. The splitting-up approximation scheme.

4.1. The algorithm. Let k be a positive integer which will tend to $+\infty$, and set $\delta = \frac{T}{k+1}$.

For each $w(\cdot) = (w_1(\cdot),\dots,w_m(\cdot)) \in C(0,T;\mathbb{R}^m)$, we shall define a process $z_k(t)$ depending on k and $w(\cdot)$. Consider an interval $[r\delta,(r+1)\delta)$, $r = 0,1,\dots,k$; then z_k is defined on this interval by the relation

$$\begin{cases} \dfrac{dz_k(t)}{dt} + A(t)z_k(t) = 0, \\ z_k(r\delta) = z_k^r, \end{cases} \tag{4.1}$$

where

$$\begin{aligned} z_k^{r+1} &= z_k((r+1)\delta - 0) + \int_{r\delta}^{(r+1)\delta} f(t,z_k(t))\, dt \\ &\quad + g(r\delta,\bar{z}_k^r)\cdot(w((r+1)\delta) - w(r\delta)), \\ z_k^0 &= y_0, \end{aligned}$$

and

$$\bar{z}_k^r = \frac{1}{\delta}\int_{r\delta}^{(r+1)\delta} z_k(t)\, dt.$$

By (2.2), (4.1) has a unique solution in $L^2(r\delta,(r+1)\delta;H^1)\cap C(r\delta,(r+1)\delta;L^2)$ once $z_k^r(\in L^2)$ is given. So (4.1) completely defines z_k on $[0,T)$. We set for completeness $z_k(T)=z_k^{k+1}$ and note that z_k is discontinuous at points $r\delta$ $(r=1,\dots,k+1)$ and has left-hand-side limits.

Hence we can define a map $\Psi_k : C(0,T;\mathbb{R}^m)\to L^2(0,T;H^1)$ by

$$\Psi_k(w)=z_k=\text{the solution of (4.1) corresponding to } w\in C(0,T;\mathbb{R}^m).$$

It is easy to see that Ψ_k is a continuous map from $C(0,T;\mathbb{R}^m)$ to $L^2(0,T;H^1)$, where $C(0,T;\mathbb{R}^m)$ is equipped with the uniform topology and $L^2(0,T;H^1)$ with the strong topology.

Let $B=(B_1,\dots,B_m)$ be an m-dimensional standard Brownian motion defined on a probability space $(\Omega,\mathcal{F},P)$, and put $z_k=\Psi_k(B)$.

4.2. Estimations. According to Bensoussan [1], we shall prove the following a priori estimates.

LEMMA 4.1. *There is a constant $C>0$ such that*

$$E\left[\int_0^T \|z_k(t)\|^2\,dt\right]\le C, \tag{4.2}$$

$$\sup_{0\le t\le T} E[|z_k(t)|^2]\le C, \tag{4.3}$$

$$\sup_{0\le t\le T} E[|z_k(t)|^4]\le C \tag{4.4}$$

for any $k=1,2,\dots$.

Remark 4.1. Hereafter we denote by C a constant independent of $k=1,2,\dots$ and $r=0,1,\dots,k+1$.

Proof. From the energy equality related to (4.1), we get

$$|z_k(t)|^2+2\int_{r\delta}^t\langle A(s)z_k(s),z_k(s)\rangle\,ds=|z_k^r|^2,\qquad t\in[r\delta,(r+1)\delta). \tag{4.5}$$

Using (2.2), we have

$$|z_k(t)|^2+\alpha\int_{r\delta}^t\|z_k(s)\|^2\,ds\le|z_k^r|^2+|\lambda|\int_{r\delta}^t|z_k(s)|^2\,ds. \tag{4.6}$$

Gronwall's inequality derives

$$|z_k(t)|^2\le e^{|\lambda|(t-r\delta)}\,|z_k^r|^2,\qquad t\in[r\delta,(r+1)\delta). \tag{4.7}$$

It follows that

$$\alpha\int_{r\delta}^{(r+1)\delta}\|z_k(t)\|^2\,dt\le e^{|\lambda|\delta}\,|z_k^r|^2 \tag{4.8}$$

and

$$|\bar{z}_k^r|^2\le\frac{e^{|\lambda|\delta}-1}{|\lambda|\delta}\,|z_k^r|^2. \tag{4.9}$$

Define the process $\tilde{z}_k$ by

$$\begin{aligned}\tilde{z}_k(t) = z_k((r+1)\delta - 0) &+ \int_{r\delta}^{t} f(s, z_k(s))\,ds \\ &+ g(r\delta, \bar{z}_k^r)\cdot(B(t) - B(r\delta))\end{aligned} \tag{4.10}$$

for $t \in [r\delta, (r+1)\delta)$, $r = 0, 1, \dots, k$.

It follows from Itô's formula that

$$\begin{aligned}d|\tilde{z}_k(t)|^2 &= 2(f(t, z_k(t)), \tilde{z}_k(t))\,dt \\ &\quad + 2(g(r\delta, \bar{z}_k^r), \tilde{z}_k(t))\,dB(t) + |g(r\delta, \bar{z}_k^r)|^2\,dt\end{aligned} \tag{4.11}$$

for $t \in [r\delta,\ (r+1)\delta)$, $r = 0, 1, \dots, k$.

Using (A.2), (A.3), (4.7), and (4.9), we get

$$\begin{aligned}E|\tilde{z}_k(t)|^2 &= E|z_k((r+1)\delta - 0)|^2 \\ &\quad + 2E\left[\int_{r\delta}^{t} (f(s, z_k(s)), \tilde{z}_k(s))\,ds\right] + E\left[\int_{r\delta}^{t} |g(r\delta, \bar{z}_k^r)|^2\,dt\right] \\ &\le e^{|\lambda|\delta} E|z_k^r|^2 + C\{1 + E|z_k^r|^2\}(t - r\delta) + E\left[\int_{r\delta}^{t} |\tilde{z}_k(s)|^2\,ds\right].\end{aligned} \tag{4.12}$$

So Gronwall's inequality derives

$$\begin{aligned}E|z_k^{r+1}|^2 &= E|\tilde{z}_k((r+1)\delta - 0)|^2 \\ &\le e^{(|\lambda|+1)\delta} E|z_k^r|^2 + C\{1 + E|z_k^r|^2\}(e^{\delta} - 1) \\ &\le (C\delta + 1)E|z_k^r|^2 + C\delta,\end{aligned} \tag{4.13}$$

which yields

$$\begin{aligned}E|z_k^r|^2 &\le (1 + C\delta)^r\{|y_0|^2 + 1\} \\ &\le \left(1 + \frac{CT}{k+1}\right)^{k+1}\{|y_0|^2 + 1\}\end{aligned} \tag{4.14}$$

for any $r = 0, 1, \dots, k+1$. Hence we obtain

$$E|z_k^r|^2 \le C \tag{4.15}$$

for any $k = 1, 2, \dots$ and $r = 0, 1, \dots, k+1$.

Combining (4.15) with (4.7) and (4.8), we get (4.2) and (4.3). For the proof (4.4), we note that from (4.7)

$$|z_k(t)|^4 \le e^{2|\lambda|(t - r\delta)}\,|z_k^r|^4, \qquad t \in [r\delta, (r+1)\delta). \tag{4.16}$$

Applying Itô's formula to (4.11), we get

$$\begin{aligned}&d|\tilde{z}_k(t)|^4 \\ &\quad = 2|\tilde{z}_k(t)|^2\{2(f(t, z_k(t)), \tilde{z}_k(t)) + |g(r\delta, \bar{z}_k^r)|^2\}\,dt \\ &\quad\quad + 4|\tilde{z}_k(t)|^2(g(r\delta, \bar{z}_k^r), \tilde{z}_k(t))\,dB(t) \\ &\quad\quad + 4|(g(r\delta, \bar{z}_k^r), \tilde{z}_k(t))|^2\,dt\end{aligned} \tag{4.17}$$

for $t \in [r\delta, (r+1)\delta)$, $r = 0, 1, \dots, k$.

By the same argument as above, we get

$$E|z_k^{r+1}|^4 = E|\tilde{z}_k((r+1)\delta - 0)|^4 \le (C\delta + 1)E|z_k^r|^4 + C\delta, \tag{4.18}$$

which yields

$$E|z_k^r|^4 \le C \tag{4.19}$$

for any $k = 1, 2, \dots$ and $r = 0, 1, \dots, k+1$.

This implies (4.4).

Remark 4.2. Combining (4.12) and (4.15), we can easily see that

$$\sup_{0 \le t \le T} E|\tilde{z}_k(t)|^2 \le C \quad \forall\, k = 1, 2, \dots. \tag{4.20}$$

Similarly it follows from (4.17) and (4.19) that

$$\sup_{0 \le t \le T} E|\tilde{z}_k(t)|^4 \le C \quad \forall\, k = 1, 2, \dots. \tag{4.21}$$

LEMMA 4.2. *There is a constant $C > 0$ such that*

$$E\left[\sup_{0 \le t \le T} |z_k(t)|^2\right] \le C \tag{4.22}$$

for any $k = 1, 2, \dots$.

Proof. Using (4.7), we have

$$\begin{aligned}\sup_{0 \le t \le T} |z_k(t)|^2 &\le e^{|\lambda|\delta} \max_{0 \le r \le k+1} |z_k^r|^2 \\ &\le e^{|\lambda|T} \max_{0 \le r \le k+1} |z_k^r|^2.\end{aligned} \tag{4.23}$$

Consider next the process $\tilde{z}_k$ defined in the proof of Lemma 4.1. Combining (4.6) with (4.11), we get

$$\begin{aligned}|z_k^{r+1}|^2 \le |z_k^r|^2 &+ |\lambda| \int_{r\delta}^{(r+1)\delta} |z_k(t)|^2\, dt \\ &+ 2\int_{r\delta}^{(r+1)\delta} (f(t, z_k(t)), \tilde{z}_k(t))\, dt \\ &+ 2\int_{r\delta}^{(r+1)\delta} (g(r\delta, \bar{z}_k^r), \tilde{z}_k(t))\, dB(t) \\ &+ \int_{r\delta}^{(r+1)\delta} |g(r\delta, \bar{z}_k^r)|^2\, dt.\end{aligned} \tag{4.24}$$

Adding these relations, we deduce

$$\begin{aligned}|z_k^r|^2 &\le |y_0|^2 + |\lambda| \int_0^{r\delta} |z_k(t)|^2\, dt + 2\int_0^{r\delta} (f(t, z_k(t)), \tilde{z}_k(t))\, dt \\ &\quad + 2\int_0^{r\delta} (\bar{g}(t), \tilde{z}_k(t))\, dB(t) + \int_0^{r\delta} |\bar{g}(t)|^2\, dt \\ &\le |y_0|^2 + |\lambda| \int_0^T |z_k(t)|^2\, dt + \int_0^T |f(t, z_k(t))|^2\, dt + \int_0^T |\tilde{z}_k(t)|^2\, dt \\ &\quad + \int_0^T |\bar{g}(t)|^2\, dt + 2\int_0^{r\delta} (\bar{g}(t), \tilde{z}_k(t))\, dB(t),\end{aligned} \tag{4.25}$$

where $\bar{g}(t) = g(r\delta, \bar{z}_k^r)$, $t \in [r\delta, (r+1)\delta)$, $r = 0, 1, \ldots, k$.

From (4.3) and (4.20), we deduce

$$\begin{aligned} E\left[\max_{0\le r\le k+1} |z_k^r|^2\right] \\ \le C + 2E\left[\sup_{0\le t\le T}\left|\int_0^t (\bar{g}(s), \tilde{z}_k(s))\, dB(s)\right|\right]. \end{aligned} \tag{4.26}$$

Using the Burkholder–Gundy inequality, we have

$$\begin{aligned} &E\left[\sup_{0\le t\le T}\left|\int_0^t (\bar{g}(s), \tilde{z}_k(s))\, dB(s)\right|\right] \\ &\le CE\left[\left\{\int_0^T (\bar{g}(s), \tilde{z}_k(s))^2\, ds\right\}^{1/2}\right] \\ &\le C\left\{E\left[\int_0^T |\tilde{z}_k(t)|^2\, dt\right] + E\left[\sum_{r=0}^k |\bar{z}_k^r|^2 \int_{r\delta}^{(r+1)\delta} |\tilde{z}_k(t)|^2\, dt\right]\right\}^{1/2} \\ &\le C\left\{E\left[\int_0^T |\tilde{z}_k(t)|^2\, dt\right]\right. \\ &\quad \left. + E\left[\sum_{r=0}^k \frac{1}{\delta}\int_{r\delta}^{(r+1)\delta} |z_k(t)|^2\, dt \int_{r\delta}^{(r+1)\delta} |\tilde{z}_k(t)|^2\, dt\right]\right\}^{1/2} \\ &\le C\left\{E\left[\int_0^T |\tilde{z}_k(t)|^2\, dt\right] + E\left[\int_0^T |z_k(t)|^4\, dt\right] + E\left[\int_0^T |\tilde{z}_k(t)|^4\, dt\right]\right\}^{1/2} \\ &\le C \quad \text{by (4.4), (4.20), and (4.21).} \end{aligned} \tag{4.27}$$

From (4.23), the desired result (4.22) follows.

LEMMA 4.3. *There is a constant $C > 0$ such that*

$$E\left[\sup_{|\theta|\le\eta}\int_0^T \|z_k(t+\theta) - z_k(t)\|_*^2\, dt\right] \le C\eta \tag{4.28}$$

for any $0 < \eta \le 1$, $k = 1, 2, \ldots$, where z_k is extended by 0 outside of $[0, T]$ and $\|\cdot\|_*$ is the norm of the dual space $H^{-1} = (H^1)^*$.

Proof. Assume $\theta > 0$. A similar calculation is done whenever $\theta < 0$. We write

$$I = E\left[\sup_{0\le\theta\le\eta}\int_0^T \|z_k(t+\theta) - z_k(t)\|_*^2\, dt\right] \le I_1 + I_2, \tag{4.29}$$

where

$$I_1 = E\left[\sup_{0\le\theta\le\eta}\int_0^{T-\eta} \|z_k(t+\theta) - z_k(t)\|_*^2\, dt\right]$$

and

$$I_2 = E\left[\sup_{0\le\theta\le\eta}\int_{T-\eta}^T \|z_k(t+\theta) - z_k(t)\|_*^2\, dt\right].$$

From (4.22), we can see

$$I_2 \le C\eta. \tag{4.30}$$

Now we deal with I_1. Using (4.1) and (4.10), we have

$$\begin{aligned} &z_k(t+\theta) - z_k(t) + \int_t^{t+\theta} A(s) z_k(s) ds \\ &\quad = \int_{[t/\delta]\delta}^{[(t+\theta)/\delta]\delta} f(s, z_k(s))\, ds + \int_{[t/\delta]\delta}^{[(t+\theta)/\delta]\delta} \bar{g}(s)\, dB(s) \end{aligned} \tag{4.31}$$

for $t \in [0, T-\eta]$ and $0 \le \theta \le \eta$, where $\bar{g}$ is the process defined in (4.25) and $[t]$ denotes the integral part of t.

Next we have

$$\begin{aligned} \left\| \int_t^{t+\theta} A(s) z_k(s)\, ds \right\|_* &\le \int_t^{t+\theta} \|A(s) z_k(s)\|_*\, ds \\ &\le \theta^{1/2} \left\{ \int_t^{t+\theta} \|A(s) z_k(s)\|_*^2\, ds \right\}^{1/2}. \end{aligned} \tag{4.32}$$

Using (4.2), we get

$$\begin{aligned} &E\left[\sup_{0 \le \theta \le \eta} \int_0^{T-\eta} \left\| \int_t^{t+\theta} A(s) z_k(s)\, ds \right\|_*^2 dt \right] \\ &\quad \le \eta E\left[\int_0^{T-\eta} \int_0^T \left\| A(s) z_k(s) \right\|_*^2 ds\, dt \right] \\ &\quad \le C\eta. \end{aligned} \tag{4.33}$$

For the first term of the right-hand side of (4.31), we have

$$\begin{aligned} &\sup_{0 \le \theta \le \eta} \left\| \int_{[t/\delta]\delta}^{[(t+\theta)/\delta]\delta} f(s, z_k(s))\, ds \right\|_*^2 . \\ &\quad \le C \left\{ \int_{[t/\delta]\delta}^{[(t+\eta)/\delta]\delta} (1 + |z_k(s)|)\, ds \right\}^2 . \end{aligned} \tag{4.34}$$

Using (4.22), we get

$$\begin{aligned} &E\left[\sup_{0 \le \theta \le \eta} \int_0^{T-\eta} \left\| \int_{[t/\delta]\delta}^{[(t+\theta)/\delta]\delta} f(s, z_k(s))\, ds \right\|_*^2 dt \right] \\ &\quad \le CE\left[\sup_{0 \le s \le T} (1 + |z_k(s)|)^2 \int_0^{T-\eta} \left(\left[\frac{t+\eta}{\delta}\right] \delta - \left[\frac{t}{\delta}\right] \delta \right)^2 dt \right] \\ &\quad \le C \int_0^{T-\eta} \left(\left[\frac{t+\eta}{\delta}\right] \delta - \left[\frac{t}{\delta}\right] \delta \right)^2 dt \\ &\quad \le C\eta. \end{aligned} \tag{4.35}$$

Finally, using the Burkholder–Gundy inequality and (4.3), we have

$$
\begin{aligned}
&E\left[\sup_{0\le\theta\le\eta}\int_0^{T-\eta}\left\|\int_{[t/\delta]\delta}^{[(t+\theta)/\delta]\delta}\bar{g}(s)\,dB(s)\right\|_*^2\,dt\right]\\
&\le\int_0^{T-\eta}E\left[\sup_{0\le\theta\le\eta}\left|\int_{[t/\delta]\delta}^{[(t+\theta)/\delta]\delta}\bar{g}(s)\,dB(s)\right|^2\right]dt\\
&\le C\int_0^{T-\eta}E\left[\int_{[t/\delta]\delta}^{[(t+\eta)/\delta]\delta}|\bar{g}(s)|^2\,ds\right]dt\\
&\le C\int_0^{T-\eta}\left(\left[\frac{t+\eta}{\delta}\right]\delta-\left[\frac{t}{\delta}\right]\delta\right)dt\\
&\le C\eta.
\end{aligned}\tag{4.36}
$$

Combining (4.33), (4.35), and (4.36) with (4.31), we obtain $I_1 \le C\eta$, and this completes the proof.

LEMMA 4.4. *There is a constant $C > 0$ such that*

$$\sup_{0\le t\le T} E|z_k(t)|_\gamma^2 \le C \tag{4.37}$$

for any $k = 1, 2, \dots$, where γ is the same number as in (A.4).

Proof. We define the operator $\bar{A}(t)$ by

$$
\begin{aligned}
\bar{A}(t)u(x) = &-\sum_{i,j=1}^d \frac{\partial}{\partial x_i}\left(a_{ij}(t,x)\frac{\partial}{\partial x_j}u(x)\right)\\
&+\sum_{j=1}^d \bar{b}_{0j}(t,x)\frac{\partial}{\partial x_j}u(x)+\sum_{i=1}^d\frac{\partial}{\partial x_i}(\bar{b}_{i0}(t,x)u(x))\\
&+\bar{c}(t,x)u(x),
\end{aligned}\tag{4.38}
$$

where

$$
\begin{aligned}
\bar{b}_{0j}(t,x) &= b_j(t,x)+\sum_{i=1}^d a_{ij}(t,x)R_i(x),\\
\bar{b}_{i0}(t,x) &= \sum_{j=1}^d a_{ij}(t,x)R_j(x),\\
\bar{c}(t,x) &= c(t,x)-\sum_{i,j=1}^d a_{ij}(t,x)R_i(x)R_j(x)\\
&\quad-\sum_{i=1}^d b_i(t,x)R_i(x),
\end{aligned}
$$

and

$$R_i(x)=\frac{\gamma x_i}{1+|x|^2}\qquad (i=1,\dots,d).$$

Then $\langle\bar{A}(t)(Ru),\varphi\rangle = \langle A(t)u, R\varphi\rangle$ for any $u \in H^1$ and $\varphi \in C_0^\infty(\mathbb{R}^d)$ where $R(x) = (1+|x|^2)^{\gamma/2}$. Hence $q_k(t) = Rz_k(t)$ satisfies the following equation:

$$
\begin{cases}
\dfrac{dq_k(t)}{dt}+\bar{A}(t)q_k(t)=0,\\
q_k(r\delta)=q_k^r, \qquad t\in[r\delta,(r+1)\delta),
\end{cases}\tag{4.39}
$$

where

$$q_k^{r+1} = q_k((r+1)\delta - 0) + \int_{r\delta}^{(r+1)\delta} Rf(t, z_k(t))\,dt$$
$$+Rg(r\delta, \bar{z}_k^r) \cdot (B((r+1)\delta) - B(r\delta))$$

and

$$q_k^0 = Ry_0.$$

By virtue of assumptions (A.4) and (A.5), we can repeat an argument similar to (4.5) – (4.15) and obtain

$$\sup_{0 \le t \le T} E|q_k(t)|^2 \le C. \tag{4.40}$$

This yields (4.37) and completes the proof.

5. Tightness property. Now we introduce the fundamental lemma which is the whole-space $\mathbb{R}^d$ version of Bensoussan's compactness result (see Bensoussan [1, Prop. 3.1]). We consider a subset Z of $L^2(0,T;H^1(\mathbb{R}^d))$ depending on three constants, K, L, and M, and two sequences, μ_n and ν_n, with μ_n, $\nu_n > 0$ and μ_n, $\nu_n \to 0$ as $n \to 0$. The set Z is defined as follows:

$$Z = \left\{ u \in L^2(0,T;H^1(\mathbb{R}^d));\ \int_0^T \|u(t)\|^2\,dt \le K,\ \int_0^T |u(t)|_\gamma^2\,dt \le L, \right.$$
$$\left. \text{and} \sup_{|\theta| \le \mu_n} \int_0^T \|u(t+\theta) - u(t)\|_*^2\,dt \le \nu_n M \quad \forall n \ge 1 \right\}, \tag{5.1}$$

where u is extended by 0 outside of $[0,T]$.

LEMMA 5.1. *The set Z is a compact subset of $L^2(0,T;L^2(\mathbb{R}^d))$.*

Proof. Let $\{u_k\}$ be a sequence in Z. By the first condition of Z, we can extract a subsequence, still denoted by $\{u_k\}$, such that

$$u_k \to u \quad \text{in } L^2(0,T;H^1) \text{ weakly.} \tag{5.2}$$

Clearly the limit u satisfies the first and third conditions of Z. Put

$$D_p = \{x \in \mathbb{R}^d;\ |x| < p\}, \qquad p = 1, 2, \ldots.$$

First we shall prove that

$$u_k \to u \quad \text{in } L^2((0,T) \times D_p) \text{ strongly.} \tag{5.3}$$

Since the injection $H^1(D_p) (= W^{1,2}(D_p))$ into $L^2(D_p)$ is compact, for any $\varepsilon > 0$ there is a constant $C(\varepsilon)$ such that

$$|\varphi|_{D_p}^2 \le \varepsilon \|\varphi\|_{D_p}^2 + C(\varepsilon)\|\varphi\|_*^2 \quad \forall \varphi \in H^1(D_p), \tag{5.4}$$

where $|\cdot|_{D_p}$ and $\|\cdot\|_{D_p}$ are the norm of $L^2(D_p)$ and $H^1(D_p)$, respectively. (Apply Proposition 4.1 of Lions [3, p. 59] to the triplet $(H^1(D_p), L^2(D_p), (H^1(\mathbb{R}^d))^*)$.)

Hence

$$\begin{aligned}\int_0^T &|u_k(t)-u(t)|_{D_p}^2\,dt\\ &\le \varepsilon\int_0^T \|u_k(t)-u(t)\|_{D_p}^2\,dt + C(\varepsilon)\int_0^T \|u_k(t)-u(t)\|_*^2\,dt\\ &\le 2K\varepsilon + C(\varepsilon)\int_0^T \|u_k(t)-u(t)\|_*^2\,dt.\end{aligned}\tag{5.5}$$

Therefore, to prove (5.3), it is sufficient to prove that

$$\int_0^T \|u_k(t)-u(t)\|_*^2\,dt \to 0.\tag{5.6}$$

Consider a function $\psi \in C_0^\infty(\mathrm{R})$ with $\psi \ge 0$, $\int_{-\infty}^{\infty}\psi(t)\,dt = 1$, $\mathrm{supp}(\psi) = [-1,1]$, and the mollifier

$$\begin{aligned}\psi_\varepsilon * u(t) &= \frac{1}{\varepsilon}\int_{-\infty}^{\infty}\psi\left(\frac{t-s}{\varepsilon}\right)u(s)ds\\ &= \int_{-1}^{1} u(t-\varepsilon s)\psi(s)\,ds.\end{aligned}$$

Put $\tilde{u}_k = u_k - u$. Since u_k and u satisfy the third condition of Z, we see

$$\begin{aligned}\int_0^T &\|\psi_{\mu_n} * \tilde{u}_k(t) - \tilde{u}_k(t)\|_*^2\,dt\\ &\le \int_{-1}^{1}\left\{\int_0^T \|\tilde{u}_k(t-\mu_n s) - \tilde{u}_k(t)\|_*^2\,dt\right\}\psi(s)\,ds\\ &\le 4\nu_n M.\end{aligned}\tag{5.7}$$

Now, from (5.2), we have

$$\psi_{\mu_n} * \tilde{u}_k(t) = \frac{1}{\mu_n}\int_0^T \psi\left(\frac{t-s}{\mu_n}\right)\tilde{u}_k(s)\,ds \to 0 \quad \text{as}\, k\to\infty\tag{5.8}$$

in $H^1(D_p)$ weakly for any $n \ge 1$, $t \in [0,T]$.

Recalling that for the triplet $(H^1(D_p), L^2(D_p), (H^1(\mathrm{R}^d))^*)$ the injection $H^1(D_p)$ into $L^2(D_p)$ is compact, we have

$$\psi_{\mu_n} * \tilde{u}_k(t) \to 0 \quad \text{as}\, k\to\infty \quad \text{in}\,(H^1(\mathrm{R}^d))^* \quad \text{strongly.}\tag{5.9}$$

Since

$$\|\psi_{\mu_n} * \tilde{u}_k(t)\|_*^2 \le C(n)\int_0^T \|\tilde{u}_k(t)\|^2\,dt \le 4C(n)K,$$

we get

$$\psi_{\mu_n} * \tilde{u}_k \to 0 \quad \text{as}\, k\to\infty \quad \text{in}\, L^2(0,T;(H^1(\mathrm{R}^d))^*) \quad \text{strongly.}\tag{5.10}$$

(5.7) and (5.10) imply (5.6). Hence we obtain (5.3). From (5.3), it is easy to see that u satisfies the second condition of Z and thus $u \in Z$.

Finally

$$\begin{aligned}
&\int_0^T |u_k(t) - u(t)|^2\,dt \\
&\quad \le \int_0^T |u_k(t) - u(t)|^2_{D_p}\,dt + \frac{1}{(1+p^2)^\gamma}\int_0^T |u_k(t) - u(t)|^2_\gamma\,dt \\
&\quad \le \int_0^T |u_k(t) - u(t)|^2_{D_p}\,dt + \frac{4L}{(1+p^2)^\gamma}.
\end{aligned} \tag{5.11}$$

Therefore we deduce from (5.3) that

$$u_k \to u \quad \text{in } L^2(0,T;L^2(\mathbb{R}^d)) \quad \text{strongly,} \tag{5.12}$$

and this completes the proof.

Put $S = C(0,T;\mathbb{R}^m) \times L^2(0,T;L^2(\mathbb{R}^d))$. We denote by π_k the image measure of $(B, z_k) = (B, \Psi_k(B))$ on S. By virtue of the above compactness result, we can prove the following tightness property.

LEMMA 5.2. *The family* $\{\pi_k\}_{k\ge 1}$ *is uniformly tight.*

Proof. Put

$$W_\varepsilon = \left\{ w \in C(0,T;\mathbb{R}^m);\ \sup_{0\le t\le T} |w(t)| \le q_\varepsilon \ \text{ and } \sup_{|t-s|\le T/n^6} |w(t) - w(s)| \le \frac{r_\varepsilon}{n}\,\forall n \ge 1 \right\}$$

and $Z_\varepsilon =$ the set defined by (5.1) for constants K_ε, L_ε, and M_ε, and sequences $\nu_n = \frac{1}{n}$, $\mu_n = \frac{1}{n^3}$, where q_ε, r_ε, K_ε, L_ε, and M_ε are constants to be chosen later, depending on a given $\varepsilon > 0$. From Ascoli–Arzelà's theorem and Lemma 5.1, $W_\varepsilon \times Z_\varepsilon$ is a compact subset of S.

Then

$$\begin{aligned}
&\pi_k((W_\varepsilon \times Z_\varepsilon)^c) \\
&\quad \le P\left(\sup_{0\le t\le T} |B(t)| > q_\varepsilon\right) + \sum_{n=1}^\infty P\left(\sup_{|t-s|\le T/n^6} |B(t) - B(s)| > \frac{r_\varepsilon}{n}\right) \\
&\qquad + P\left(\int_0^T \|z_k(t)\|^2\,dt > K_\varepsilon\right) + P\left(\int_0^T |z_k(t)|^2_\gamma\,dt > L_\varepsilon\right) \\
&\qquad + \sum_{n=1}^\infty P\left(\sup_{|\theta|\le\mu_n} \int_0^T \|z_k(t+\theta) - z_k(t)\|^2_*\,dt > \nu_n M_\varepsilon\right) \\
&\quad \le \frac{1}{q_\varepsilon} E[\sup_{0\le t\le T} |B(t)|] \\
&\qquad + \sum_{n=1}^\infty \sum_{i=0}^{n^6-1} \left(\frac{3n}{r_\varepsilon}\right)^4 E\left[\sup_{iT/n^6\le t\le (i+1)T/n^6} |B(t) - B(iT/n^6)|^4\right] \\
&\qquad + \frac{1}{K_\varepsilon} E\left[\int_0^T \|z_k(t)\|^2\,dt\right] + \frac{1}{L_\varepsilon} E\left[\int_0^T |z_k(t)|^2_\gamma\,dt\right] \\
&\qquad + \sum_{n=1}^\infty \frac{1}{\nu_n M_\varepsilon} E\left[\sup_{|\theta|\le\mu_n} \int_0^T \|z_k(t+\theta) - z_k(t)\|^2_*\,dt\right].
\end{aligned} \tag{5.13}$$

By Lemmas 4.1, 4.3, and 4.4, we have

$$\begin{aligned}&\pi_k((W_\varepsilon \times Z_\varepsilon)^c)\\ &\le C\left\{\frac{1}{q_\varepsilon}+\frac{1}{r_\varepsilon^4}\sum_{n=1}^{\infty}\frac{1}{n^2}+\frac{1}{K_\varepsilon}+\frac{1}{L_\varepsilon}+\frac{1}{M_\varepsilon}\sum_{n=1}^{\infty}\frac{1}{n^2}\right\}\\ &\le C\left\{\frac{1}{q_\varepsilon}+\frac{1}{r_\varepsilon^4}+\frac{1}{K_\varepsilon}+\frac{1}{L_\varepsilon}+\frac{1}{M_\varepsilon}\right\}<\varepsilon\end{aligned} \tag{5.14}$$

for a convenient choice of q_ε, r_ε, K_ε, L_ε, and M_ε.

This completes the proof.

6. Proof of Theorem 3.1. By Prokhorov's theorem, $\{\pi_k\}_{k\ge 1}$ is relatively compact. Hence there is a subsequence, still denoted by $\{\pi_k\}_{k\ge 1}$, and a probability measure π on S such that $\{\pi_k\}_{k\ge 1}$ converges to π weakly. Moreover, by Skorokhod's theorem, there exist S-valued random variables (W_k, y_k) and (W, y) on a suitable probability space $(\tilde{\Omega}, \tilde{\mathcal{F}}, \tilde{P})$ such that

$$\text{the probability law of } (W_k, y_k) = \pi_k, \tag{6.1}$$

$$\text{the probability law of } (W, y) = \pi, \tag{6.2}$$

and with probability 1,

$$W_k \longrightarrow W \quad \text{uniformly on } [0,T], \tag{6.3}$$

$$y_k \longrightarrow y \quad \text{in } L^2(0,T;L^2(\mathbb{R}^d)) \quad \text{strongly.} \tag{6.4}$$

By the same argument as in the proof of Theorem 1.1 in Bensoussan [1], we can see that

$$y_k = \Psi_k(W_k) \quad \text{a.s.} \tag{6.5}$$

and

$$W(t) \text{ is an } \mathcal{F}_t = \sigma\{(W(s), y(s));\ s \le t\}\text{-Brownian motion.} \tag{6.6}$$

From (6.5), Lemmas 4.1 and 4.2 hold for y_k. Therefore, by extracting a new subsequence, still denoted by y_k, we have

$$\begin{aligned} y_k \longrightarrow y \ &\text{in } L^2(\tilde{\Omega}; L^2(0,T;H^1)) \text{ weakly,}\\ &\text{in } L^2(\tilde{\Omega}; L^\infty(0,T;L^2(\mathbb{R}^d))) \text{ weak star,}\\ &\text{in } L^\infty(0,T;L^4(\tilde{\Omega};L^2(\mathbb{R}^d))) \text{ weak star.}\end{aligned} \tag{6.7}$$

From this and Lemmas 4.1 and 4.2, we see that

$$\tilde{E}\left[\int_0^T \|y(t)\|^2\,dt\right] \le C, \tag{6.8}$$

$$\sup_{0\le t\le T} \tilde{E}|y(t)|^4 \le C, \tag{6.9}$$

$$\tilde{E}\left[\sup_{0\le t\le T} |y(t)|^2\right] \le C. \tag{6.10}$$

Combining (6.4) with (4.4) and (6.9), we have

$$y_k \to y \quad \text{in } L^2(\tilde{\Omega} \times (0,T); L^2(\mathbb{R}^d)), \tag{6.11}$$

and thus by extracting a new subsequence, we obtain

$$y_k(t) \to y(t) \quad \text{in } L^2(\mathbb{R}^d) \quad \text{for almost all } (\tilde{\omega}, t). \tag{6.12}$$

By the assumption (A.2) and estimates (4.4) and (6.9), we have

$$f(\cdot\,, y_k(\cdot)) \to f(\cdot\,, y(\cdot)) \quad \text{in } L^2(\tilde{\Omega} \times (0,T); L^2(\mathbb{R}^d)). \tag{6.13}$$

Put

$$\bar{y}_k(t) = \frac{1}{\delta}\int_{r\delta}^{(r+1)\delta} y_k(s)\,ds, \quad t \in [r\delta, (r+1)\delta), \quad r = 0, 1, \dots, k,$$

and

$$\bar{y}^{(k)}(t) = \frac{1}{\delta}\int_{r\delta}^{(r+1)\delta} y(s)\,ds, \quad t \in [r\delta, (r+1)\delta), \quad r = 0, 1, \dots, k.$$

Then, by (6.11) we have

$$\begin{aligned} &\tilde{E}\left[\int_0^T |\bar{y}_k(t) - \bar{y}^{(k)}(t)|^2\,dt\right] \\ &\quad\le \tilde{E}\left[\int_0^T |y_k(t) - y(t)|^2\,dt\right] \to 0 \quad \text{as } k \to \infty. \end{aligned} \tag{6.14}$$

On the other hand, by Lebesgue's theorem

$$\bar{y}^{(k)}(t) \to y(t) \quad \text{in } L^2(\tilde{\Omega} \times \mathbb{R}^d) \quad \text{for almost all } t. \tag{6.15}$$

From this and (6.9), it follows that

$$\bar{y}^{(k)} \to y \quad \text{in } L^2(\tilde{\Omega} \times (0,T); L^2(\mathbb{R}^d)), \tag{6.16}$$

which yields

$$\bar{y}_k \to y \quad \text{in } L^2(\tilde{\Omega} \times (0,T); L^2(\mathbb{R}^d)). \tag{6.17}$$

Moreover, by extracting a new subsequence, we obtain

$$\bar{y}_k(t) \to y(t) \quad \text{in } L^2(\mathbb{R}^d) \quad \text{for almost all } (\tilde{\omega}, t). \tag{6.18}$$

By assumption (A.3) and estimates (4.4) and (6.9), we have

$$g\left(\left[\frac{t}{\delta}\right]\delta, \bar{y}_k(t)\right) \to g(t, y(t)) \quad \text{in } L^2(\tilde{\Omega} \times (0,T); L^2(\mathbb{R}^d)). \tag{6.19}$$

By the same argument as in Bensoussan [1], we can deduce from (6.3) and (6.19) that

$$\int_0^{[t/\delta]\delta} g\left(\left[\frac{s}{\delta}\right]\delta, \bar{y}_k(s)\right) dW_k(s) \to \int_0^t g(s, y(s))\, dW(s) \tag{6.20}$$

$$\text{in } L^2(\tilde{\Omega}; L^2(\mathbb{R}^d)) \text{ weakly.}$$

Since y_k satisfies (4.1), we have

$$\begin{aligned} y_k(t) + \int_0^t A(s) y_k(s)\, ds = y_0 &+ \int_0^{[t/\delta]\delta} f(s, y_k(s))\, ds \\ &+ \int_0^{[t/\delta]\delta} g\left(\left[\frac{s}{\delta}\right]\delta, \bar{y}_k(s)\right) dW_k(s). \end{aligned} \tag{6.21}$$

Using (6.7), (6.13), and (6.20), we can pass to the limit and obtain

$$\begin{aligned} y(t) + \int_0^t A(s) y(s)\, ds = y_0 &+ \int_0^t f(s, y(s))\, ds \\ &+ \int_0^t g(s, y(s))\, dW(s). \end{aligned} \tag{6.22}$$

Hence (W, y) is a solution of (2.3). This completes the proof of Theorem 3.1.

Acknowledgments. The author would like to thank the referees, who read the manuscript carefully and gave him helpful comments.

REFERENCES

[1] A. Bensoussan, *Some existence results for stochastic partial differential equations*, in Stochastic Partial Differential Equations and Applications (Trento, 1990), Pitman Res. Notes Math. Ser. 268, Longman Scientific and Technical, Harlow, UK, 1992, pp. 37–53.

[2] A. Bensoussan, R. Glowinski, and A. Rascanu, *Approximation of the Zakai equation by the splitting up method*, SIAM J. Control Optim., 28 (1990), pp. 1420–1431.

[3] J. L. Lions, *Equations différentielles opérationnelles et problèmes aux limites*, Springer-Verlag, Berlin, 1961.

[4] E. Pardoux, *Equations aux dérivées partielles stochastiques non linéaires monotones*, Thése, Université Paris XI, 1975.

[5] J. B. Walsh, *An Introduction to Stochastic Partial Differential Equations*, Lecture Notes in Math. 1180, Springer-Verlag, New York, Berlin, 1986, pp. 266–437.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 33, No. 6, pp. 1731–1757, November 1995

© 1995 Society for Industrial and Applied Mathematics
006

SOLUTION OF OPTIMAL CONTROL PROBLEMS BY A POINTWISE PROJECTED NEWTON METHOD*

C. T. KELLEY† AND E. W. SACHS‡

Abstract. In the context of optimal control of ordinary differential equations, we prove local superlinear convergence and constraint identification results for an extension of the projected Newton method of Bertsekas. The estimates are also valid for discretized versions of the method-problem pair.

Key words. projected Newton iteration, optimal control

AMS subject classifications. 47H17, 49K15, 49M15, 65J15, 65K10

1. Introduction.

In many areas of optimal control the problems are formulated with simple constraints on the control. For these problems, the gradient projection type algorithms have proven to be quite successful, because they are able to take into account the structure of the underlying optimization problem. Another interesting feature of these methods is that they often can be formulated in infinite-dimensional spaces, which is important for the application to optimal control problems.

In general, let H denote a Hilbert space and for some closed convex subset $U \in H$ consider the optimization problem

$$\text{Minimize} \quad \phi(u) \quad \text{subject to} \quad u \in U. \tag{1.1}$$

If $\mathcal{P} : H \longrightarrow U$ denotes the projection onto the feasible set, then the gradient projection method iterates are given by

$$u_+ = \mathcal{P}(u_c - \alpha_c \nabla\phi(u_c)), \tag{1.2}$$

where $\alpha > 0$ is determined by a step-size rule. In Hilbert space, this algorithm has been formulated and investigated by Goldstein [10] and Levitin and Polyak [13]. The books [4] and [3] discussed the convergence properties of gradient projection methods. In [7] a thorough convergence analysis of the gradient projection method was presented which yields various convergence rates of the algorithm under various assumptions.

Since the gradient projection method as presented in (1.2) is based on and identical for $U = H$ with the steepest descent method, its convergence properties exhibit locally a rather slow rate. This was the motivation to extend Newton's method to a projection method. There are basically two routes by which this goal can be achieved.

If one considers in the unconstrained case a Newton step as the solution of the minimization of a quadratic approximation of ϕ at the current iterates u_c, then for a problem of the type (1.1) one would have to solve

$$\text{Minimize} \ (\nabla\phi(u_c), u - u_c) + \frac{1}{2}(u - u_c, \nabla^2\phi(u_c)(u - u_c)) \ \text{subject to} \ u \in U. \tag{1.3}$$

* Received by the editors June 4, 1993; accepted for publication (in revised form) April 7, 1994.

† North Carolina State University, Department of Mathematics, Box 8205, Raleigh, North Carolina 27695-8205. The research of this author was supported by National Science Foundation grant DMS-9024622, Air Force Office of Scientific Research grant AFOSR-FQ8671-9101094, and North Atlantic Treaty Organization grant CRG 920067. Computing was partially supported by an allocation of time from the North Carolina Supercomputing Center.

‡ Universität Trier, FB IV – Mathematik, 54286 Trier, Germany. The research of this author was supported by the Volkswagen–Stiftung and North Atlantic Treaty Organization grant CRG 920067.

This algorithm has been analyzed in [13] and [6]. The disadvantage of the method (1.3) is that at each step a quadratic problem with constraints needs to be solved. The simplicity of the constraints cannot be used in a direct way through the projection $\mathcal{P}$ except in solving the quadratic problem.

The other route to extend Newton's method to the constrained case is as follows. Instead of projecting the steepest descent direction onto the feasible set, one projects the Newton direction onto the feasible set:

$$u_+ = \mathcal{P}(u_c - \alpha_c(\nabla^2\phi(u_c))^{-1}\nabla\phi(u_c)). \tag{1.4}$$

This method utilizes again the simple projection but has the drawback that it does not always produce a descent in the objective function. Bertsekas [1] and [2] introduced for the finite-dimensional case with simple constraints, such as upper and lower bounds on the variables, a projected Newton method which alleviated this problem. For $H = R^n$ let

$$u_+ = \mathcal{P}(u_c - \alpha_c D_c^{-1}\nabla\phi(u_c)), \tag{1.5}$$

where D_c is a properly chosen matrix such that descent in the objective function is ensured. Let C_J denote the map that sets the components which lie in J of a vector $u \in R^n$ to zero. Then Bertsekas suggested that

$$D_c = C_J^T \nabla^2\phi(u_c) C_J + C_{J^c}^T I C_{J^c},$$

where J contains the components of u_c which are active and the corresponding components of $\nabla\phi(u_c)$ point outside the feasible set. J^c denotes the complement of J in the index set. This algorithm combined with a proper step-size rule can be shown to converge locally at a quadratic rate since it identifies all active constraints after finitely many steps and becomes Newton's method for an unconstrained problem. The assumptions required for superlinear convergence of the method proposed in this paper include assumptions of the type of second-order sufficiency (2.4) and of nondegeneracy (2.3). The latter assumption might not be needed in an approach similar to (1.3), but a larger problem has to be solved at each step.

Since optimal control problems are problems formulated in function space, an analysis of the projected Newton method in this framework would give some insight for the case of fine discretizations. As shown in [16], this issue is important because the identification of finite indices is only mesh independent if proper measures are taken. The goal of this paper is to extend the algorithm to the infinite-dimensional setting of optimal control problems.

The class of problems we seek to solve is

$$\text{minimize} \int_0^T L(x(t),\, u(t),\, t)\, dt$$

over $u \in U$ such that $x \in W_N^{1,\infty}[0,T]$ solves

$$\dot{x}(t) = f(x(t), u(t), t), \quad x(0) = x_0, \quad t \in (0,T).$$

Here $f : R^N \times R \times [0,T] \to R^N$ and $L : R^N \times R \times [0,T] \to R$ and for $t \in [0,T]$ we let U be given by

$$U(t) = \{u \in L^\infty[0,T] \,|\, u_{\min}(t) \le u(t) \le u_{\max}(t)\}$$

with $u_{\min}$ and $u_{\max}$ in $L^\infty[0,T]$.

The projection $\mathcal{P}$ is the map on $L^\infty[0,T]$ given by

$$\mathcal{P}(u)(t) = \begin{cases} u_{\min}(t), & \text{if } u(t) \leq u_{\min}(t), \\ u(t), & \text{if } u_{\min}(t) \leq u(t) \leq u_{\max}(t), \\ u_{\max}(t), & \text{if } u(t) \geq u_{\max}(t). \end{cases} \tag{1.6}$$

We use the notation

$$L_N^\infty[0,T] = L^\infty([0,T];R^N), \quad W_N^{1,\infty}[0,T] = W^{1,\infty}([0,T];R^N)$$

for the spaces of R^N valued functions on $[0,T]$ having components in $L^\infty[0,T]$ and $W^{1,\infty}[0,T]$, respectively. If $w : [0,T] \to R^N$ has components w_i, the norms on $L_N^\infty[0,T]$ and $W_N^{1,\infty}[0,T]$ are given by

$$\|w\|_{L_N^\infty[0,T]} = \sum_{i=1}^N \|w_i\|_{L^\infty[0,T]} \text{ and } \|w\|_{W_N^{1,\infty}[0,T]} = \sum_{i=1}^N \|w_i\|_{W^{1,\infty}[0,T]}.$$

The L^∞ and $W^{1,\infty}$ are defined by

$$\|u\|_{L^\infty[0,T]} = \text{ess-sup}_{t\in[0,T]}|u(t)| \text{ and } \|u\|_{W^{1,\infty}[0,T]} = \|u\|_{L^\infty[0,T]} + \|du/dt\|_{L^\infty[0,T]}.$$

We will work in the spaces

$$X = W_N^{1,\infty}[0,T] \oplus W_N^{1,\infty}[0,T] \oplus L^\infty[0,T] \text{ and } Y = L_N^\infty[0,T] \oplus L_N^\infty[0,T] \oplus L^\infty[0,T].$$

For the unconstrained case, the first-order necessary conditions for optimality can be defined in terms of the Hamiltonian function $H : R^N \times R^N \times R \times R \to R$

$$H(p,x,u,t) = f^T(x,u,t)\,p + L(x,u,t), \quad (p,x,u,t) \in R^{2N+2}.$$

Usually, $p \in W_N^{1,\infty}[0,T]$ denotes the solution of the adjoint equation

$$-\dot{p} = f_x^T p + L_x^T, \quad p(T) = 0.$$

For simplicity we will also use the notation for the Hamiltonian

$$H(p,x,u)(t) := H(p(t),x(t),u(t),t), \quad t \in [0,T]$$

and likewise for the partial derivatives of H.

The first-order necessary conditions for the unconstrained case can be expressed as the system of nonlinear equations

$$F(z) = F(p,x,u) = \begin{pmatrix} \dot{x} - f(x,u,\cdot) \\ \dot{p} + H_x(p,x,u) \\ H_u(p,x,u) \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix} \tag{1.7}$$

for $z = (p,x,u)^T \in X$ and $F : X \to Y$. The advantage of solving the system (1.7) over applying a Newton-like method directly to $\nabla f = H_u = 0$ is that the linear equations for the Newton steps in (1.7) can be expressed as linear equations for the new iterates, without solving a nonlinear differential equation.

For the constrained case, the third equation in (1.7) must be modified to take the constraints into account. The system of nonlinear equations we consider here is

$$\mathcal{F}(z) = \mathcal{F}(p,x,u) = \begin{pmatrix} \dot{x} - f(x,u,\cdot) \\ \dot{p} + H_x(p,x,u) \\ u - \mathcal{P}(u - H_u(p,x,u)) \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}. \tag{1.8}$$

To formulate the algorithm we introduce some more notation: For $I \subset [0,T]$, $A = [0,T] \setminus I$, and $z \in X$ define

$$Q_I z = (p, x, \chi_I u)^T,$$

$$\mathcal{F}_I(z) = Q_I F(p, x, \chi_I u + \chi_A u^*).$$

Note that in the definition of $\mathcal{F}_I$, $Q_I F$ and not $Q_I \mathcal{F}$ is used. This is important not only to make $\mathcal{F}_I'$ well defined but also for the success of the algorithm described in §4.

$\mathcal{F}_I$ can be regarded as a map from

$$X_I = W_N^{1,\infty}[0,T] \oplus W_N^{1,\infty}[0,T] \oplus L^\infty[I]$$

to

$$Y_I = L_N^\infty[0,T] \oplus L_N^\infty[0,T] \oplus L^\infty[I].$$

If the third component of $\mathcal{F}_I$ is understood to be identically zero on A, we may also regard $\mathcal{F}_I$ as a map from X to Y. We will use $\mathcal{F}_I$ as both a theoretical and computational tool. When used in computation we must have knowledge of u^* on A. This is analogous to identification of the active set in finite-dimensional problems [1], [2]. In the infinite-dimensional setting considered in this paper, complete identification of A^* is not possible. However, as we will show in §4, a useful subset of A^* can be identified.

The authors extended in [12] Bertsekas' gradient projection method to constrained compact fixed-point problems. It was combined with a multilevel algorithm of Atkinson and Brakhage and applied to a parabolic boundary control problem with simple bound constraints on the control. In this paper, we do not assume any compactness of the nonlinear map and analyze the resulting algorithm in infinite dimensions. As we will see in §4, relaxing the assumptions with regard to the compactness gives rise to an additional smoothing step in the algorithm so that a proper identification of the active set can be done at the subsequent iteration.

Section 3 contains a series of lemmas that are rather technical but lead to an important estimate in Theorem 3.5 in which the norm of the residual $\mathcal{F}$ can be used in an upper bound on the distance of the current iterate to the solution in the strong X-norm. The transfer from a projection-based method, which has its natural formulation in a Hilbert space like L^2, to an L^∞-type norm in X poses in the analysis of the convergence various difficulties. In another context this aspect has been the focus of other research activities, see, e.g., [9], [8]. The estimate in Theorem 3.5 enables us to show a result on the identification of the set of active indices. It estimates the measure of the set on which the active set at the current iterate differs from the active set of the solution by the distance of the iterate from the solution in the X-norm. This result is the key for the convergence analysis of the algorithm.

In the following algorithm we set $u_c = u^*$ on $\bar{A}$, which is a well-defined step by Lemma 3.7, and then apply a projected Newton iteration with $\bar{A}(z_c)$ as the active set. This yields an intermediate iterate $(x_{1/2}, p_{1/2})$ for the state and costate. Then a smoothing step for x and p is added which properly determines the set $\bar{A}(z_+)$ at the next iterate. The iteration is formally given by the following algorithm.

ALGORITHM proj_newt($\mathcal{F}, z_c, z_+$). Choose $\bar{p} \in (1/2, 1)$

1. Compute

$$\bar{A}(z_c) = \{t \,|\, |H_u(z_c)(t)| \geq \|\mathcal{F}(z_c)\|_Y^{\bar{p}}\}.$$

2. Set $u_c = u^*$ on $\bar{A}(z_c)$.
3. Compute the projected Newton step

$$s = -Q_{\bar{I}} \mathcal{F}_{\bar{I}}'(z_c)^{-1} Q_{\bar{I}} F(z_c).$$

4. Set $z_{1/2} = \bar{P}(z_c + s)$ and $u_+ = u_{1/2}$, where

$$\bar{P} = \begin{pmatrix} I & 0 & 0 \\ 0 & I & 0 \\ 0 & 0 & \mathcal{P} \end{pmatrix}.$$

5. Compute

$$x_+(t) = \int_0^t f(x_{1/2}(s), u_+(s), s)\, ds + x_0,$$
$$p_+(t) = \int_t^T H_x(p_{1/2}(s), x_{1/2}(s), u_+(s))\, ds.$$

Apart from the smoothing step, this algorithm differs from the gradient projection method given by Bertsekas even in the finite-dimensional case. To clarify this point consider a optimal control problem where in step (1.5) the new state x_+ in Bertsekas' algorithm has to be computed by solving the nonlinear system equation. In Algorithm proj_newt the correction for the new state is computed from a linearized version of the system equation, see, e.g., (6.6).

In §4 we prove the local convergence of the iterates of the algorithm to the solution which is superlinear. The order of the rate of convergence is dependent on how stringently the identification of the active set is carried out. The convergence is of q-order $2\bar{p}$ depending on the choice of $\bar{p} > 1/2$ in $\bar{A}(z)$ in the algorithm. The norm which we use in the convergence statement is the X-norm which is stronger than the Y-norm.

The contents of §5 relate the assumptions of the lemmas and theorems to other assumptions which have been used in the literature. In particular, the relation to second-order sufficiency conditions is clarified. Also, the role of the growth conditions in [6] is indicated in our context.

Section 6 contains comments on the implementation of the algorithm. Furthermore, we present for an example numbers illustrating the convergence rate estimates of the convergence theorem. In particular, the dependence on $\bar{p}$ can be observed.

2. Notation. In this section we introduce some more notation, assumptions, and immediate consequences of these.

Since U is given by

$$U(t) = \{u \in L^\infty[0,T] \,|\, u_{\min}(t) \leq u(t) \leq u_{\max}(t)\}$$

we define the 2-point set

$$\partial U(t) = \{u_{\min}(t), u_{\max}(t)\}.$$

We assume the existence of a solution to the first-order necessary conditions.

ASSUMPTION 2.1. *There exists $z^* = (p^*, x^*, u^*)$ such that $\mathcal{F}(z^*) = 0$.*

Since $z^* \in X$, its range is contained in the bounded set

$$\mathcal{R} = \{\xi \in R^N \times R^N \times R \,|\, \|\xi\|_\infty \leq \|z^*\|_X\}.$$

Also $(x, u) \in \mathcal{R}_1 \subset R^N \times R$ where

$$\mathcal{R}_1 = \{\zeta \in R^N \times R \,|\, \|\zeta\|_\infty \leq \|z^*\|_X\}.$$

ASSUMPTION 2.2. *There is an open set $\mathcal{R}_0 \supset \mathcal{R}_1$ such that f, L and their first and second partial derivatives with respect to x and u are uniformly Lipschitz continuous on $\mathcal{R}_0 \times [0, T]$.*

Lemma 2.1 is an immediate consequence of Assumption 2.2 and the fact that for $z, w \in X$, $F'(z) - F'(w)$ is a multiplication operator and not a differential operator.

In the rest of the paper we denote function space norms by $\|\cdot\|$ and norms in R^k by

$$|x| = \max\{|x_j| \mid j = 1, ..., k\}.$$

LEMMA 2.1. *There are $\sigma^*, L_F, M_F > 0$ such that for all $z, w \in \mathcal{N} = \{v \in X \,|\, \|v - z^*\|_X < \sigma^*\}$ and $t \in [0, T]$,*

$$|F(z)(t) - F(w)(t)| \leq M_F |z(t) - w(t)|, \tag{2.1}$$

$$\begin{aligned} &\|F'(z) - F'(w)\|_{\mathcal{L}(X,Y)} \leq L_F \|z - w\|_Y \text{ and} \\ &\|F'(z) - F'(w)\|_{\mathcal{L}(Y,Y)} \leq L_F \|z - w\|_Y. \end{aligned} \tag{2.2}$$

We define active and inactive sets for u by

$$A(u) = \{t \,|\, u(t) \in \partial U(t)\} \text{ and } I(u) = [0, T] \setminus A(u) = \{t \,|\, u(t) \in \text{int}(U)(t)\}. \tag{2.3}$$

In (2.3) $\text{int}(U)(t)$ is defined to be the interval

$$\text{int}(U)(t) = (u_{\min}(t), u_{\max}(t)).$$

We let

$$A^* = A(u^*) \text{ and } I^* = I(u^*).$$

If $S \subset R^k$ for some k and $t \in R^k$ we denote the distance from t to S by

$$\text{dist}(t, S) = \inf\{s \in S \,|\, |t - s|\}.$$

As in [12] we make structural assumptions on the active set at the solution.

ASSUMPTION 2.3. *There is $\nu \in (0, 1)$ such that $u_{\max}(t) \geq u_{\min}(t) + \nu$ for all $t \in [0, T]$. A^* is the closure of a finite union of open sets. On each component of A^* either $u = u_{\max}$ or $u = u_{\min}$.*

Moreover, there is c_1 such that

$$\begin{aligned} &|H_u(p^*, x^*, u^*)(t)| \geq c_1 \mathrm{dist}(t, \partial A^*) \textit{ for all } t \in A^* \textit{ and} \\ &\mathrm{dist}(t, A^*) \leq c_1^{-1} \mathrm{dist}(u^*(t), \partial U(t)) \textit{ for all } t \in I^*. \end{aligned} \tag{2.4}$$

This assumption is a condition on the slope of the function H_u, which is also called switching function in the control context. In §5 we relate Assumption 2.3 to the growth condition in [7]. We also give a connection to a similar condition in [16] which has been used for the finite identification of active indices in the gradient projection method.

The fact that we consider problems of dimension 1 with regard to the control together with the assumption on the structure of A^* yields the following lemma, which we give without proof.

LEMMA 2.2. *There is $c_0 > 0$ such that for all $\delta > 0$ the sets*

$$E_\delta = \{t \in R \,|\, \mathrm{dist}(t, \partial A^*) < \delta\}$$

are uniformly bounded in measure by

$$\mu(E_\delta) \leq c_0 \delta. \tag{2.5}$$

We define projections P_1 and P_2 for $z = (p, x, u)^T \in Y$ by

$$P_1 z = (p, x, 0)^T, \qquad P_2 z = (0, 0, u)^T.$$

The observations that $P_1 F = P_1 \mathcal{F}$ and $Q_J P_1 = P_1 Q_J = P_1$ for any $J \subset [0, T]$ lead to the following result.

PROPOSITION 2.3. *Let Assumptions 2.1 and 2.2 hold. If $I \subset I^*$ then $\mathcal{F}_I(z^*) = 0$ and for all $t \in [0, T]$*

$$|P_1 \mathcal{F}_I(z)(t)| \leq \|P_1 \mathcal{F}(z)\|_Y + M_F |(I - Q_I)(z - z^*)(t)|.$$

Proof. Since

$$P_1 \mathcal{F}_I(z) = P_1 Q_I F(Q_I z + (I - Q_I) z^*)$$

and

$$P_1 Q_I \mathcal{F}(z) = P_1 Q_I F(z)$$

we have the result by Lemma 2.1. □

Nonsingularity assumptions are also complicated by the constraints.

ASSUMPTION 2.4. *There are K_η and $\tilde{\eta} > 0$ such that if*

$$I \subset \{t \,|\, |H_u(p^*, x^*, u^*)(t)| \leq \tilde{\eta}\} \tag{2.6}$$

then $\mathcal{F}_I'(z^*)$ is a nonsingular map from X_I to Y_I. Moreover,

$$\|\mathcal{F}_I'(z^*)^{-1}\|_{\mathcal{L}(Y_I, X_I)} \leq K_\eta \tag{2.7}$$

and for any measurable set $S \subset [0, T]$

$$\|P_1 \mathcal{F}_I'(z^*)^{-1}(I - Q_S)\|_{\mathcal{L}(Y_I, Y_I)} \leq K_\eta \mu(S). \tag{2.8}$$

There is $\tilde{\alpha}$ such that

$$H_{uu}(p^*, x^*, u^*)(t) \geq \tilde{\alpha}$$

for all $t \in I$.

This assumption is related to second-order sufficiency conditions for optimal control problems. Details are discussed in §5.

3. Identification of the active set.

We begin by summarizing the definitions of the many projections that will be used in the following sections. Beginning with the definition of $\mathcal{P}$ in (1.6) as the projection of u onto the feasible set, we define $\bar{P}$ $z = (p, x, u)^T \in Y$ by

$$\bar{P}(z) = (p, x, \mathcal{P}(u))^T.$$

For a measurable set $I \in [0, T]$ we define

$$Q_I(z) = (p, x, \chi_I u)^T,$$

where χ_I is the characteristic function of the set I. Finally we define P_1 and P_2, which decompose z into the state-costate and control parts by

$$P_1(z) = (p, x, 0)^T, \qquad P_2 z = (0, 0, u)^T.$$

For $z \in X$ we let

$$e = (e_p, e_x, e_u)^T = z - z^*.$$

We want to show that provided $\|e\|_X$ is sufficiently small, it can be estimated by a constant multiple of $\|\mathcal{F}\|_Y$. This estimate is important in that it allows us to identify a subset of A^*. We will require a sequence of lemmas.

LEMMA 3.1. *Assume that Assumptions 2.1–2.4 hold. Then there are B_0, c_2, and $\tau_0 > 0$ such that if $z = \bar{P}(z) \in X$ is such that $\|e\|_X \leq \tau_0$, then for $S(z) = \{t \,|\, u(t) - H_u(p, x, u)(t) \notin U(t)\} \cap I^* \cap I(u)$,*

$$|P_2 \mathcal{F}_{I(u) \cap I^*}(z)(t)| \leq B_0(\|\mathcal{F}(z)\|_Y + \chi_{S(z)}(t)\|e\|_Y),$$

where

$$\mu(S(z)) \leq c_2 \|e\|_Y.$$

Proof. We assume that $\tau_0 < \min(\sigma^*, \nu)$, where σ^* is the diameter of the set $\mathcal{N}$ in Lemma 2.1. We set $J = I^* \cap I(u)$ in this proof and set $\|e\|_Y = \sigma \leq \|e\|_X$.

Observe that

$$P_2 \mathcal{F}_J(z) = (0, 0, \chi_J H_u(p, x, \chi_J u + \chi_{J^c} u^*))^T = (0, 0, \chi_J H_u(p, x, u))^T$$

and therefore

$$P_2 \mathcal{F}_J(z) = (0, 0, (1 - \chi_{S(z)})\chi_J H_u(p, x, u) + \chi_{S(z)} H_u(p, x, u)(t))^T. \tag{3.1}$$

On $S(z)$ we have $H_u(p^*, x^*, u^*) = 0$, since $S(z) \subset J \subset I^*$. Hence for $t \in S(z)$

$$\begin{aligned} |\chi_{S(z)} H_u(p, x, u)(t)| &= |\chi_{S(z)}(H_u(p, x, u) - H_u(p^*, x^*, u^*))(t)| \\ &\leq \chi_{S(z)} M_F \|e\|_Y. \end{aligned} \tag{3.2}$$

For $t \in S(z)^c$ we have $u - H_u(p,x,u) \in U$ and

$$(1 - \chi_{S(z)})(0, 0, H_u(p,x,u))^T = (1 - \chi_{S(z)}) P_2 \mathcal{F}(z);$$

hence

$$|(1 - \chi_{S(z)}) H_u(p,x,u)| \leq \|\mathcal{F}\|_Y.$$

This proves the first part of the assertion with $B_0 = \max\{1, M_F\}$.

We now complete the proof with an estimate of the measure of $S(z)$. Let $t \in S(z)$. The estimate (3.2) and $H_u(p^*, x^*, u^*)(t) = 0$ yield

$$\|\chi_{S(z)}(u^* - H_u(p^*, x^*, u^*) - (u - H_u(p,x,u)))\|_\infty \leq (1 + M_F)\sigma.$$

Since $u - H_u(p,x,u) \notin U(t)$ we must have

$$\mathrm{dist}(u^*(t), \partial U(t)) \leq (1 + M_F)\sigma.$$

Hence, by Assumption 2.3,

$$\mathrm{dist}(t, A^*) \leq c_1^{-1}(1 + M_F)\sigma$$

for any $t \in S(z)$. Therefore

$$S(z) \subset \{t \,|\, \mathrm{dist}(t, \partial A^*) \leq c_1^{-1}(1 + M_F)\sigma\},$$

and, further, by Lemma 2.2,

$$\mu(S(z)) \leq c_0 c_1^{-1}(1 + M_F)\sigma.$$

This completes the proof with $c_2 = c_0 c_1^{-1}(1 + M_F)$. □

COROLLARY 3.2. *Assume that Assumptions* 2.1–2.4 *hold. Then if* $\|e\|_X \leq \tau_0$

$$\|P_1 \mathcal{F}'_{I(u)\cap I^*}(z^*)^{-1} P_2 \mathcal{F}_{I(u)\cap I^*}(z)\|_Y \leq B_1(\|\mathcal{F}(z)\|_Y + \|e\|_Y^2). \tag{3.3}$$

Proof. The result follows directly from Assumption 2.4 and Lemma 3.1 with $B_1 = K_\eta B_0(1 + c_2)$. □

LEMMA 3.3. *Assume that Assumptions* 2.1–2.4 *hold. Then there are* B_5 *and* $\tau_1 > 0$ *such that if* $z = \bar{P}(z) \in X$ *is such that* $\|e\|_X \leq \tau_1$, *then*

$$\|P_1 \mathcal{F}'_{I(u)\cap I^*}(z^*)^{-1} P_1 \mathcal{F}_{I(u)\cap I^*}(z)\|_Y \leq B_5(\|\mathcal{F}(z)\|_Y + \|e\|_Y^2). \tag{3.4}$$

Proof. We assume that $\tau_1 \leq \tau_0$ with τ_0 chosen as in Lemma 3.1. We set $I = I(u)$ and $A = A(u)$ in this proof. We let $\|e\|_Y = \sigma \leq \|e\|_X \leq \tau_1$ and $\|\mathcal{F}(z)\|_Y = \delta$. We let $J = I^* \cap I$.

By Proposition 2.3

$$|P_1 \mathcal{F}_J(z)| \leq \delta + M_F(1 - \chi_J)|e_u| = \delta + M_F \chi_{S_U}|e_u|,$$

for all $t \in [0,T]$, where S_U is the support of $(1 - \chi_J)e_u$. Our next task is to study S_U.

Assumption 2.3 implies that if we choose $\tau_1 < \nu$ then $u = u^*$ on $E = A(u) \cap A^*$. Hence $(1 - \chi_J)e_u$ is nonzero only in the set

$$E_1 = ([0,T] \setminus (E \cup J)) = (I^* \cap A) \cup (I \cap A^*)$$

and hence $S_U \subset E_1$. We consider three cases. First, if $t \in (I^* \cap A)$ then $H_u(p^*, x^*, u^*) = 0$. If $t \in (I \cap A^*)$ and $u - H_u(p, x, u) \in \operatorname{int}(U)$, then $|H_u(p, x, u)(t)| = |P_2 \mathcal{F}(z)(t)| \leq \delta$ and therefore

$$|H_u(p^*, x^*, u^*)(t)| \leq \delta + M_F \sigma.$$

If we now let

$$E_2 = \{t \in E_1 \,|\, u - H_u(p, x, u) \in \operatorname{int}(U)\} \cup (I^* \cap A)$$

then for all $t \in E_2$ and $B_2 = 1 + M_F$

$$|H_u(p^*, x^*, u^*)(t)| \leq B_2(\delta + \sigma). \tag{3.5}$$

We must consider a third case:

$$t \in E_3 = \{t \in I \cap A^* \,|\, u - H_u(p, x, u) \notin \operatorname{int}(U)\}.$$

Since

$$\begin{aligned} |\mathcal{P}(u - H_u(p, x, u)) - u^*| &= |\mathcal{P}(u - H_u(p, x, u)) - \mathcal{P}(u^* - H_u(p^*, x^*, u^*))| \\ &\leq (1 + M_F)\sigma \end{aligned} \tag{3.6}$$

for all $t \in [0, T]$, we may reduce τ_1 if needed so that $(1 + M_F)\tau_1 < \nu$ to conclude that if $\sigma < \tau_1$ and if $t \in E_3$ then

$$u^* = \mathcal{P}(u - H_u(p, x, u));$$

hence

$$P_2 \mathcal{F}(z) = (0, 0, u - \mathcal{P}(u - H_u(p, x, u)))^T = (0, 0, e_u)^T. \tag{3.7}$$

This implies that

$$|\chi_{E_3} e_u| \leq \|\mathcal{F}(z)\|_Y = \delta.$$

At this point we have

$$\begin{aligned} |P_1 \mathcal{F}_J(z)| \leq \delta + M_F(1 - \chi_J)|e_u| &= \delta + M_F \chi_{E_2}|e_u| + M_F \chi_{E_3}|e_u| \\ &\leq (1 + M_F)\delta + M_F \chi_{E_2}|e_u|. \end{aligned} \tag{3.8}$$

The Banach lemma, Assumption 2.4, and Lemma 2.1 imply that if $\sigma < \sigma^*$ then

$$\|\mathcal{F}_J'^{\,-1}\|_{\mathcal{L}(Y_I, X_I)} \leq K_\eta / (1 - \sigma M_F).$$

Hence, reducing τ_1 if needed so that $\tau_1 M_F < 1/2$, we have

$$\|(\mathcal{F}_J')^{-1} P_1 \mathcal{F}_J(z)\|_Y \leq 2K_\eta((1 + M_F)\delta + M_F \sigma \mu(E_2)). \tag{3.9}$$

We now estimate the measure of E_2. Using (2.4) we see that if $t \in E_2$ then either $t \in I \cap A^*$ and from (3.5)

$$\operatorname{dist}(t, \partial A^*) \leq c_1^{-1} B_2(\sigma + \delta)$$

or $t \in I^* \cap A$ and

$$\operatorname{dist}(t, \partial A^*) \le c_1^{-1} \operatorname{dist}(u^*(t), \partial U(t)) \le c_1^{-1}|u(t) - u^*(t)| \le c_1^{-1}\sigma.$$

Hence, setting $B_3 = c_1^{-1}\max(B_2, 1)$

$$\operatorname{dist}(t, \partial A^*) \le c_1^{-1} B_3(\sigma + \delta) \text{ for all } t \in E_2.$$

Then Lemma 2.2 implies that

$$\mu(E_2) \le B_4(\sigma + \delta), \tag{3.10}$$

where $B_4 = B_3 c_0 c_1^{-1}$. If we reduce τ_1 if needed so that $M_F B_4 \tau_1 \le 1$ and set $B_5 = 2K_\eta(2 + M_F + M_F B_4)$, then the proof is complete. □

LEMMA 3.4. *Assume that Assumptions* 2.1–2.4 *hold. Then there are* B_9 *and* $\tau_2 > 0$ *such that if* $z = \bar{P}(z) \in X$ *is such that* $\|e\|_X \le \tau_2$, *then*

$$\|P_2 e\|_Y \le B_9(\|\mathcal{F}(z)\|_Y + \|P_2 e\|_Y^2 + \|P_1 e\|_Y). \tag{3.11}$$

Proof. We assume that $\tau_2 \le \tau_1$ with τ_1 from Lemma 3.3. Let $e_p = p - p^*$, $e_x = x - x^*$, and $e_u = u - u^*$. We set $I = I(u)$ and $A = A(u)$ in this proof. We let $\|e\|_Y = \sigma \le \|e\|_X \le \tau_1$, $\|P_1 e\|_Y = \beta$, and $\|\mathcal{F}(z)\|_Y = \delta$. We let $J = I^* \cap I$.

We define a set $\mathcal{S}$ by

$$\mathcal{S} = \{t \,|\, u - H_u(p, x, u) \in U(t)\}.$$

Note that if $t \in \mathcal{S}$ then the third component of $\mathcal{F}(z)$ is H_u. Therefore $|H_u(p, x, u)| \le \delta$ for all $t \in \mathcal{S}$. Therefore, since $H_u(p^*, x^*, u^*) = 0$ on I^*, both $H_u(p^*, x^*, u^*) = 0$ and $H_u(p, x, u) = O(\delta)$ on the set $\mathcal{S} \cap I^*$. Hence, for all $t \in \mathcal{S} \cap I^*$ we have

$$\begin{aligned} 0 = H_u(p^*, x^*, u^*) &= H_u(p, x, u^*) + O(\beta) \\ &= H_u(p, x, u) - H_{uu}(p^*, x^*, u^*)e_u + O(\sigma^2 + \beta). \end{aligned}$$

The bound away from 0 of H_{uu} implies that there is B_6 such that

$$\|\chi_{\mathcal{S} \cap I^*} e_u\|_\infty \le B_6(\delta + \sigma^2 + \beta).$$

On $\mathcal{S} \cap A^*$, $|H_u(p, x, u)| \le \delta$ and hence

$$H_u(p^*, x^*, u^*) + H_{uu}(p^*, x^*, u^*)e_u = O(\delta + \sigma^2 + \beta).$$

As for $H_u(p^*, x^*, u^*)$ we know that

$$H_u(p^*, x^*, u^*)(u - u^*) \ge 0.$$

If $u^* = u_{\max}$, say, then $e_u \le 0$ and therefore $H_u(p^*, x^*, u^*) \le 0$. Hence

$$0 \le -H_u(p^*, x^*, u^*) = H_{uu}(p^*, x^*, u^*)e_u + O(\delta + \sigma^2 + \beta).$$

Since $e_u \le 0$ we must have $|e_u| = O(\delta + \sigma^2 + \beta)$. Applying a similar argument to the case where $u = u_{\min}$ implies that

$$\|\chi_{\mathcal{S} \cap A^*} e_u\|_\infty \le B_7(\delta + \sigma^2 + \beta).$$

We must now estimate $|e_u|$ on $\mathcal{S}^c = [0,T] \setminus \mathcal{S}$. If $t \in \mathcal{S}^c$ we have $u - H_u(p,x,u) \notin U(t)$ and therefore $\mathcal{P}(u - H_u(p,x,u)) \in \partial U(t)$. On $A^* \cap \mathcal{S}^c$, $u^* = \mathcal{P}(u - H_u(p,x,u))$ if σ_0 is sufficiently small by (3.6). Therefore (3.7) holds and so

$$\|\chi_{A^* \cap \mathcal{S}^c} e_u\|_\infty \leq \delta.$$

On $I^* \cap \mathcal{S}^c$, $0 = H_u(p^*, x^*, u^*)$. Since $\mathcal{P}(u - H_u(p,x,u)) \in \partial U(t)$, either $\mathcal{P}(u - H_u(p,x,u)) = u_{\max}$ or $\mathcal{P}(u - H_u(p,x,u)) = u_{\min}$. If $\mathcal{P}(u - H_u(p,x,u)) = u_{\max}$, say, then

$$|u - \mathcal{P}(u - H_u(p,x,u))| = |u - u_{\max}| \leq \delta. \tag{3.12}$$

Hence

$$u_{\max} + \delta - H_u(p,x,u) \geq u - H_u(p,x,u) \geq u_{\max}$$

and therefore $H_u(p,x,u) \leq \delta$. Similarly if $\mathcal{P}(u - H_u(p,x,u)) = u_{\min}$, $H_u(p,x,u) \geq -\delta$. Now, if $\mathcal{P}(u - H_u(p,x,u)) = u_{\max}$ then $H_u(p,x,u) \leq \delta$ and therefore

$$-\delta \leq u - \mathcal{P}(u - H_u(p,x,u)) = u - u_{\max} \leq u - u^*. \tag{3.13}$$

Hence $e_u \geq -\delta$. Also,

$$\begin{aligned} 0 = H_u(p^*, x^*, u^*) &= H_u(p,x,u) - H_{uu}(p^*, x^*, u^*)e_u + O(\delta + \sigma^2 + \beta) \\ &\leq -H_{uu}(p^*, x^*, u^*)e_u + \delta + O(\delta + \sigma^2 + \beta) \\ &= -H_{uu}(p^*, x^*, u^*)e_u + O(\delta + \sigma^2 + \beta). \end{aligned} \tag{3.14}$$

Since $H^*_{uu} \geq \tilde{\alpha}$ on I^* by Assumption 2.4, (3.14) implies

$$e_u \leq \tilde{\alpha}^{-1} O(\delta + \sigma^2 + \beta) = O(\delta + \sigma^2 + \beta). \tag{3.15}$$

We may use (3.13) and (3.15) to conclude that

$$-\delta \leq e_u \leq O(\delta + \sigma^2 + \beta)$$

and therefore $e_u = O(\delta + \sigma^2 + \beta)$. The estimate is exactly the same if $u = u_{\min}$. Hence there is B_8 such that

$$\|\chi_{I^* \cap \mathcal{S}^c} e_u\|_\infty \leq B_8(\delta + \sigma^2 + \beta).$$

This completes the proof with $B_9 = \max\{B_6, B_7, B_8\}$. □

THEOREM 3.5. *Assume that Assumptions* 2.1–2.4 *hold. Then there are* $K_X, \sigma_0 > 0$ *such that if* $z = \bar{P}(z) \in X$ *satisfies* $\|z - z^*\|_X \leq \sigma_0$ *then*

$$\|z - z^*\|_X \leq K_X \|\mathcal{F}(z)\|_Y. \tag{3.16}$$

Proof. We assume that $\sigma_0 < \tau_1$ with τ_1 by Lemma 3.3. We let $\|e\|_Y = \sigma \leq \|e\|_X \leq \sigma_0$ and $\|\mathcal{F}(z)\|_Y = \delta$. We let $J = I^* \cap I(u)$. By Proposition 2.3, $\mathcal{F}_J(z^*) = 0$.

$$\begin{aligned} \mathcal{F}_J(z) &= \mathcal{F}_J(z^*) + \int_0^1 \mathcal{F}'_J(z^* + tQ_J e)Q_J e\, dt \\ &= \mathcal{F}'_J(z^*)Q_J e + \int_0^1 (\mathcal{F}'_J(z^* + tQ_J e) - \mathcal{F}'_J(z^*))Q_J e\, dt. \end{aligned} \tag{3.17}$$

We now have, using (3.17),

$$Q_J e = \mathcal{F}_J'(z^*)^{-1}\left(\mathcal{F}_J(z) - \int_0^1 (\mathcal{F}_J'(z^* + tQ_J e) - \mathcal{F}_J'(z^*))Q_J e\, dt\right). \tag{3.18}$$

By Lemma 2.1 the integral term satisfies

$$\left\|\int_0^1 (\mathcal{F}_J'(z^* + tQ_J e) - \mathcal{F}_J'(z^*))Q_J e\, dt\right\|_Y \leq L_F\sigma^2/2, \tag{3.19}$$

where L_F is the bound in Lemma 2.1 and hence with (2.7)

$$\left\|\mathcal{F}_J'(z^*)^{-1}\int_0^1 (\mathcal{F}_J'(z^* + tQ_J e) - \mathcal{F}_J'(z^*))Q_J e\, dt\right\|_X \leq K_\eta L_F\sigma^2. \tag{3.20}$$

It remains to estimate $\mathcal{F}_J'(z^*)^{-1}\mathcal{F}_J(z)$.

By Lemma 3.3 and Corollary 3.2 we have

$$\|P_1\mathcal{F}_J'(z^*)^{-1}\mathcal{F}_J(z)\|_Y \leq B_{10}(\delta + \sigma^2), \tag{3.21}$$

where $B_{10} = B_1 + B_5$. At this point we can estimate (3.18) using (3.20), (3.21), and $\|x\|_Y \leq \|x\|_X$ for $x \in X$ as follows:

$$\|P_1 e\|_Y = \|P_1 Q_J e\|_Y \leq B_{11}(\delta + \sigma^2), \tag{3.22}$$

where $B_{11} = B_{10} + K_\eta L_F$.

By the definition of P_2, Lemma 3.4, and Assumption 2.4 we have

$$\|P_2 e\|_X = \|P_2 e\|_Y \leq B_9(1 + B_{11})(\delta + \sigma^2). \tag{3.23}$$

From this we conclude with (3.22) that

$$\sigma = \|e\|_Y \leq B_{12}(\delta + \sigma^2),$$

where $B_{12} = B_{11} + B_9(1 + B_{11})$. Hence, reducing σ_0 if necessary so that $B_{12}\sigma \leq B_{12}\sigma_0 \leq 1/2$,

$$\|e\|_Y = \sigma \leq 2B_{12}\delta \text{ and } \sigma^2 \leq \sigma 2B_{12}\delta \leq \delta. \tag{3.24}$$

To obtain an estimate for $\|e\|_X$, not $\|e\|_Y$, we use (3.23) and the second two parts of (3.24):

$$\|e\|_X \leq \|P_1 e\|_X + \|P_2 e\|_X \leq \|P_1 Q_J e\|_X + 2B_9(1 + B_{11})\delta. \tag{3.25}$$

We estimate $\|P_1 Q_J e\|_X$ in two parts based on (3.18). Lemma 3.1 and Proposition 2.3 together with (3.24) imply that

$$\begin{aligned}\|\mathcal{F}_J(z)\|_Y &\leq \|P_1\mathcal{F}(z)\|_Y + M_F\|(I - Q_J)e\|_Y + B_0(\|\mathcal{F}(z)\|_Y + \|e\|_Y)\\ &\leq (1 + B_0)\delta + (M_F + B_0)\|e\|_Y \leq B_{13}\delta,\end{aligned}$$

where

$$B_{13} = (1 + 2M_F B_{12}) + B_0(1 + 2B_{12}).$$

Therefore by Assumption 2.4

$$\|\mathcal{F}_J'(z^*)^{-1}\mathcal{F}_J(z)\|_X \leq K_\eta\|\mathcal{F}_J(z)\|_Y \leq K_\eta B_{13}\delta. \tag{3.26}$$

Hence we can conclude the proof by estimating (3.25) further with (3.18), (3.26), (3.20), and (3.24) to obtain (3.16) with $K_X = K_\eta(B_{13} + L_F) + 2B_9(1 + B_{11})$. □

As a consequence we have a result on identification of the active set A^*.

THEOREM 3.6. *Assume that Assumptions* 2.1–2.4 *hold. For all* $\bar{p} \in (0,1)$ *there is* σ_1 *such that if* $\|z - z^*\|_X < \sigma_1, z \neq z^*$ *and*

$$\bar{A}(z) = \{t \,|\, |H_u(z)(t)| \geq \|\mathcal{F}(z)\|_Y^{\bar{p}}\}, \tag{3.27}$$

then $\bar{A}(z) \subset A^*$ *and there is* c_μ *such that*

$$\mu(A^* \setminus \bar{A}(z)) \leq c_\mu\|z - z^*\|_X^{\bar{p}}. \tag{3.28}$$

Proof. Let $\sigma_1 < \sigma_0$ as in Theorem 3.5, so that the consequences of Theorem 3.5 hold. Let $\|\mathcal{F}(z)\|_Y = \delta$ and $\|z - z^*\|_X = \sigma$. Let L_H denote the Lipschitz constant of H_u. For $t \in \bar{A}(z)$ we have

$$|H_u(z^*)(t)| \geq |H_u(z)(t)| - L_H\sigma \geq \delta^{\bar{p}} - L_H\sigma \geq (K_X^{-1}\sigma)^{\bar{p}} - L_H\sigma > 0$$

if

$$\sigma^{1-\bar{p}} \leq K_X^{-\bar{p}}/L_H.$$

Hence if

$$\sigma_1 \leq (K_X^{-\bar{p}}/L_H)^{1/(1-\bar{p})},$$

then $t \in \bar{A}(z)$ implies that $H_u(z^*)(t) > 0$ and therefore that $t \in A^*$. We now set

$$\sigma_1 = \min(\sigma_0, (K_X^{-\bar{p}}/L_H)^{1/(1-\bar{p})}).$$

If $t \in A^* \setminus \bar{A}$ then, using (2.4) from Assumption 2.3,

$$\begin{aligned} c_1\text{dist}(t, \partial A^*) \leq |H_u(z^*)(t)| &\leq |H_u(z^*)(t) - H_u(z)| + |H_u(z)| < L_H\sigma + \delta^{\bar{p}} \\ &\leq L_H\sigma + M_F^{\bar{p}}\sigma^{\bar{p}}. \end{aligned} \tag{3.29}$$

This implies that $t \in E_\zeta$ where

$$\zeta = (L_H\sigma + M_F^{\bar{p}}\sigma^{\bar{p}})/c_1.$$

Hence

$$\mu(A^* \setminus \bar{A}(z)) \leq \mu(E_\zeta) \leq c_0\zeta.$$

If we set

$$c_\mu = c_0(L_H\sigma_1^{1-\bar{p}} + M_F^{\bar{p}})/c_1$$

the proof is complete. □

The final result in this section is that if z is sufficiently near z^* then u can be set to u^* on $\bar{A}(z)$ in a well-defined way.

LEMMA 3.7. *Assume that Assumptions* 2.1–2.4 *hold. Then for all* $\bar{p} \in (0,1)$ *there is* σ_2 *such that if* $\|z - z^*\|_X < \sigma_2, z \neq z^*$ *and* $t \in \bar{A}(z)$ *then*

$$|u(t) - u^*(t)| < |u(t) - w(t)|$$

for all $w \neq u^*$, $w(t) \in \partial U(t)$. *Therefore the assignment of* u *to* $u_{\min}$ *or* $u_{\max}$, *whichever is closer to* u, *on* $\bar{A}$ *is well defined and decreases the* X*-norm of* $z - z^*$.

Proof. Let $\sigma_2 \leq \sigma_1$ so that the conclusions of Theorem 3.6 hold. If $t \in \bar{A}(z) \subset A^*$ then either $u^*(t) = u_{\max}(t)$ or $u^*(t) = u_{\min}(t)$. Without loss of generality we assume that $u^*(t) = u_{\max}(t)$. Letting $\|z - z^*\|_X = \sigma$ we have $|u(t) - u_{\max}(t)| \leq \sigma$ and

$$|u(t) - u_{\min}(t)| \geq |u^*(t) - u_{\min}(t)| - |u(t) - u^*(t)| \geq \nu - \sigma,$$

where ν is from Assumption 2.3. This completes the proof if $\sigma_2 < \nu/2$. □

4. The algorithm.

Let z be such that the conclusions of Theorem 3.5 hold and let $\|\mathcal{F}(z)\|_Y = \delta$. Let $\bar{p} \in (0,1)$ and let $\bar{A}(z)$ be given by (3.27). Let $\bar{I} = \bar{A}^c = [0,T]\backslash\bar{A}$.

Note that if $v = Q_{\bar{I}}z + (I - Q_{\bar{I}})(z - \mathcal{F}(z))$ then $v_u = u^*$ on $\bar{A}(z)$ and hence $\mathcal{F}_I$ can be computed since the value of z^* on $\bar{A}$ is known.

The variant of the projected Newton algorithm that we propose here makes the transition from a current iterate z_c to a new point z_+ by setting $u_c = u^*$ on $\bar{A}$, which is a well-defined step by Lemma 3.7, and then applying a projected Newton iteration with $\bar{A}(z_c)$ as the active set. The iteration is formally given by the following algorithm.

ALGORITHM 4.1. Algorithm proj_newt($\mathcal{F}, z_c, z_+$)

1. Compute

$$\bar{A}(z_c) = \{t \,|\, |H_u(z_c)(t)| \geq \|\mathcal{F}(z_c)\|_Y^{\bar{p}}\}.$$

2. Set $u_c = u^*$ on $\bar{A}(z_c)$.
3. Compute the projected Newton step

$$\begin{aligned} s &= -(I - Q_{\bar{I}})\mathcal{F}(z_c) - Q_{\bar{I}}\mathcal{F}'_{\bar{I}}(z_c)^{-1}Q_{\bar{I}}\mathcal{F}_{\bar{I}}(z_c) \\ &= -Q_{\bar{I}}\mathcal{F}'_{\bar{I}}(z_c)^{-1}Q_{\bar{I}}F(z_c) \\ &= -(Q_{\bar{I}}F'(z_c)Q_{\bar{I}})^{-1}Q_{\bar{I}}F(z_c). \end{aligned} \tag{4.1}$$

4. Set $z_{1/2} = \bar{P}(z_c + s)$ and $u_+ = u_{1/2}$, where

$$\bar{P} = \begin{pmatrix} I & 0 & 0 \\ 0 & I & 0 \\ 0 & 0 & \mathcal{P} \end{pmatrix}.$$

5. Compute

$$\begin{aligned} x_+(t) &= \int_0^t f(x_{1/2}(s), u_+(s), s)\, ds + x_0, \\ p_+(t) &= \int_t^T H_x(p_{1/2}(s), x_{1/2}(s), u_+(s))\, ds. \end{aligned}$$

Remark. The term $(I - Q_{\bar{I}})\mathcal{F}(z_c)$ in the right side of (4.1) in step 3 of Algorithm 4.1 vanishes because of the change in u_c on $\bar{A}$ in step 2 in the algorithm. We include it in the first line of (4.1) to emphasize the similarity of the algorithm we propose with the projected Newton algorithm in [2]. Hence after the overwriting of u_c with u^* on $\bar{A}$ in step 2 we need only compute the step on $\bar{I}$. Another effect of step 2 is the relation

$$Q_{\bar{I}}\mathcal{F}_{\bar{I}}(z_c) = Q_{\bar{I}}F(z_c),$$

which follows from $u = u^*$ on $\bar{A}$.

The convergence behavior of the iteration is given by the following theorem.

THEOREM 4.1. *Let the assumptions of Theorem* 3.5 *hold. There are* $K_N > 0$ *and* $\sigma_3 > 0$ *such that if* $\|e_c\|_X < \sigma_3$ *and* z_+ *is given by Algorithm* 4.1 *then*

$$\|e_+\|_X \leq K_N \|e_c\|_X^{2\bar{p}}. \tag{4.2}$$

Proof. First let $\sigma_3 \leq \sigma_2$ so that the conclusions of our previous results are valid. The proof begins by estimating $\|P_1 e_+\|_X$ in terms of $\|e_c\|_X$ and $\|P_2 e_+\|$. This first step reduces the proof to an estimate of $\|P_2 e_+\|_X$ in terms of $\|e_c\|$, which only involves the u-component of the error.

Step 5 of Algorithm 4.1 serves as a smoothing step. By definition we have $\|e_+\|_X = \|e_+\|_Y + \|\frac{d}{dt} P_1 e_+\|_Y$. Assumption 2.2 yields with a constant $B_{14} > 0$

$$\begin{aligned}\left\|\frac{d}{dt}P_1 e_+\right\|_Y &= \|f(x_{1/2}, u_+) - f(x^*, u^*)\|_{L^\infty} + \|H_x(p_{1/2}, x_{1/2}, u_+) - H_x(p^*, x^*, u^*)\|_{L^\infty} \\ &\leq B_{14}\|e_{1/2}\|_Y.\end{aligned}$$

Similarly,

$$\begin{aligned} e_{x+}(t) &= \int_0^t (f(x_{1/2}(s), u_+(s)) - f(x^*(s), u^*(s)))ds \quad \text{and} \\ e_{p+}(t) &= \int_t^T H_x(p_{1/2}(s), x_{1/2}(s), u_+(s)) - H_x(p^*(s), x^*(s), u^*(s)))ds. \end{aligned} \tag{4.3}$$

Hence there is B_{15} such that

$$\|P_1 e_+\|_Y \leq B_{15}\|e_{1/2}\|_Y.$$

Therefore

$$\|e_+\|_X \leq (B_{14} + B_{15})\|e_{1/2}\|_Y.$$

Step 2 of Algorithm 4.1 forces $u_c = u^*$ on $\bar{A} \subset A^*$ by Theorem 3.6. $Q_{\bar{I}}s$ can be viewed as a perturbation of the Newton step $\bar{s}$ for the map $Q_{\bar{I}}(\mathcal{F}_{\bar{I}}(z) - \mathcal{F}_{\bar{I}}(z^*))$ from the point z_c. We have

$$\bar{s} = -Q_{\bar{I}}\mathcal{F}'_{\bar{I}}(z_c)^{-1}Q_{\bar{I}}(\mathcal{F}_{\bar{I}}(z_c) - \mathcal{F}_{\bar{I}}(z^*))$$

and $\bar{z}_{1/2} = z_c + \bar{s}$ satisfies

$$\|Q_{\bar{I}}\bar{e}_{1/2}\|_X \leq B_{16}\|Q_{\bar{I}}e_c\|_X^2. \tag{4.4}$$

Here, for this proof,

$$B_{16} = K_\eta L_F/2$$

as is standard for Newton's method.

Now,

$$Q_{\bar I} e_{1/2} = Q_{\bar I} \bar e_{1/2} + \mathcal{F}'_{\bar I}(z_c)^{-1} \mathcal{F}_{\bar I}(z^*). \tag{4.5}$$

To estimate $\mathcal{F}'_{\bar I}(z_c)^{-1}\mathcal{F}_{\bar I}(z^*)$ we note that because $\bar A \subset A^*$, $I^* \subset \bar I$, $\mathcal{F}_{\bar I}(z^*)$ vanishes on I^*. $\mathcal{F}_{\bar I}(z^*)$ also vanishes on $\bar A \subset A^*$ by definition. Therefore

$$\mathcal{F}_{\bar I}(z^*) = P_2 \mathcal{F}_{\bar I}(z^*) = \begin{pmatrix} 0 \\ 0 \\ \chi_{\bar I \setminus I^*} H_u(p^*, x^*, u^*) \end{pmatrix}.$$

Hence, as in (3.9), we may require σ_3 to be small enough so that

$$\|P_1 \mathcal{F}'_{\bar I}(z_c)^{-1} \mathcal{F}_{\bar I}(z^*)\|_Y \le 2K_\eta \|\mathcal{F}_{\bar I}(z^*)\|_Y \mu(\bar I \setminus I^*). \tag{4.6}$$

We may estimate $\|\mathcal{F}_{\bar I}(z^*)\|_Y$ by using the fact that

$$\mathcal{F}_{\bar I}(z^*) = P_2 \mathcal{F}_{\bar I}(z^*)$$

and the definition of $\bar A$ to find

$$\begin{aligned} \|\mathcal{F}_{\bar I}(z^*)\|_Y &\le \|\mathcal{F}_{\bar I}(z^*) - \mathcal{F}_{\bar I}(z_c)\|_Y \|\mathcal{F}_{\bar I}(z_c)\|_Y \\ &\le M_F \|e_c\|_X + M_F^{\bar p} \|e_c\|_Y^{\bar p} \\ &\le (M_F \sigma_2^{1-\bar p} + M_F^{\bar p}) \|e_c\|_X^{\bar p}. \end{aligned}$$

The fact that $\bar I \setminus I^* = A^* \setminus \bar A$ and Theorem 3.6 imply

$$\mu(\bar I \setminus I^*) \le c_\mu \|e_c\|_X^{\bar p}. \tag{4.7}$$

From (4.5), (4.6), (4.7) we obtain

$$\begin{aligned} \|P_1 e_{1/2}\|_Y = \|P_1 Q_{\bar I} e_{1/2}\|_Y &\le \|P_1 Q_{\bar I} \bar e_{1/2}\|_Y + \|P_1 \mathcal{F}'_{\bar I}(z_c)^{-1} \mathcal{F}_{\bar I}(z^*)\|_Y \\ &\le B_{16} \|e_c\|_X^2 + K_\eta c_\mu (M_F \sigma_2^{1-\bar p} + M_F^{\bar p}) \|e_c\|_X^{2\bar p} \le B_{17} \|e_c\|_X^{2\bar p}, \end{aligned} \tag{4.8}$$

where

$$B_{17} = B_{16} \sigma_2^{2-2\bar p} + c_\mu K_\eta (M_F \sigma_2^{1-\bar p} + M_F^{\bar p}).$$

This gives

$$\|P_1 e_+\|_X \le B_{15} B_{17} \|e_c\|_X^{2\bar p} + \|P_2 e_+\|_X. \tag{4.9}$$

We must now consider the equation for the third component $P_2 z_+$ of the projected Newton iterate. If $t \in \bar A \subset A^*$ then $u_+ = u_c = u^*$ and the third components of the step, the current error, and the new error vanish.

By (4.1), on $\bar I$ we have

$$f_u(x_c, p_c, u_c) s_p + H_{ux}(x_c, p_c, u_c) s_x + H_{uu}(p_c, x_c, u_c) s_u = -H_u(p_c, x_c, u_c).$$

By Taylor's theorem and the fact that $f_u = H_{up}$ we have

$$\begin{aligned} H_{uu}(p_c, x_c, u_c) s_u &= -(H_u(p_c, x_c, u_c) + H_{up}(p_c, x_c, u_c) s_p + H_{ux}(p_c, x_c, u_c) s_x) \\ &= -H_u(p_{1/2}, x_{1/2}, u_c) - \Delta_1, \end{aligned}$$

where

$$\Delta_1 = \int_0^1 (H_{up}(p_c + ts_p, x_c + ts_x, u_c) - H_{up}(p_c, x_c, u_c))s_p \\ +(H_{ux}(p_c + ts_p, x_c + ts_x, u_c) - H_{ux}(p_c, x_c, u_c))s_x \, dt.$$

By Assumption 2.2

$$\|\Delta_1\|_\infty \leq 2M_H \|P_1 s\|_X^2,$$

where M_H is an upper bound for the Lipschitz constants of $f_u = H_{up}$, H_{ux}, and H_{uu}.

Now,

$$H_{uu}(p_c, x_c, u_c)s_u = -H_u(p^*, x^*, u_c) + \Delta_2,$$

where

$$\Delta_2 = H_u(p^*, x^*, u) - H_u(p_{1/2}, x_{1/2}, u) - \Delta_1.$$

Since

$$\|P_1 s\|_X \leq K_\eta \delta \leq K_\eta L_F \|e_c\|_X$$

and (4.8) implies

$$\|H_u(p_{1/2}, x_{1/2}, u) - H_u(p^*, x^*, u)\|_\infty \leq L_H \|P_1 e_{1/2}\|_Y \leq L_H B_{17} \|e_c\|_X^{2\bar{p}},$$

we have that Δ_2 can be bounded by

$$\|\Delta_2\|_\infty \leq B_{18} \|e_c\|_X^{2\bar{p}},$$

where

$$B_{18} = 2M_H K_\eta^2 L_F^2 \sigma_0^{2-2\bar{p}} + L_H B_{17}.$$

We expand $H_u(p^*, x^*, u_c)$ about u^* and apply Taylor's theorem again to obtain

$$H_u(p^*, x^*, u) = H_u(p^*, x^*, u^*) + H_{uu}(p^*, x^*, u^*)e_u + O(\|e_c\|_Y^2)$$

and hence

$$H_{uu}(p_c, x_c, u_c)s_u = -H_u(p^*, x^*, u^*) - H_{uu}(p^*, x^*, u^*)e_u + \Delta_3, \tag{4.10}$$

where

$$\|\Delta_3\|_\infty \leq B_{19} \|e_c\|_X^{2\bar{p}},$$

for some $B_{19} > 0$.

Let $\tilde{u}_+ = u_c + s_u$. Equation (4.10) may be rewritten as

$$\tilde{u}_+ = u^* - (H_{uu}(p_c, x_c, u_c))^{-1} H_u(p^*, x^*, u^*) \\ -[1 - (H_{uu}(p_c, x_c, u_c))^{-1} H_{uu}(p^*, x^*, u^*)]e_u + \Delta_3. \tag{4.11}$$

Since

$$(1 - H_{uu}(p_c, x_c, u_c))^{-1} H_{uu}(p^*, x^*, u^*))e_u = O(\|e_c\|_Y^2)$$

we have, for all $t \in \bar{I}$,

$$\tilde{u}_+ = u^* - (H_{uu}(p_c, x_c, u_c))^{-1} H_u(p^*, x^*, u^*) + \Delta_4, \tag{4.12}$$

where

$$\|\Delta_4\|_\infty \leq B_{20} \|e_c\|_X^{2\bar{p}}$$

for some $B_{20} > 0$.

Since $H_{uu}(p_c, x_c, u_c) > \tilde{\alpha}$ for all $t \in \bar{I}$ we have that

$$u^* = \mathcal{P}(u^* - (H_{uu}(p_c, x_c, u_c))^{-1} H_u(p^*, x^*, u^*)),$$

for all $t \in \bar{I}$. Therefore

$$\begin{aligned} u_+ = \mathcal{P}(\tilde{u}_+) &= \mathcal{P}(u^* - (H_{uu}(p_c, x_c, u_c))^{-1} H_u(p^*, x^*, u^*) + \Delta_4) \\ &= u^* + \Delta_5, \end{aligned}$$

where

$$\Delta_5 = \mathcal{P}(u^* - (H_{uu}(p_c, x_c, u_c))^{-1} H_u(p^*, x^*, u^*) + \Delta_4) - u^*$$

satisfies

$$\|\Delta_5\|_\infty \leq \|\Delta_4\|_\infty \leq B_{20} \|e_c\|_X^{2\bar{p}}.$$

Hence

$$\|P_2 e_+\|_X \leq B_{20} \|e_c\|_X^{2\bar{p}}. \tag{4.13}$$

We combine this with (4.9) to obtain

$$\begin{aligned} \|e_+\|_X &\leq \|P_1 e_+\|_X + \|P_2 e_+\|_X \\ &\leq B_{15} B_{17} \|e_c\|_X^{2\bar{p}} + 2\|P_2 e_+\|_X \\ &\leq (B_{15} B_{17} + 2B_{20}) \|e_c\|_X^{2\bar{p}}. \end{aligned}$$

Setting $K_N = B_{15} B_{17} + 2B_{20}$ completes the proof. □

5. Assumptions. In this section we review the assumptions posed in §2 and relate them to other conditions used in the context of optimal control problems with ordinary differential equations.

Since the theory developed uses an L^∞-framework in contrast to L^2, there is no problem in establishing the proper differentiability assumptions of the mappings. Recall that $\mathcal{F}_I : X_I \to Y_I$ for some measurable set $I \subset [0, T]$ with $A = [0, T] \setminus I$ is defined by

$$\mathcal{F}_I(z) = Q_I F \begin{pmatrix} p \\ x \\ \chi_I u + \chi_A u^* \end{pmatrix} = \begin{pmatrix} \dot{x} - f(x, \chi_I u + \chi_A u^*, t) \\ \dot{p} + H_x(x, \chi_I u + \chi_A u^*, t) \\ \chi_I(H_u(x, \chi_I u + \chi_A u^*, t)) \end{pmatrix} \tag{5.1}$$

for $z \in X_I$. Therefore the derivative is given by

$$\mathcal{F}_I'(z)(\zeta) = Q_I F' \begin{pmatrix} p \\ x \\ \chi_I u + \chi_A u^* \end{pmatrix} \begin{pmatrix} \pi \\ \xi \\ \chi_I \nu \end{pmatrix} = \begin{pmatrix} \dot{\xi} - f_x \xi - f_u \chi_I \nu \\ \dot{\pi} + f_x^T \pi + H_{xx} \xi + H_{xu} \chi_I \nu \\ \chi_I (f_u^T \pi + H_{ux} \xi + H_{uu} \chi_I \nu) \end{pmatrix}, \tag{5.2}$$

where $\zeta = (\pi, \xi, \nu) \in X$. In (5.2) we have omitted the arguments for the derivatives of the functions.

The regularity assumptions in Assumption 2.4 are related to second-order sufficiency conditions in optimal control. In the papers [15] and [14] second-order sufficiency conditions of the following type are used. A strengthened Legendre–Clebsch condition is posed with the existence of a solution to a Riccati equation, both appropriately altered to the case of control constraints. We assume the existence of a solution $Z(t) \in R^{n \times n}$ on [0,T] of the Riccati equation

$$-\dot{Z} = Z f_x + f_x^T Z + H_{xx} - (H_{xu} + Z f_u) H_{uu}^+ (H_{ux} + f_u^T Z), \quad Z(T) = 0, \tag{5.3}$$

where H_{uu}^+ is defined as

$$H_{uu}^+ = \chi_I H_{uu}^{-1} \chi_I.$$

LEMMA 5.1. *Let $z^* = (p^*, x^*, u^*) \in X$ be given. Assume that for I given by* (2.6) *there exists a solution $Z \in W_{n\times n}^{1,\infty}[0,T]$ of* (5.3) *and for some $\delta > 0$*

$$H_{uu}(t) \geq \delta \quad a.e. \text{ on } I. \tag{5.4}$$

Then (2.7) *and* (2.8) *of Assumption* 2.4 *hold.*

Proof. For given $(a,b,c)^T \in Y^\infty$ let γ be the solution of the initial value problem

$$\dot{\gamma} - (-f_x^T + H_{xu} H_{uu}^+ f_u^T)\gamma = b - H_{xu} H_{uu}^+ c - Z(a + f_u H_{uu}^+ c), \quad \gamma(T) = 0. \tag{5.5}$$

With γ known, denote by ξ the solution of

$$\dot{\xi} + (f_u H_{uu}^+ f_u^T Z - f_x + f_u H_{uu}^+ f_u^T)\xi = a + f_u H_{uu}^+ c - f_u H_{uu}^+ f_u^T \gamma, \quad \xi(0) = 0. \tag{5.6}$$

Define π by

$$\pi = Z\xi + \gamma \tag{5.7}$$

and ν on I by

$$\nu = H_{uu}^+ (c - f_u^T \pi - H_{ux} \xi), \tag{5.8}$$

which can be defined also as a function on $[0,T]$ by extension with 0.

Then one can verify that $(\pi, \xi, \nu)^T \in X_I$ solve the system

$$\mathcal{F}_I'(z^*)(\pi, \xi, \nu)^T = (a, b, c)^T, \tag{5.9}$$

which proves the surjectivity of $\mathcal{F}_I'(z^*)$. The continuous dependence of solutions of initial value problems on the right-hand side of the differential equation allows us to deduce from (5.5) with (5.4) that $\|\gamma\|_{W^{1,\infty}}$ depends continuously on $\|(a,b,c)\|_{Y_I}$. Using this fact one obtains the same statement from (5.6) for ξ. Finally, one estimates the L^∞-norm of ν by (5.8) and this altogether yields the estimate

$$\|\mathcal{F}_I'(z^*)^{-1}(a,b,c)\|_{X_I} = \|(\pi, \xi, \nu)\|_{X_I} \leq B_{21} \|(a,b,c)\|_{Y_I} \tag{5.10}$$

for some positive number B_{21}. This proves (2.7).

To show (2.8), let S be a measurable subset of $[0,T]$ and let $(a,b,c)^T \in Y^\infty$. Then

$$(I - Q_{S^c})(a,b,c)^T = (0,0,\chi_S c)$$

and let $(\pi,\xi,\nu)^T \in X_I$ be the solution of

$$\mathcal{F}_I'(z^*)(\pi,\xi,\nu)^T = (0,0,\chi_S c)^T = (I - Q_{S^c})(a,b,c)^T. \tag{5.11}$$

We define γ and ξ as solutions of (5.5) and (5.6), respectively, with (a,b,c) replaced by $(0,0,\chi_S c)$. Then we obtain from the modified (5.5) that for some constant B_{22} we have

$$\|\gamma\|_\infty \le B_{22}\|\chi_S \nu\|_1 \le B_{22} T \mu(S)\|\nu\|_\infty$$

and a similar estimate follows from modified (5.6) for $\|\xi\|_\infty$. Hence we obtain with (5.11)

$$\|P_1 \mathcal{F}_I'(z^*)^{-1}(I - Q_{S^c})(a,b,c)\|_{Y_I} = \|P_1(\pi,\xi,\nu)\|_{Y_I} = \|(\pi,\xi,0)\|_{Y_I} \le B_{23}\mu(S)\|\nu\|_\infty$$

which implies (2.8). □

The last condition in Assumption 2.4 is a trivial consequence of the assumption on the strengthened Legendre–Clebsch condition (5.4) in Lemma 5.1 if we choose I properly. We can relax the second-order sufficiency conditions to hold only on I^* under a proper smoothness assumption on the control.

LEMMA 5.2. *Let $z^* = (p^*, x^*, u^*) \in X$ be given such that $u^* \in C[0,T]$ and let Assumption 2.3 hold. Assume that for I^* we have for some $\delta > 0$*

$$H_{uu}(t) \ge \delta \quad a.e. \ on\ I^* \tag{5.12}$$

and that there exists a solution $Z \in W^{1,\infty}_{n\times n}[0,T]$ of (5.3) with $H^+_{uu} = \chi_{I^*} H^{-1}_{uu} \chi_{I^*}$. Then (2.7) and (2.8) of Assumption 2.4 hold.

Proof. Note that p^*, x^* are continuous as solutions of differential equations. With the assumption on u^* we have $H_{uu} \in C[0,T]$. Furthermore, Assumption 2.3 yields that for small ρ

$$I_\rho = \{t \in [0,T] \mid |H_u(p^*, x^*, u^*)| \le \rho\}$$

and we have that

$$\lim_{\rho \to 0} \mu(I_\rho \setminus I^*) = 0.$$

Since by (5.12) the continuous function H_{uu} is greater or equal to $\delta > 0$ on I^*, we can choose $\rho = \tilde{\sigma} > 0$ small enough so that for an appropriately small $\tilde{\alpha}$

$$H_{uu}(t) \ge \tilde{\alpha} > 0 \text{ on } I_{\tilde{\sigma}}.$$

We have assumed the existence of a solution of the Riccati equation where H^+_{uu} has support only on I^*. If $\tilde{\sigma}$ is small enough then the Riccati equation with H^+_{uu} and support on $I_{\tilde{\sigma}}$ also has a bounded solution on $[0,T]$. This completes the proof. □

Next we discuss the statements in Assumption 2.3. For the finite-dimensional case, a typical nondegeneracy condition would require that each component of $H_u(p^*, x^*, u^*)_i$

is nonzero if the corresponding component u_i^* of the optimal control lies in the active set A^*. The additional difficulty occurring for the infinite-dimensional problem is that H_u can approach zero in different ways. Here we had to impose a requirement that $|H_u|$ grows at a similar rate as the distance from the boundary of the active set when moving away from the boundary. Obviously, this condition reduces to the previously mentioned nondegeneracy condition in finite dimensions.

We can relate (2.4) to a condition on the zeroes of the switching function H_u. A similar condition was used in [5, Thm. 6.2] and also in [16, (2.14)]. To state this more precisely we prove the following lemma.

LEMMA 5.3. *Assume that for* $0 \le t_1 \le \cdots \le t_{2r+1} \le T$

$$\{t \in [0,T] \mid H_u(x^*,p^*,u^*)(t) = 0\} = \bigcup_{i=0}^{r} [t_{2i}, t_{2i+1}] \tag{5.13}$$

and that the function $g(t) := |H_u(x^*,p^*,u^*)(t)|$ *is continuous and has one-sided derivatives with*

$$g'_-(t_{2i}), g'_+(t_{2i+1}) > 0 \quad \text{for } i = 1,\dots,r. \tag{5.14}$$

Then the first line of (2.4) *in Assumption* 2.3 *holds.*

Proof. The assumption (5.13) yields that A^* consists of finitely many subintervals. Let $A_i^* = [t_{2i-1}, t_{2i}]$ denote such an interval. Then (5.14) implies that there are $\epsilon, m > 0$ such that

$$g(t) \ge m|t - t_{2i-1}| \text{ on } [t_{2i-1}, t_{2i-1} + \epsilon] \quad \text{and} \quad g(t) \ge m|t - t_{2i}| \text{ on } [t_{2i} - \epsilon, t_{2i}].$$

By definition g is positive on (t_{2i-1}, t_{2i}). Therefore, we can choose $\epsilon > 0$ so small that

$$g(t) \ge m\epsilon \text{ on } [t_{2i-1} + \epsilon, t_{2i} - \epsilon].$$

With $m^* = \min\{m, 2\epsilon/(t_{2i} - t_{2i-1})\}$ we obtain

$$g(t) \ge m|t - t_{2i-1}| \text{ on } [t_{2i-1}, (t_{2i-1}+t_{2i})/2] \text{ and } g(t) \ge m|t - t_{2i}| \text{ on } [(t_{2i-1}+t_{2i})/2, t_{2i}],$$

i.e., $g(t) \ge m \operatorname{dist}(t, \partial A^*)$ on A_i^*. □

To reconsider the second line of (2.4), we need more information about H_u which vanishes identically on I^*. The following lemma addresses a class of problems where the objective function contains a quadratic control term as is the case in many applications.

LEMMA 5.4. *Suppose that the objective function contains a quadratic control term*

$$L(x,u) = \bar{L}(x,u) + \frac{\alpha}{2}u^2,$$

with some $\alpha > 0$. *If the function* $f_u(x^*,u^*)^T p^* + \bar{L}_u(x^*,u^*)$ *has a nonzero slope when entering and leaving the active set, then the second line of* (2.4) *is also true.*

The form of L implies $H_u = f_u^T p + \bar{L}_u + \alpha u^*$ and

$$u^* = -\frac{1}{\alpha}(f_u(x^*,u^*)^T p^* + \bar{L}_u(x^*,u^*)) \text{ on } I^*.$$

The proof of this lemma is similar to the one given for Lemma 5.3.

6. Implementation. In this section we want to touch upon some of the details of the implementation of the algorithm. The unconstrained version of the algorithm presented in the previous sections can be given as the following system of nonlinear equations:

$$F(z) = F(p,x,u) = \begin{pmatrix} \dot{x} - f(x,u,t) \\ \dot{p} + H_x(p,x,u) \\ H_u(p,x,u) \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix} \tag{6.1}$$

for $z = (p,x,u)^T$ which satisfy the boundary conditions $x(0) = x_0$ and $p(T) = 0$.

The Fréchet derivative of F is given by

$$F'(p,x,u) \begin{pmatrix} \pi \\ \xi \\ \nu \end{pmatrix} = \begin{pmatrix} 0 & D - f_x & -f_u \\ D + f_x^T & H_{xx} & H_{xu} \\ f_u^T & H_{ux} & H_{uu} \end{pmatrix} \begin{pmatrix} \pi \\ \xi \\ \nu \end{pmatrix} \tag{6.2}$$

with $D = \frac{d}{dt}$ and all other components as multiplication operators.

If one considers the constrained optimal control problem with the corresponding system of nonlinear equations $\mathcal{F}(z) = 0$, then the projected Newton step requires a decision on the set of active indices. In the implementation we used the rule

$$\bar{A}(z) = \{t \in [0,T] : |H_u(z)(t)| \geq \|.5\mathcal{F}(z)\|_Y^{\bar{p}}\}.$$

Then the projected Newton step is computed by solving for (π,ξ,ν) with $\pi(T) = \xi(0) = 0$

$$\begin{aligned} Q_I F'(p,x,u) Q_I \begin{pmatrix} \pi \\ \xi \\ \nu \end{pmatrix} &= \begin{pmatrix} 0 & D - f_x & -f_u \chi_I \\ D + f_x^T & H_{xx} & H_{xu}\chi_I \\ \chi_I f_u^T & \chi_I H_{ux} & \chi_I H_{uu} \chi_I \end{pmatrix} \begin{pmatrix} \pi \\ \xi \\ \nu \end{pmatrix} \\ = -Q_I F(p,x,u) &= \begin{pmatrix} -\dot{x} + f(x,u,t) \\ -\dot{p} - f_x^T p - L_x \\ -\chi_I (f_u^T p + L_u) \end{pmatrix}, \end{aligned} \tag{6.3}$$

where

$$I = \{t \in [0,T] : |H_u(t)| < \|\mathcal{F}(z)\|_Y^{\bar{p}}\}.$$

The new control is then computed as

$$u_+ = \mathcal{P}(u + \chi_I \nu - \chi_A H_u).$$

The intermediate new state and adjoint variable are given by

$$(p_{1/2},\ x_{1/2}) = (p,\ x) + (\pi,\ \xi).$$

If one observes that in (6.3) π and ξ appear with derivatives, we can derive a differential equation for the sum $x_{1/2} = x + \xi$ and likewise for $p_{1/2}$. Hence, a more efficient way to compute $(p_{1/2}, x_{1/2})$ is to solve the following differential equations (all unsubscripted quantities are evaluated along the current values (p_c, x_c, u_c)):

$$\begin{aligned} \dot{x}_{1/2} - f_x x_{1/2} &= -f_x x + f + f_u \chi_I \nu, \\ \dot{p}_{1/2} + f_x^T p_{1/2} + H_{xx} x_{1/2} &= H_{xx} x - L_x - H_{xu} \chi_I \nu \end{aligned} \tag{6.4}$$

with boundary conditions $x_{1/2}(0) = x_0, p_{1/2}(T) = 0$. In addition, if we use the invertibility of $\chi_I H_{uu} \chi_I$ for all $t \in I$, then ν can be expressed in terms of $x_{1/2}, p_{1/2}, x, p$ as follows:

$$\chi_I \nu = -(H_{uu}^I)^+ (f_u^T p_{1/2} + H_{ux}(x_{1/2} - x) + L_u),$$

where we substitute

$$(H_{uu}^I)^+ = \begin{cases} (\chi_I H_{uu} \chi_I)^{-1}(t) & \text{if} \quad (\chi_I H_{uu} \chi_I)(t) \neq 0, \\ 0 & \text{if} \quad (\chi_I H_{uu} \chi_I)(t) = 0. \end{cases} \tag{6.5}$$

Hence a linear 2-point boundary value problem needs to be solved at each iteration. Solve for $p_{1/2}, x_{1/2}$ with $p_{1/2}(T) = 0, x_{1/2}(0) = 0$

$$\begin{aligned} &\dot{x}_{1/2} + (-f_x + f_u (H_{uu}^I)^+ H_{ux}) x_{1/2} + f_u (H_{uu}^I)^+ f_u^T p_{1/2} \\ &\quad = (-f_x + f_u (H_{uu}^I)^+ H_{ux}) x + f - f_u (H_{uu}^I)^+ L_u, \\ &\dot{p}_{1/2} + (f_x^T - H_{xu} (H_{uu}^I)^+ f_u^T) p_{1/2} + (H_{xx} - H_{xu} (H_{uu}^I)^+ H_{ux}) x_{1/2} \\ &\quad = (H_{xx} - H_{xu} (H_{uu}^I)^+ H_{ux}) x - L_x + H_{xu} (H_{uu}^I)^+ L_u. \end{aligned} \tag{6.6}$$

At the end of each iteration a smoothing step has to be carried out as follows:

$$\begin{aligned} x_+(t) &= \textstyle\int_0^t f(x_{1/2}(s), u_+(s), s)\, ds + x_0, \\ p_+(t) &= \textstyle\int_t^T H_x(p_{1/2}(s), x_{1/2}(s), u_+(s))\, ds. \end{aligned} \tag{6.7}$$

Termination and the identification of the active set is based on the size of the residual which can be computed as follows:

$$\begin{aligned} F(z_+) &= \begin{pmatrix} \dot{x}_+ - f(x_+, u_+, t) \\ \dot{p}_+ + H_x(p_+, x_+, u_+) \\ u_+ - \mathcal{P}(u_+ - H_u(p_+, x_+, u_+)) \end{pmatrix} \\ &= \begin{pmatrix} f(x_{1/2}, u_+, t) - f(x_+, u_+, t) \\ H_x(p_+, x_+, u_+) - H_x(p_{1/2}, x_{1/2}, u_+) \\ u_+ - \mathcal{P}(u_+ - H_u(p_+, x_+, u_+)) \end{pmatrix} \end{aligned} \tag{6.8}$$

We also list the size of the step $\|z_+ - z_c\|_X$ which we calculate by (the intermediate iterate $x_{1/2}$ has a corresponding iterate in the previous step denoted by $x_{-1/2}$)

TABLE 6.1
$\bar{p} = 0.6$.

k	ρ_k	ρ_k/ρ_{k-1}	$\rho_k/\rho_{k-1}^{2\bar{p}}$	σ_k
1	0.8669D+00	0.867	0.867	0.3308D+01
2	0.1746D+00	0.201	0.207	0.1331D+01
3	0.1054D+00	0.604	0.856	0.1656D+00
4	0.7004D-01	0.664	1.042	0.1706D+00
5	0.3290D-01	0.470	0.799	0.1463D+00
6	0.9885D-02	0.300	0.595	0.8485D-01
7	0.1604D-02	0.162	0.409	0.2830D-01
8	0.1328D-03	0.083	0.300	0.4816D-02
9	0.3456D-05	0.026	0.155	0.4380D-03

$$
\begin{aligned}
(6.9)\quad & \|z_+ - z_c\|_X \\
&= \max\{\|x_+ - x_c\|_\infty + \|\dot{x}_+ - \dot{x}_c\|_\infty, \|p_+ - p_c\|_\infty + \|\dot{p}_+ - \dot{p}_c\|_\infty, \|u_+ - u_c\|_\infty\} \\
&= \max\{\|x_+ - x_c\|_\infty + \|f(x_{1/2}, u_+, \cdot) - f(x_{-1/2}, u_c, \cdot)\|_\infty, \\
&\qquad \|p_+ - p_c\|_\infty + \|H_x(p_{1/2}, x_{1/2}, u_+) - H_x(p_{-1/2}, x_{-1/2}, u_c)\|_\infty, \|u_+ - u_c\|_\infty\}.
\end{aligned}
$$

We use the next example to illustrate the results. Let

$$
\begin{pmatrix} \dot{x}_1 \\ \dot{x}_2 \end{pmatrix} = \begin{pmatrix} (1 - x_2^2)x_1 - x_2 + u \\ x_1 \end{pmatrix}, \quad L(x, u) = \frac{1}{2}(x_1^2 + x_2^2 + u^2),
$$

and $T = 3, x(0) = (0, 1)^T$. Furthermore let

$$
0 \leq u(t) \leq 0.8,
$$

and starting data $x_0 \equiv (0, 1)^T, p_0 \equiv (0, 0)^T, u_0 \equiv 0$.

In Tables 6.1, 6.2, and 6.3 we tabulate, for different values of $\bar{p}$, the progress of the iteration for the example above. The 2-point boundary value problem (6.4) was solved with the trapezoid rule extrapolation approach used in [11] and the integration in (4.3) was done with the trapezoid rule. A uniform mesh of 1400 points was used.

For each iterate k we tabulate the norm of the nonlinear residual

$$
\rho = \|(\dot{x} - f, \dot{p} + H_x, u - \mathcal{P}(u - H_u))\|_\infty,
$$

the ratio $\rho_{k+1}/\rho_k^{\bar{p}}$, and the norm of the step $\sigma = \|(p_{k+1} - p_k, x_{k+1} - x_k, u_{k+1} - u_k)\|_\infty$.

We see from the tables that the convergence follows the predictions of Theorem 4.1 until the residual can be reduced no further as a result of truncation error effects. In several numerical experiments with different numbers of mesh points, a sharp increase in $\rho_{k+1}/\rho_k^{\bar{p}}$ seemed to be an indicator that truncation error effects were dominating the computation.

With Table 6.4 we want to illustrate the effect of Assumption 2.3 on the rate of convergence. If this assumption does not hold we are no longer guaranteed a local

Table 6.2
$\bar{p} = 0.75$.

k	ρ_k	ρ_k/ρ_{k-1}	$\rho_k/\rho_{k-1}^{2\bar{p}}$	σ_k
1	0.8669D+00	0.867	0.867	0.2135D+01
2	0.1746D+00	0.201	0.216	0.1331D+01
3	0.7894D-01	0.452	1.083	0.2370D+00
4	0.3058D-01	0.387	1.379	0.2007D+00
5	0.5309D-02	0.174	0.993	0.9241D-01
6	0.2431D-03	0.046	0.629	0.1671D-01
7	0.1638D-05	0.007	0.432	0.7524D-03
8	0.1691D-05	1.033	806.927	0.9862D-06
9	0.1691D-05	1.000	769.034	0.5488D-12

Table 6.3
$\bar{p} = 0.9$.

k	ρ_k	ρ_k/ρ_{k-1}	$\rho_k/\rho_{k-1}^{2\bar{p}}$	σ_k
1	0.8669D+00	0.867	0.867	0.2135D+01
2	0.1746D+00	0.201	0.226	0.1331D+01
3	0.6306D-01	0.361	1.460	0.3048D+00
4	0.1512D-01	0.240	2.187	0.1895D+00
5	0.9063D-03	0.060	1.715	0.4991D-01
6	0.3483D-05	0.004	1.044	0.2929D-02
7	0.1691D-05	0.485	11287.238	0.9399D-05
8	0.1691D-05	1.000	41449.044	0.3974D-11
9	0.1691D-05	1.000	41448.895	0.8188D-13

Table 6.4
$\bar{p} = 0.6$.

	$u_{\max}$=0.95		$u_{\max}$=0.8	
k	ρ_k	ρ_k/ρ_{k-1}	ρ_k	ρ_k/ρ_{k-1}
1	0.9107D+00		0.9107D+00	
2	0.1799D+00	0.198	0.2503D+00	0.275
3	0.6510D-02	0.036	0.1936D+00	0.774
4	0.4760D-02	0.731	0.1215D+00	0.628
5	0.2568D-02	0.540	0.5380D-01	0.443
6	0.9994D-03	0.389	0.1357D-01	0.252
7	0.2761D-03	0.276	0.1707D-02	0.126
8	0.5716D-04	0.207	0.1279D-03	0.075
9	0.9666D-05	0.169	0.8496D-05	0.066
10	0.2229D-05	0.231	0.4072D-05	0.479

superlinear rate of convergence as in Theorem 4.1. In Lemma 5.4 we give a sufficient condition for Assumption 2.3 to be true. For the example under consideration we have

$$f_u(x^*, u^*)^T p^* + \bar{L}_u(x^*, u^*) = p_1^*,$$

which looks like a parabola $-(t-2)^2 + 1$ with negative curvature. Here we impose only upper bounds on the controls. If the upper bound is relatively high, like 0.95, the slope of p_1^* at the boundary of the active set is small. The assumption in Lemma 5.4 is still satisfied but we can see a slower rate of convergence locally compared to an example where the upper bound on the control is set to 0.8, yielding a steeper slope of p_1^* at the boundary of the active set.

We have 1400 discretization points and select $\bar{p} = 0.6$.

Acknowledgments. The authors appreciate the work of the two very careful referees.

REFERENCES

[1] D. P. Bertsekas, *On the Goldstein–Levitin–Polyak gradient projection method*, IEEE Trans. Automat. Control, AC-21 (1976), pp. 174–184.

[2] ———, *Projected Newton methods for optimization problems with simple constraints*, SIAM J. Control Optim., 20 (1982), pp. 221–246.

[3] J. W. Daniel, *The Approximate Minimization of Functionals*, Prentice-Hall, Englewood Cliffs, NJ, 1971.

[4] V. F. Demyanov and A. M. Rubinov, *Approximate Methods in Optimization Theory*, Elsevier, New York, 1970.

[5] J. C. Dunn, *Rates of convergence for conditional gradient algorithms near singular and nonsingular extremals*, SIAM J. Control Optim., 17 (1979), pp. 187–211.

[6] ———, *Newton's method and the Goldstein step-length rule for constrained minimization problems*, SIAM J. Control Optim., 18 (1980), pp. 659–674.

[7] ———, *Global and asymptotic convergence rate estimates for a class of projected gradient processes*, SIAM J. Control Optim., 19 (1981), pp. 368–400.

[8] T. Tian and J. C. Dunn, *On the gradient projection method for optimal control problems with nonnegative L^2 inputs*, SIAM J. Control Optim., 32 (1994), pp. 516–537.

[9] J. C. Dunn and T. Tian, *Variants of the Kuhn-Tucker sufficient conditions in cones of nonnegative functions*, SIAM J. Control Optim., 30 (1992), pp. 1361 –1384.

[10] A. A. Goldstein, *Convex programming in Hilbert space*, Bull. Amer. Math. Soc., 70 (1964), pp. 709–710.

[11] C. T. Kelley and E. W. Sachs, *A pointwise quasi-Newton method for unconstrained optimal control problems*, Numer. Math., 55 (1989), pp. 159–176.

[12] ———, *Multilevel algorithms for constrained compact fixed point problems*, SIAM J. Sci. Comput., 15 (1994), pp. 645–667.

[13] E. S. Levitin and B. T. Polyak, *Constrained optimization methods*, USSR Comput. Math. Phys., 6 (1966), pp. 1–50.

[14] H. Maurer, *The Two-Norm Approach for Second Order Sufficiency Conditions in Mathematical Programming and Optimal Control*, Tech. report 6/92-N, Institut für Angewande Mathematik und Informatik, Universität Münster, 1992.

[15] D. Orrell and V. Zeidan, *Another Jacobi sufficiency criterion for optimal control with smooth constraints*, J. Optim. Theory Appl., 58 (1988), pp. 282–300.

[16] E. W. Sachs, *Convergence of algorithms for perturbed optimization problems*, Ann. Oper. Res., 27 (1990), pp. 311–342.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 33, No. 6, pp. 1758–1777, November 1995

© 1995 Society for Industrial and Applied Mathematics
007

ON THE ADAPTIVE CONTROL OF JUMP PARAMETER SYSTEMS VIA NONLINEAR FILTERING*

PETER E. CAINES† AND JI-FENG ZHANG‡

Abstract. In this paper we first present an error analysis for the process of estimates generated by the Wonham filter when it is used for the estimation of the (finite set-valued) jump-Markov parameters of a random parameter linear stochastic system and further give bounds on certain functions of these estimates. We then consider a certainty equivalence adaptive linear-quadratic Gaussian feedback control law using the estimates generated by the nonlinear filter and demonstrate the global existence of solutions to the resulting closed-loop system. A stochastic Lyapunov analysis establishes that the certainty equivalence law stabilizes the Markov jump parameter linear system in the mean square average sense. The conditions for this result are that certain products of (i) the parameter process jump rate and (ii) the solution of the control Riccati equation and its second derivatives should be less than certain given bounds. An example is given where the controlled linear system has state dimension 2. Finally, the stabilizing properties of certainty equivalence laws which depend on (i) the maximum likelihood estimate of the parameter value and (ii) a modified version of this estimate are established under certain conditions.

Key words. jump parameter, nonlinear filter, adaptive control, stochastic systems, maximum likelihood

AMS subject classifications. 93E11, 93E15, 93E35

1. Introduction.

The hybrid system considered in this work is taken to have the following form:

$$dx_t = [A(\theta_t)x_t + B(\theta_t)u_t]dt + dw_t, \tag{1.1}$$

where $x_t \in I\!R^n$ and $u_t \in I\!R^m$ are the state and input of the system, $\{w_t, \mathcal{F}_t\}$ is a standard Wiener process in $I\!R^n$ with respect to a probability space $(\Omega, P, \mathcal{F})$, and $\theta_t \in \{1, 2, \dots, N\}$ is the N-state jump-Markov parameter process subject to

$$\Phi_t = \Phi_0 + \Pi \int_0^t \Phi_s ds + m_t. \tag{1.2}$$

Here, $\Phi_t = [1\!\!1_{\{\theta_t=1\}}, 1\!\!1_{\{\theta_t=2\}}, \dots, 1\!\!1_{\{\theta_t=N\}}]^\tau$ is the indicator process for θ_t, Π is the transition probability rate matrix, m_t is a zero-mean L^2 martingale, measurable with respect to an increasing σ-field $\mathcal{F}_t$. Φ_0 is $\mathcal{F}_0$-measurable and $E\Phi_0 = p_0$.

For $\theta = i$, $A(\theta) = A_i$, and $B(\theta) = B_i$, where the A_i's and B_i's are, respectively, $I\!R^{n\times n}$ and $I\!R^{n\times m}$ matrices such that $\|A_i - A_j\| + \|B_i - B_j\| \neq 0$ for $i \neq j$. Here and hereafter, $\|X\| \triangleq [\lambda_{\max}(X^\tau X)]^{1/2}$, where $\lambda_{\max}(A)$ denotes the largest eigenvalue of a matrix A.

The model (1.1), (1.2) is particularly appropriate for the analysis of the control of time varying systems, since (1.1) has a variable structure. As indicated by the dependence of all matrix parameters on the indicator process Φ_t, it can be used as

* Received by the editors October 14, 1993; accepted for publication (in revised form) May 27, 1994. This research was supported by Natural Sciences and Engineering Research Council of Canada grant A 1329 and the National Natural Science Foundation of China.

† Department of Electrical Engineering, McGill University, 3480 University Street, Montreal H3A 2A7, Canada, and Canadian Institute for Advanced Research.

‡ Department of Electrical Engineering, McGill University, 3480 University Street, Montreal H3A 2A7, Canada, and Institute of Systems Science, Academia Sinica, Beijing 100080, People's Republic of China.

a model for systems subject to random failures and structural changes. Moreover, (1.2) is a general model for jump-Markov parameter processes (see, e.g., Liptser and Shiryaev (1977)).

Control problems for such systems in a nonadaptive setting have been the subject of considerable theoretical research for the past two decades and Sworder and Chou (1985) and Ezzine and Haddad (1989) have given surveys of previous work on this topic.

Generally speaking, the previous works can be classified into three groups: one group (see, e.g., Sworder and Chou (1985); Ezzine and Haddad (1989); Mariton and Bertrand (1985); Mariton (1986); Ji and Chizeck (1990); Feng, Loparo, Ji, and Chizeck (1992)) deals with the case where the system state process x and the jump parameter process Φ can be observed completely at any time instant. The second group (see, e.g., Wonham (1965), Rishel (1981), Caines and Chen (1985), Chen and Caines (1989), Helmes and Rishel (1990), Caines and Nassiri-Toussi (1991)) is concerned with the adaptive case where the system state process x can be observed, but the jump parameter process Φ cannot be directly observed and is consequently estimated. This may, for instance, be carried out by an application of the Wonham filter (see, e.g., Caines and Chen (1985), Chen and Caines (1989), Caines and Nassiri-Toussi (1991)). The third group (see, e.g., Sworder (1991)) discusses the adaptive case where neither the system state process x nor the jump parameter process Φ can be observed.

Among the first group, it is worth mentioning that Ji and Chizeck (1990) and Feng, Loparo, Ji, and Chizeck (1992) examine the relationship between appropriately defined controllability and stabilizability properties, and establish necessary and sufficient conditions for (i) system stabilization and (ii) infinite time jump linear quadratic (JLQ) optimal controls to exist. However, in most situations, direct observation of system parameters is impossible and this leads to the use of adaptive control. Caines and Chen (1985) used the Wonham filter and a dynamic programming approach to obtain a finite-horizon adaptive optimal control law for a general jump-Markov system. In a continuation of this work, Caines and Nassiri-Toussi (1991) and Nassiri-Toussi and Caines (1991) carried out a stochastic Lyapunov analysis of a certainty equivalence stabilizing control law and gave an analysis of the resulting ergodic behavior of the system. It is shown that, under rather strong conditions on the magnitude of the jumps of the parameters and the rate of the jump parameter process, a certainty equivalence linear feedback regulator (using the parameter estimates generated by the Wonham filter) gives rise to stable ergodic behavior of the system (1.1), (1.2). In some special cases, where the system is deterministic or where indirect observations of the parameter are available, special solutions to this problem have also been given in Sworder and Chou (1985), while the general adaptive control problem for stochastic jump-Markov parameter systems is addressed in Rishel (1981), Caines and Chen (1985), Chen and Caines (1989), Helmes and Rishel (1990), Sworder (1991), Caines and Nassiri-Toussi (1991), Nassiri-Toussi and Caines (1991), and Dufour and Bertrand (1993). It should be remarked that Rishel (1981) was the first to use the Wonham filter to find the equations of the optimal linear quadratic Gaussian (LQG) controller for a system depending upon a (constant in time) unobserved finite set-valued random variable. More recently, Helmes and Rishel (1990) have given an explicit solution to this problem for the case of minimizing the expectation of the quadratic state deviation at a final time plus the integrated square of the control action. Sworder (1991) presents an approximation to the quadratic-optimal regulator problem for a situation in which there is an unconventional measurement architecture; the solution is in a form quite similar to that obtained in the complete observation case, but the gain equation is made more complicated by the presence of noise. Finally, in a recent paper, Dufour and Bertrand (1993) responded to an announcement (Caines and Zhang (1992)) of the results of the present paper by giving a form of averaged control law (with respect to the conditional densities) that adaptively stabilizes the jump parameter system in question whenever it satisfies a simple set of algebraic sufficient conditions.

The object of this paper is to establish the existence of stabilizing adaptive feedback controllers for jump parameter systems under relatively weak conditions.

In §2 of this paper, the Wonham filter for estimating the indicator process Φ from observations on x and u is presented, and the error behavior of the filter is analyzed. Theorem 2.1 gives a formula for the mean square estimation error of Φ and Corollaries 2.1 and 2.2 give bounds for the expectation of certain weighted integrals of the estimates; these are required in the subsequent stability analysis. Section 3 contains the principal adaptive control result of the paper. By use of a stochastic Lyapunov technique it is shown that an adaptive LQG certainty equivalence feedback control law, which employs parameter estimates generated by the nonlinear filter, stabilizes the system in an average mean square sense. This result is subject to the condition that (i) the rate of the jump process of the system and (ii) the magnitude of the solution to the control Riccati equation and its second derivative are such that two products of these quantities fall below specified bounds (see (3.8)). It is to be noted that there is no condition on the size of the jumps of the parameters. In §4, a nontrivial example of this theory is given concerning the adaptive control of a two-dimensional linear system with jump-Markov system matrices $\{A_i,\ 1 \leq i \leq N\}$. Finally, in §5, the stabilizing properties of certainty equivalence laws which depend on (i) the maximum likelihood estimate of the parameter value and (ii) a modified version of this estimate are established under certain conditions.

2. The nonlinear filter and preliminary results.

Suppose that (i) A_i and B_i are known for $i = 1, \dots, N$, (ii) $E\|x_0\|^2 < \infty$, (iii) the cross quadratic variation of m and w, i.e., $d\langle m, w\rangle_t/dt \equiv 0$, and (iv) u_t is an m-dimensional $\mathcal{F}_t^x \triangleq \sigma\{x_s,\ s \leq t\}$-measurable control process. Set

$$\widehat{\Phi}_t = [\widehat{\Phi}_t(1), \dots, \widehat{\Phi}_t(N)]^\tau \triangleq E(\Phi_t|\mathcal{F}_t^x), \quad \forall t \geq 0, \tag{2.1}$$

$$H_t = [A_1 x_t + B_1 u_t, \dots, A_N x_t + B_N u_t], \tag{2.2}$$

and

$$\mathrm{Diag}\widehat{\Phi}_t = \begin{bmatrix} \widehat{\Phi}_t(1) & & 0 \\ & \ddots & \\ 0 & & \widehat{\Phi}_t(N) \end{bmatrix}. \tag{2.3}$$

Then the nonlinear Wonham filter for the values of the parameter indicator process Φ_t is given by (see, e.g., Chen and Caines (1989))

$$d\widehat{\Phi}_t = \Pi \widehat{\Phi}_t dt + (\mathrm{Diag}\widehat{\Phi}_t - \widehat{\Phi}_t \widehat{\Phi}_t^\tau) H_t^\tau d\overline{w}_t, \tag{2.4}$$

where $\{\overline{w}_t,\ \mathcal{F}_t^x\}$ is the Wiener process of innovations defined by the innovation representation of x_t:

$$d\overline{w}_t = dx_t - H_t \widehat{\Phi}_t dt. \tag{2.5}$$

THEOREM 2.1. *The conditional mean square estimation error of the filter* (2.4) *for the system* (1.1) *satisfies*

$$
\begin{aligned}
E\|\widetilde{\Phi}_t\|^2 = E\|\widetilde{\Phi}_0\|^2 &+ 2E\int_0^t \widetilde{\Phi}_s^\tau \Pi \widetilde{\Phi}_s ds - 2E\int_0^t \Phi_s^\tau \Pi \Phi_s ds \\
&-E\int_0^t \mathrm{Tr}\left([\widehat{\Phi}_s\widehat{\Phi}_s^\tau - \mathrm{Diag}\widehat{\Phi}_s](H_s^\tau H_s)[\widehat{\Phi}_s\widehat{\Phi}_s^\tau - \mathrm{Diag}\widehat{\Phi}_s]\right) ds, \qquad (2.6)
\end{aligned}
$$

where $\widetilde{\Phi}_t \overset{\triangle}{=} \Phi_t - \widehat{\Phi}_t$, *and* $\mathrm{Tr}(X)$ *denotes the trace of matrix* X.

Proof. By (2.2), (1.1) can be rewritten as

$$dx_t = H_t\Phi_t dt + dw_t,$$

which together with (2.5) results in

$$d\overline{w}_t = H_t\widetilde{\Phi}_t dt + dw_t.$$

Therefore, by (1.2) and (2.4), we have

$$
\begin{aligned}
d\widetilde{\Phi}_t &= \Pi\widetilde{\Phi}_t dt + [\widehat{\Phi}_t\widehat{\Phi}_t^\tau - \mathrm{Diag}\widehat{\Phi}_t]H_t^\tau d\overline{w}_t + dm_t \\
&= \Pi\widetilde{\Phi}_t dt + [\widehat{\Phi}_t\widehat{\Phi}_t^\tau - \mathrm{Diag}\widehat{\Phi}_t]H_t^\tau H_t\widetilde{\Phi}_t dt \\
&\quad +[\widehat{\Phi}_t\widehat{\Phi}_t^\tau - \mathrm{Diag}\widehat{\Phi}_t]H_t^\tau dw_t + dm_t,
\end{aligned}
$$

which combined with Ito's formula (see, e.g., Schwartz (1984)) leads to

$$
\begin{aligned}
\widetilde{\Phi}_t^\tau\widetilde{\Phi}_t = \widetilde{\Phi}_0^\tau\widetilde{\Phi}_0 &+ 2\int_0^t \widetilde{\Phi}_s^\tau\Pi\widetilde{\Phi}_s ds \\
&+2\int_0^t \widetilde{\Phi}_s^\tau(\widehat{\Phi}_s\widehat{\Phi}_s^\tau - \mathrm{Diag}\widehat{\Phi}_s)H_s^\tau H_s\widetilde{\Phi}_s ds \\
&+2\int_0^t \widetilde{\Phi}_s^\tau[\widehat{\Phi}_s\widehat{\Phi}_s^\tau - \mathrm{Diag}\widehat{\Phi}_s]H_s^\tau dw_s + 2\int_0^t \widetilde{\Phi}_s^\tau dm_s \\
&+\int_0^t \mathrm{Tr}\left([\widehat{\Phi}_s\widehat{\Phi}_s^\tau - \mathrm{Diag}\widehat{\Phi}_s](H_s^\tau H_s)[\widehat{\Phi}_s\widehat{\Phi}_s^\tau - \mathrm{Diag}\widehat{\Phi}_s]\right) ds \\
&+\sum_{0<s\leq t}(\widetilde{\Phi}_s - \widetilde{\Phi}_{s-})^\tau(\widetilde{\Phi}_s - \widetilde{\Phi}_{s-}). \qquad (2.7)
\end{aligned}
$$

Since $\widehat{\Phi}_t$, as a solution of (2.4), is continuous, $(\widetilde{\Phi}_s - \widetilde{\Phi}_{s-}) = \Phi_s - \Phi_{s-}$. From this we see that

$$\sum_{0<s\leq t}(\widetilde{\Phi}_s - \widetilde{\Phi}_{s-})^\tau(\widetilde{\Phi}_s - \widetilde{\Phi}_{s-}) = \sum_{0<s\leq t}(\Phi_s - \Phi_{s-})^\tau(\Phi_s - \Phi_{s-}) = 2J_t, \qquad (2.8)$$

where J_t is the number of the jump points of Φ_s in $[0, t]$.

Substituting (2.8) into (2.7) and taking expectations on both sides, we see that

$$
\begin{aligned}
E\widetilde{\Phi}_t^\tau\widetilde{\Phi}_t = E\widetilde{\Phi}_0^\tau\widetilde{\Phi}_0 &+ 2E\int_0^t \widetilde{\Phi}_s^\tau\Pi\widetilde{\Phi}_s ds + 2EJ_t \\
&+2E\int_0^t \widetilde{\Phi}_s^\tau(\widehat{\Phi}_s\widehat{\Phi}_s^\tau - \mathrm{Diag}\widehat{\Phi}_s)H_s^\tau H_s\widetilde{\Phi}_s ds \\
&+E\int_0^t \mathrm{Tr}\left([\widehat{\Phi}_s\widehat{\Phi}_s^\tau - \mathrm{Diag}\widehat{\Phi}_s](H_s^\tau H_s)[\widehat{\Phi}_s\widehat{\Phi}_s^\tau - \mathrm{Diag}\widehat{\Phi}_s]\right) ds. \qquad (2.9)
\end{aligned}
$$

From (1.2) and Ito's formula it follows that

$$\Phi_t^\tau \Phi_t = \Phi_0^\tau \Phi_0 + 2\int_0^t \Phi_s^\tau \Pi \Phi_s ds + 2\int_0^t \Phi_s^\tau dm_s + 2J_t,$$

which, together with $\Phi_t^\tau \Phi_t = \Phi_0^\tau \Phi_0 = 1$, implies

$$EJ_t = -E\int_0^t \Phi_s^\tau \Pi \Phi_s ds. \tag{2.10}$$

Notice that

$$E(\widetilde{\Phi}_t \widetilde{\Phi}_t^\tau | \mathcal{F}_t^x) = E(\Phi_t \Phi_t^\tau | \mathcal{F}_t^x) - \widehat{\Phi}_t \widehat{\Phi}_t^\tau = \mathrm{Diag}\widehat{\Phi}_t - \widehat{\Phi}_t \widehat{\Phi}_t^\tau.$$

Then, by (2.9) and (2.10), we can conclude that

$$\begin{aligned} E\widetilde{\Phi}_t^\tau \widetilde{\Phi}_t &= E\widetilde{\Phi}_0^\tau \widetilde{\Phi}_0 + 2E\int_0^t \widetilde{\Phi}_s^\tau \Pi \widetilde{\Phi}_s ds - 2E\int_0^t \Phi_s^\tau \Pi \Phi_s ds \\ &\quad -E\int_0^t \mathrm{Tr}\left([\widehat{\Phi}_s \widehat{\Phi}_s^\tau - \mathrm{Diag}\widehat{\Phi}_s](H_s^\tau H_s)[\widehat{\Phi}_s \widehat{\Phi}_s^\tau - \mathrm{Diag}\widehat{\Phi}_s]\right) ds, \end{aligned}$$

i.e., (2.6) holds. □

COROLLARY 2.1. (2.6) implies that

$$E\int_0^t \sum_{i=1}^N [\widehat{\Phi}_s(i)]^2 \| A_{\widehat{\Phi}_s} x_s + B_{\widehat{\Phi}_s} u_s - A_i x_s - B_i u_s \|^2 ds \le 1 + 4\|\Pi\| t, \tag{2.11}$$

where $\widehat{\Phi}_t$ and $\widehat{\Phi}_t(i)$ are defined in (2.1), and

$$A_{\widehat{\Phi}_s} = \sum_{i=1}^N \widehat{\Phi}_s(i) A_i, \qquad B_{\widehat{\Phi}_s} = \sum_{i=1}^N \widehat{\Phi}_s(i) B_i. \tag{2.12}$$

Proof. Let

$$H_{t,i} = A_i x_t + B_i u_t \quad \text{and} \quad H_{t,\widehat{\Phi}_t} = A_{\widehat{\Phi}_t} x_t + B_{\widehat{\Phi}_t} u_t. \tag{2.13}$$

Then, by (2.1)–(2.3), we have

$$\begin{aligned} H_t[\widehat{\Phi}_t \widehat{\Phi}_t^\tau - \mathrm{Diag}\widehat{\Phi}_t] &= [H_{t,\widehat{\Phi}_t} \widehat{\Phi}_t(1), \ldots, H_{t,\widehat{\Phi}_t} \widehat{\Phi}_t(N)] - H_t \mathrm{Diag}\widehat{\Phi}_t \\ &= [(H_{t,\widehat{\Phi}_t} - H_{t,1})\widehat{\Phi}_t(1), \ldots, (H_{t,\widehat{\Phi}_t} - H_{t,N})\widehat{\Phi}_t(N)]. \end{aligned}$$

Thus, by (2.13) and (2.6) we get

$$\begin{aligned} & E\int_0^t \sum_{i=1}^N [\widehat{\Phi}_s(i)]^2 \| A_{\widehat{\Phi}_s} x_s + B_{\widehat{\Phi}_s} u_s - A_i x_s - B_i u_s \|^2 ds \\ & = E\int_0^t \sum_{i=1}^N [\widehat{\Phi}_s(i)]^2 \| H_{s,\widehat{\Phi}_s} - H_{s,i} \|^2 ds \\ & = E\int_0^t \mathrm{Tr}\left([\widehat{\Phi}_s \widehat{\Phi}_s^\tau - \mathrm{Diag}\widehat{\Phi}_s](H_s^\tau H_s)[\widehat{\Phi}_s \widehat{\Phi}_s^\tau - \mathrm{Diag}\widehat{\Phi}_s]\right) ds \\ & \le E\widetilde{\Phi}_0^\tau \widetilde{\Phi}_0 + 2E\int_0^t \widetilde{\Phi}_s^\tau \Pi \widetilde{\Phi}_s ds - 2E\int_0^t \Phi_s^\tau \Pi \Phi_s ds \\ & \le 1 + 4\|\Pi\| t, \end{aligned} \tag{2.14}$$

where we have used the fact that $E\|\widetilde{\Phi}_t\|^2 = 1 - E\|\widehat{\Phi}_t\|^2 \le 1$ and $\|\Phi_t\|^2 = 1$ for $t \ge 0$ in order to get the last inequality.

This completes the proof of Corollary 2.1. □

COROLLARY 2.2. *For any constant $\eta > 0$, we have*

$$\begin{aligned} &E\int_0^T \|x_t\|^2 \sum_{i=1}^N [\widehat{\Phi}_t(i)]^2 \|A_{\widehat{\Phi}_t} x_s + B_{\widehat{\Phi}_t} u_s - A_i x_t - B_i u_t\|^2 dt \\ &\quad \le E\|x_0\|^2 + (\eta + 2\|\Pi\|)E\int_0^T \|x_t\|^2 dt + 2E\int_0^T |x_t^\tau H_t \widehat{\Phi}_t| dt \\ &\qquad +4\eta^{-1}(1 + 4\|\Pi\|T) + NT. \end{aligned} \tag{2.15}$$

Proof. From Ito's formula and (2.4) it follows that

$$\begin{aligned} d(\widehat{\Phi}_t^\tau \widehat{\Phi}_t) &= 2\widehat{\Phi}_t^\tau \Pi \widehat{\Phi}_t dt + 2\widehat{\Phi}_t^\tau \left(\mathrm{Diag}\widehat{\Phi}_t - \widehat{\Phi}_t \widehat{\Phi}_t^\tau\right) H_t^\tau d\overline{w}_t \\ &\quad + \mathrm{Tr}\left[\left(\mathrm{Diag}\widehat{\Phi}_t - \widehat{\Phi}_t\widehat{\Phi}_t^\tau\right) H_t^\tau H_t \left(\mathrm{Diag}\widehat{\Phi}_t - \widehat{\Phi}_t\widehat{\Phi}_t^\tau\right)\right] dt, \end{aligned}$$

and from (2.5),

$$d(x_t^\tau x_t) = 2x_t^\tau H_t \widehat{\Phi}_t dt + 2x_t^\tau d\overline{w}_t + N dt.$$

Therefore, by Ito's formula we have

$$\begin{aligned} &d[(1 - \widehat{\Phi}_t^\tau \widehat{\Phi}_t) x_t^\tau x_t] \\ &\quad = -x_t^\tau x_t \mathrm{Tr}\left[\left(\mathrm{Diag}\widehat{\Phi}_t - \widehat{\Phi}_t\widehat{\Phi}_t^\tau\right) H_t^\tau H_t \left(\mathrm{Diag}\widehat{\Phi}_t - \widehat{\Phi}_t\widehat{\Phi}_t^\tau\right)\right] dt \\ &\qquad -2x_t^\tau x_t \widehat{\Phi}_t^\tau \Pi \widehat{\Phi}_t dt + 2(1 - \widehat{\Phi}_t^\tau \widehat{\Phi}_t) x_t^\tau H_t \widehat{\Phi}_t dt + N(1 - \widehat{\Phi}_t^\tau \widehat{\Phi}_t) dt \\ &\qquad -4\widehat{\Phi}_t^\tau \left(\mathrm{Diag}\widehat{\Phi}_t - \widehat{\Phi}_t\widehat{\Phi}_t^\tau\right) H_t^\tau x_t dt + 2(1 - \widehat{\Phi}_t^\tau \widehat{\Phi}_t) x_t^\tau d\overline{w}_t \\ &\qquad -2x_t^\tau x_t \widehat{\Phi}_t^\tau \left(\mathrm{Diag}\widehat{\Phi}_t - \widehat{\Phi}_t\widehat{\Phi}_t^\tau\right) H_t^\tau d\overline{w}_t. \end{aligned}$$

Taking expectations of both sides, and noticing $0 < \|\widehat{\Phi}_t\| \le 1$, $(1 - \widehat{\Phi}_t^\tau \widehat{\Phi}_t) x_t^\tau x_t \ge 0$, and $4ab \le 4\eta^{-1}a^2 + \eta b^2$ (for all $a, b \ge 0, \eta > 0$) we get that for any fixed constant $\eta > 0$

$$\begin{aligned} &E\int_0^T \|x_t\|^2 \mathrm{Tr}\left[\left(\mathrm{Diag}\widehat{\Phi}_t - \widehat{\Phi}_t\widehat{\Phi}_t^\tau\right) H_t^\tau H_t \left(\mathrm{Diag}\widehat{\Phi}_t - \widehat{\Phi}_t\widehat{\Phi}_t^\tau\right)\right] dt \\ &\quad \le E\|x_0\|^2 + 2\|\Pi\| E\int_0^T \|x_t\|^2 dt + 2E\int_0^T |x_t^\tau H_t \widehat{\Phi}_t| dt + NT \\ &\qquad +4E\int_0^T \left\|\left(\mathrm{Diag}\widehat{\Phi}_t - \widehat{\Phi}_t\widehat{\Phi}_t^\tau\right) H_t^\tau x_t\right\| dt \\ &\quad \le E\|x_0\|^2 + (\eta + 2\|\Pi\|)E\int_0^T \|x_t\|^2 dt \\ &\qquad +2E\int_0^T |x_t^\tau H_t \widehat{\Phi}_t| dt + NT \\ &\qquad +4\eta^{-1} E\int_0^T \mathrm{Tr}\left[\left(\mathrm{Diag}\widehat{\Phi}_t - \widehat{\Phi}_t\widehat{\Phi}_t^\tau\right) H_t^\tau H_t \left(\mathrm{Diag}\widehat{\Phi}_t - \widehat{\Phi}_t\widehat{\Phi}_t^\tau\right)\right] dt, \end{aligned}$$

which, together with (2.14), leads to

$$E\int_0^T \|x_t\|^2 \mathrm{Tr}\left[\left(\mathrm{Diag}\widehat{\Phi}_t - \widehat{\Phi}_t\widehat{\Phi}_t^\tau\right) H_t^\tau H_t \left(\mathrm{Diag}\widehat{\Phi}_t - \widehat{\Phi}_t\widehat{\Phi}_t^\tau\right)\right] dt$$
$$\leq E\|x_0\|^2 + (\eta + 2\|\Pi\|)E\int_0^T \|x_t\|^2 dt + 2E\int_0^T |x_t^\tau H_t \widehat{\Phi}_t| dt$$
$$+4\eta^{-1}(1 + 4\|\Pi\|T) + NT, \quad \forall\eta > 0,$$

i.e., (2.15) holds. □

3. Quadratic index-based adaptive control. The following lemma is to be found in Caines and Nassiri-Toussi (1991).

LEMMA 3.1. *Let the Markov process X_t satisfy the following regular Ito stochastic differential equation:*

$$dX_t = b_t(X_t)dt + G_t(X_t)dw_t. \tag{3.1}$$

Furthermore, assume that there exist a $C^1(\mathbb{R}^+) \times C^2(\mathbb{R}^n)$ nonnegative function $V_{\cdot}(\cdot)$, a positive real number α_0, and a nonnegative function k_t, such that

$$\frac{\partial V_t(x)}{\partial t} + \mathcal{A}V_t(x) \leq -\alpha_0\|x\|^2 + k_t, \quad \forall x \in \mathbb{R}^n, \quad \forall t \geq 0,$$

where $\mathcal{A}$ is the infinitesimal generator of (3.1).

Then, if

$$\lim\ \mathrm{sup}_{t\to\infty}\frac{1}{t}E\int_0^t k_s ds < \infty \quad \text{and} \quad E[V_0(X_0)] < \infty,$$

$$\limsup_{t\to\infty}\frac{1}{t}E\int_0^t \|X_s\|^2 ds \leq \limsup_{t\to\infty}\frac{1}{\alpha_0 t}E\int_0^t k_s ds < \infty. \tag{3.2}$$

Proof. By (3.1) and Ito's formula, we know that $dV_t(X_t)$ satisfies the following equality:

$$dV_t(X_t) = \ \frac{\partial V_t(x)}{\partial t} + \mathcal{A}V_t(x)\Big) dt + \frac{\partial V_t(x)}{\partial x}G_t(X_t)dw_t.$$

With the assumptions on $V_t(X_t)$, this results in

$$V_t(X_t) \leq V_0(X_0) - \alpha_0\int_0^t \|X_s\|^2 ds + \int_0^t k_s ds + \int_0^t \frac{\partial V_s(x)}{\partial x}G_s(X_s)dw_s.$$

Taking the expectation of both sides of this inequality we get

$$E[V_t(X_t)] - E[V_0(X_0)] \leq -\alpha_0 E\int_0^t \|X_s\|^2 ds + E\int_0^t k_s ds.$$

This, combined with the positiveness of V_t, gives the desired result (3.2). □

It is well known that if (A_α, B_α) is controllable and (A_α, C) is observable (with $C^\tau C = Q$), then for all $S > 0$ the following Riccati equation has a unique, positive definite solution P_α:

$$P_\alpha A_\alpha + A_\alpha^\tau P_\alpha - P_\alpha B_\alpha S^{-1} B_\alpha^\tau P_\alpha + Q = 0. \tag{3.3}$$

LEMMA 3.2. *Suppose that (A_α, B_α) is controllable and (A_α, C) is observable (with $C^\tau C = Q$). If A_α and B_α are continuous or i-times differentiable with respect to α in an interval $[\alpha_*, \alpha^*]$, then so is the solution P_α.*

Proof. From Martensson (1971) we see that the solution P_α can actually be expressed in the following form:

$$P_\alpha = Y_\alpha X_\alpha^{-1}, \quad \text{for all } \ \alpha \in [\alpha_*, \alpha^*]$$

where the columns of the composed matrix $[\begin{smallmatrix} X_\alpha \\ Y_\alpha \end{smallmatrix}]$ are eigenvectors or generalized eigenvectors of matrix

$$\Gamma_\alpha \triangleq \begin{bmatrix} A_\alpha & -B_\alpha S^{-1} B_\alpha^\tau \\ -Q & -A_\alpha^\tau \end{bmatrix}.$$

Now, the eigenvectors (respectively, generalized eigenvectors) of a matrix are (respectively, may be chosen to be) continuous functions of its elements. Thus, if A_α and B_α are continuous with respect to α, then Y_α, X_α, and hence P_α are continuous with respect to α.

Similarly, if A_α and B_α are i-times differentiable with respect to α, then P_α is i-times differentiable with respect to α. □

We define the adaptive control law via the certainty equivalence principle and the following quadratic index:

$$\lim_{t \to \infty} \frac{1}{t} \int_0^t (x_s^\tau Q x_s + u_s^\tau S u_s) ds.$$

Hence, we will use the following adaptive control law:

$$u_t = -S^{-1} B_{\widehat{\Phi}_t}^\tau P_{\widehat{\Phi}_t} x_t, \tag{3.4}$$

where $\widehat{\Phi}_t$ is a solution of (2.4), and $P_{\widehat{\Phi}_t}$ is a solution of (3.3) with A_α and B_α replaced by $A_{\widehat{\Phi}_t}$ and $B_{\widehat{\Phi}_t}$, respectively.

Let $\Pi(i)$ denote the ith row of matrix Π and

$$\varepsilon = \sup_{\widehat{\Phi}_t \in \mathcal{D}} \max_{i,j=1,\ldots,N} \left\{ \left\| \frac{\partial^2 \left(P_{\widehat{\Phi}_t}\right)}{\partial \widehat{\Phi}_t(i) \partial \widehat{\Phi}_t(j)} \right\| \right\}, \tag{3.5}$$

$$c_1 = \sup_{\widehat{\Phi}_t \in \mathcal{D}} \left\{ \|A_{\widehat{\Phi}_t} - B_{\widehat{\Phi}_t} S^{-1} B_{\widehat{\Phi}_t}^\tau P_{\widehat{\Phi}_t}\| \right\}, \tag{3.6}$$

$$c_2 = \sup_{\widehat{\Phi}_t \in \mathcal{D}} \max_{i=1,\ldots,N} \left\{ \left\| \frac{\partial \left(P_{\widehat{\Phi}_t}\right)}{\partial \widehat{\Phi}_t(i)} \right\| \right\}, \tag{3.7}$$

where

$$\mathcal{D} \triangleq \left\{ \widehat{\Phi}_t : \ 0 \le \widehat{\Phi}_t(i) \le 1, \ \ i = 1, \ldots, N \ \text{ with } \ \sum_{i=1}^N \widehat{\Phi}_t(i) = 1 \right\}.$$

In other words, $\widehat{\Phi}_t$ ranges over the closed unit simplex $\mathcal{D}$ in $\mathbb{R}^N$.

The closed-loop system referred to in the statement of the main result below is given by the system and parameter process equations (1.1), (1.2), the filter equations (2.4), (2.5), and the Riccati and feedback equations (3.3), (3.4).

THEOREM 3.1. *Suppose that* $(A_{\widehat{\Phi}_t}, B_{\widehat{\Phi}_t})$ *is controllable for all* $\widehat{\Phi}_t$ *in the closed unit simplex* $\mathcal{D}$ *and that for some appropriate positive matrix* S*, the unique solution* $P_{\widehat{\Phi}_t}$ *to* (3.3) *combined with the matrix* Π *in* (1.2) *satisfies*

$$\|\Pi\|\varepsilon + \varepsilon c_1 < \frac{1}{4N}, \qquad c_2 \sum_{i=1}^{N} \|\Pi(i)\| < \frac{1}{2}. \tag{3.8}$$

Then, under the adaptive control law (3.4) *with* $\widehat{\Phi}_t$ *a solution of* (2.4)*, the closed-loop system has a unique strong solution* $\{x_t,\ \widehat{\Phi}_t,\ t \geq 0\}$*, and is stabilized in the following average sense:*

$$\limsup_{T\to\infty} \frac{1}{T} E \int_0^T (\|x_t\|^2 + \|u_t\|^2)dt < \infty.$$

To prove Theorem 3.1, we introduce some notation following Guo (1993). For any fixed positive number K, denote by $\mathcal{C}_K^{n+N}$ the space of $\mathbb{R}^{n+N}$-valued continuous functions on the interval $[0, K]$. When $g = \{g_t\}_{0\leq t\leq K}$ is a $\mathcal{C}_K^{n+N}$ process, we set $\|g\|_{[0,K]} = \max_{0\leq t\leq K} \|g_t\|$.

Proof. First of all, we show that the closed-loop system has a solution $\{x_t,\ \widehat{\Phi}_t,\ t \geq 0\}$. Let

(3.9)

$$z_t = \begin{bmatrix} x_t \\ \widehat{\Phi}_t \end{bmatrix},$$

(3.10)

$$a(z_t) = \begin{bmatrix} (A(\theta_t) - B(\theta_t)S^{-1}B^\tau_{\widehat{\Phi}_t} P_{\widehat{\Phi}_t})x_t \\ \Pi\widehat{\Phi}_t + \left(\mathrm{Diag}\widehat{\Phi}_t - \widehat{\Phi}_t\widehat{\Phi}_t^\tau\right) H_t^\tau \left[(A(\theta_t) - B(\theta_t)S^{-1}B^\tau_{\widehat{\Phi}_t} P_{\widehat{\Phi}_t})x_t - H_t\widehat{\Phi}_t\right] \end{bmatrix},$$

(3.11)

$$b(z_t) = \begin{bmatrix} 1 \\ \left(\mathrm{Diag}\widehat{\Phi}_t - \widehat{\Phi}_t\widehat{\Phi}_t^\tau\right) H_t^\tau \end{bmatrix}.$$

Then from (1.1), (2.4), and (3.4) the closed-loop system can be rewritten in the following form:

$$dz_t = a(z_t)dt + b(z_t)dw_t. \tag{3.12}$$

Obviously, it follows from (2.12) that $A_{\widehat{\Phi}_t}$ and $B_{\widehat{\Phi}_t}$ are differentiable with respect to each component of $\widehat{\Phi}_t$. This combined with Lemma 3.2 implies that $P_{\widehat{\Phi}_t}$ is continuous and bounded on $\mathcal{D}$, since $\mathcal{D}$ is a compact set. Thus, by (3.9)–(3.11), we can conclude that for any fixed $\Delta > 0$, there exists a constant $L^{(\Delta)}$ such that

$$\left[\|a(g_t) - a(h_t)\|^2 + \|b(g_t) - b(h_t)\|^2\right] \mathbb{1}_{\{\|g\|_{[0,K]}\leq\Delta,\ \|h\|_{[0,K]}\leq\Delta\}} \leq L^{(\Delta)}\|g_t - h_t\|^2$$

and

$$\left[\|a(g_t)\|^2 + \|b(g_t)\|^2\right] \mathbb{1}_{\{\|g\|_{[0,K]}\leq\Delta\}} \leq L^{(\Delta)}(1 + \|g_t\|^2),$$

where $\mathbb{1}_{\{\cdot\}}$ is the indicator function of the set $\{\cdot\}$.

Therefore, by Lemma 2.2 of Guo (1993) we know that there is an $\mathcal{F}_t$-time $\sigma_K > 0$ such that (3.12) has a unique strong solution $z_t(\omega)$ on $\{\omega,\ t:\ t < \sigma_K(\omega)\}$, and

$$\sup_{t<\sigma_K(\omega)} \|z_t(\omega)\| = \infty \quad \text{a.s. on } \mathcal{G} \overset{\triangle}{=} \{\omega:\ \sigma_K(\omega) < K\}. \tag{3.13}$$

We now prove $\sigma_K(\omega) = K$ a.s., i.e., $P(\mathcal{G}) = 0$.

Substituting (3.4) into (1.1) results in

$$dx_t = [A(\theta_t) - B(\theta_t)S^{-1}B^\tau_{\widehat{\Phi}_t} P_{\widehat{\Phi}_t}]x_t dt + dw_t,$$

which together with Ito's formula leads to

$$\begin{aligned}\|x_t\|^2 = &\int_0^t x_s^\tau \left([A(\theta_s) - B(\theta_s)S^{-1}B^\tau_{\widehat{\Phi}_s} P_{\widehat{\Phi}_s}]^\tau + [A(\theta_s) - B(\theta_s)S^{-1}B^\tau_{\widehat{\Phi}_s} P_{\widehat{\Phi}_s}]\right) x_s ds\\ &+\|x_0\|^2 + 2\int_0^t x_s^\tau dw_s + nt.\end{aligned} \tag{3.14}$$

Notice that, by Lemma 3.2, $\alpha_1 \overset{\triangle}{=} 2\sup_{s\geq 0,\widehat{\Phi}_s\in\mathcal{D}} \|A(\theta_s) - B(\theta_s)S^{-1}B^\tau_{\widehat{\Phi}_s} P_{\widehat{\Phi}_s}\| < \infty$ a.s., and that by Lemma 4 of Christopeit (1986) there is a random constant $0 < \alpha_2(\omega) < \infty$ a.s. such that

$$2\left|\int_0^t x_s^\tau dw_s\right| \leq \alpha_2(\omega)\int_0^t \|x_s\|^2 ds + \alpha_2(\omega), \quad \forall t \geq 0.$$

By (3.14) we get

$$\|x_t\|^2 \leq (\|x_0\|^2 + nt + \alpha_2(\omega)) + (\alpha_1 + \alpha_2(\omega))\int_0^t \|x_s\|^2 ds.$$

Thus, by the Bellman–Grownwall lemma (see e.g., Desoer and Vidyasagar (1975)) we have

(3.15)

$$\begin{aligned}\|x_t\|^2 &\leq \|x_0\|^2 + nt + \alpha_2(\omega) + (\alpha_1 + \alpha_2(\omega))\int_0^t (\|x_0\|^2 + \lambda + \alpha_2(\omega))e^{(\alpha_1+\alpha_2(\omega))(t-\lambda)}d\lambda\\ &\leq (\|x_0\|^2 + nt + \alpha_2(\omega))e^{(\alpha_1+\alpha_2(\omega))t}, \quad \forall t \geq 0.\end{aligned}$$

If $P(\mathcal{G}) > 0$, then by (3.15) and the fact that $\|\widehat{\Phi}_t\| \leq 1$ we see that

$$\begin{aligned}\sup_{0\leq t<\sigma_K(\omega)} \|z_t(\omega)\|^2 &\leq 2 + 2\sup_{0\leq t<\sigma_K(\omega)} \|x_t(\omega)\|^2\\ &\leq (\|x_0\|^2 + n\sigma_K(\omega) + \alpha_2(\omega))e^{(\alpha_1+\alpha_2(\omega))\sigma_K(\omega)} < \infty \quad \text{a.s. on } \mathcal{G},\end{aligned}$$

contradicting (3.13) and $P(\mathcal{G}) > 0$.

Noting that K can be any positive number, we see that the closed-loop system (3.12) has a unique strong solution $z_t(\omega)$ on any finite time interval.

We now prove the stability of the closed-loop system.

Let $\overline{A}_{\widehat{\Phi}_t} = A_{\widehat{\Phi}_t} - B_{\widehat{\Phi}_t} S^{-1} B^\tau_{\widehat{\Phi}_t} P_{\widehat{\Phi}_t}$ and $\overline{A}_{\widehat{\Phi}_t}(i) = A_i - B_i S^{-1} B^\tau_{\widehat{\Phi}_t} P_{\widehat{\Phi}_t}$. Then from (2.5) and (2.12) it follows that

$$dx_t = H_t \widehat{\Phi}_t dt + d\overline{w}_t = \overline{A}_{\widehat{\Phi}_t} x_t dt + d\overline{w}_t. \tag{3.16}$$

Applying the general Ito formula to $V(x_t) = x_t^\tau P_{\widehat{\Phi}_t} x_t$, and employing (3.3), (2.4), and (2.5), we have the following inequalities (see, e.g., Caines and Nassiri-Toussi (1991)):

$$\begin{aligned}
\mathcal{A}V(x_t) = & -x_t^\tau x_t - x_t^\tau P_{\widehat{\Phi}_t} B_{\widehat{\Phi}_t} S^{-1} B^\tau_{\widehat{\Phi}_t} P_{\widehat{\Phi}_t} x_t \\
& + x_t^\tau \sum_{i=1}^{N} \Pi(i) \widehat{\Phi}_t \frac{\partial \left(P_{\widehat{\Phi}_t}\right)}{\partial \widehat{\Phi}_t(i)} x_t + \mathrm{Tr} P_{\widehat{\Phi}_t} \\
& + x_t^\tau \sum_{i=1}^{N} \widehat{\Phi}_t(i) \frac{\partial \left(P_{\widehat{\Phi}_t}\right)}{\partial \widehat{\Phi}_t(i)} (\overline{A}_{\widehat{\Phi}_t} - \overline{A}_{\widehat{\Phi}_t}(i)) x_t \\
& + x_t^\tau \sum_{i=1}^{N} \widehat{\Phi}_t(i) (\overline{A}_{\widehat{\Phi}_t} - \overline{A}_{\widehat{\Phi}_t}(i))^\tau \frac{\partial \left(P_{\widehat{\Phi}_t}\right)}{\partial \widehat{\Phi}_t(i)} x_t \\
& + \frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} \widehat{\Phi}_t(i) \widehat{\Phi}_t(j) x_t^\tau \frac{\partial^2 \left(P_{\widehat{\Phi}_t}\right)}{\partial \widehat{\Phi}_t(i) \partial \widehat{\Phi}_t(j)} x_t \\
& \cdot x_t^\tau (\overline{A}_{\widehat{\Phi}_t} - \overline{A}_{\widehat{\Phi}_t}(i))^\tau (\overline{A}_{\widehat{\Phi}_t} - \overline{A}_{\widehat{\Phi}_t}(i)) x_t \\
\leq & -x_t^\tau x_t + c_2 \sum_{i=1}^{N} \|\Pi(i)\| \|x_t\|^2 + \mathrm{Tr} P_{\widehat{\Phi}_t} \\
& + 2c_2 \|x_t\| \sum_{i=1}^{N} \widehat{\Phi}_t(i) \|(\overline{A}_{\widehat{\Phi}_t} - \overline{A}_{\widehat{\Phi}_t}(i)) x_t\| \\
& + \frac{\varepsilon}{2} \|x_t\|^2 \left[\sum_{i=1}^{N} \widehat{\Phi}_t(i) \|(\overline{A}_{\widehat{\Phi}_t} - \overline{A}_{\widehat{\Phi}_t}(i)) x_t\| \right]^2 \\
\leq & -\left(\frac{3}{4} - c_2 \sum_{i=1}^{N} \|\Pi(i)\| \right) \|x_t\|^2 + \mathrm{Tr} P_{\widehat{\Phi}_t} \\
& + 4Nc_2^2 \sum_{i=1}^{N} [\widehat{\Phi}_t(i)]^2 \|(\overline{A}_{\widehat{\Phi}_t} - \overline{A}_{\widehat{\Phi}_t}(i)) x_t\|^2 \\
& + \frac{N\varepsilon}{2} \|x_t\|^2 \sum_{i=1}^{N} [\widehat{\Phi}_t(i)]^2 \|(\overline{A}_{\widehat{\Phi}_t} - \overline{A}_{\widehat{\Phi}_t}(i)) x_t\|^2,
\end{aligned}$$

where we have used the sum of squares bound $2ab \leq 1/4a^2 + 4b^2$ and a standard sum of squares bound to obtain the last inequality above and where ε and c_2 are given by (3.5) and (3.7), respectively.

By the second inequality of condition (3.8), we see

$$\beta \triangleq \frac{3}{4} - c_2 \sum_{i=1}^{N} \|\Pi(i)\| > \frac{1}{4} > 0,$$

and hence, by Lemma 3.1, we get

$$\begin{aligned}
&\limsup_{T\to\infty}\frac{1}{T}E\int_0^T \|x_t\|^2 dt\\
&\le \limsup_{T\to\infty}\frac{1}{\beta T}E\int_0^T \mathrm{Tr}P_{\widehat{\Phi}_t}dt\\
&+\limsup_{T\to\infty}\frac{4Nc_2^2}{\beta T}E\int_0^T\sum_{i=1}^N[\widehat{\Phi}_t(i)]^2\|(\overline{A}_{\widehat{\Phi}_t}-\overline{A}_{\widehat{\Phi}_t}(i))x_t\|^2dt\\
&+\limsup_{T\to\infty}\frac{N\varepsilon}{2\beta T}E\int_0^T\|x_t\|^2\sum_{i=1}^N[\widehat{\Phi}_t(i)]^2\|(\overline{A}_{\widehat{\Phi}_t}-\overline{A}_{\widehat{\Phi}_t}(i))x_t\|^2dt.
\end{aligned}\tag{3.17}$$

By (3.4), i.e., $u_t = -S^{-1}B^\tau_{\widehat{\Phi}_t}P_{\widehat{\Phi}_t}x_t$, we have

$$[\overline{A}_{\widehat{\Phi}_t}-\overline{A}_{\widehat{\Phi}_t}(i)]x_t = A_{\widehat{\Phi}_t}x_t + B_{\widehat{\Phi}_t}u_t - A_ix_t - B_iu_t.$$

Thus, from (3.17) and Corollaries 2.1 and 2.2 it follows that for any fixed $\eta > 0$,

$$\begin{aligned}
&\limsup_{T\to\infty}\frac{1}{T}E\int_0^T \|x_t\|^2 dt\\
&\le \limsup_{T\to\infty}\frac{1}{\beta T}E\int_0^T \mathrm{Tr}P_{\widehat{\Phi}_t}dt\\
&+\left\{16N\|\Pi\|\beta^{-1}c_2^2+N\varepsilon(16\|\Pi\|\eta^{-1}+N)(2\beta)^{-1}\right\}\\
&+\limsup_{T\to\infty}\frac{N(\eta+2\|\Pi\|)\varepsilon}{2\beta T}E\int_0^T\|x_t\|^2dt\\
&+\limsup_{T\to\infty}\frac{N\varepsilon}{\beta T}E\int_0^T|x_t^\tau H_t\widehat{\Phi}_t|dt.
\end{aligned}\tag{3.18}$$

It is easy to see that

$$|x_t^\tau H_t\widehat{\Phi}_t| = |x_t^\tau\overline{A}_{\widehat{\Phi}_t}x_t| \le c_1\|x_t\|^2,\tag{3.19}$$

where c_1 is defined in (3.6).

Substituting (3.19) into (3.18) we get that for all $\eta > 0$,

$$\begin{aligned}
&\limsup_{T\to\infty}\frac{1}{T}E\int_0^T \|x_t\|^2 dt\\
&\le \limsup_{T\to\infty}\frac{1}{\beta T}E\int_0^T \mathrm{Tr}P_{\widehat{\Phi}_t}dt\\
&+\left\{16N\|\Pi\|\beta^{-1}c_2^2+N\varepsilon(16\|\Pi\|\eta^{-1}+N)(2\beta)^{-1}\right\}\\
&+\limsup_{T\to\infty}\frac{N(\eta+2\|\Pi\|)\varepsilon+2N\varepsilon c_1}{2\beta T}E\int_0^T\|x_t\|^2dt.
\end{aligned}\tag{3.20}$$

Notice that (3.8) implies

$$N\beta^{-1}(\|\Pi\|\varepsilon+\varepsilon c_1) < 1.$$

So we can fix a constant $\eta > 0$ at such a value that

$$\frac{N(\eta + 2\|\Pi\|)\varepsilon + 2N\varepsilon c_1}{2\beta} < 1. \tag{3.21}$$

Recalling that $P_{\widehat{\Phi}_t}$ is bounded on $\mathcal{D}$, we get

$$\limsup_{T\to\infty} \frac{1}{T} E \int_0^T \operatorname{Tr} P_{\widehat{\Phi}_t}\, dt < \infty,$$

and hence, by (3.21) and (3.20) we have

$$\limsup_{T\to\infty} \frac{1}{T} E \int_0^T \|x_t\|^2 dt < \infty,$$

which together with (3.4) results in

$$\limsup_{T\to\infty} \frac{1}{T} E \int_0^T \|u_t\|^2 dt < \infty.$$

Therefore, Theorem 3.1 is true. □

4. An example. In this section, we present an example to demonstrate that the conditions of Theorem 3.1 are verifiable in certain nontrivial cases.

EXAMPLE 4.1. If system (1.1) is such that $n = 2$, $m = 1$, $B_1 = B_2 = \cdots = B_N = \left[\begin{smallmatrix} 0 \\ b \end{smallmatrix}\right]$ with

$$b \neq 0 \quad \text{and} \quad A_i = \begin{bmatrix} 0 & 1 \\ 0 & -a_i \end{bmatrix}$$

for a_i distinct, $i = 1, \ldots, N$, then (i) $(A_{\widehat{\Phi}_t}, B_{\widehat{\Phi}_t})$ is controllable for all $\widehat{\Phi}_t$ in the closed unit simplex, and (ii) condition (3.8) and the conclusion of Theorem 3.1 are true when the parameter S in the control Riccati equation (4.2) for $P_{\widehat{\Phi}_t}$ is sufficiently small.

Proof. The truth of (i) is evident. Concerning (ii) set

$$a_{\widehat{\Phi}_t} = \sum_{i=0}^{N} \widehat{\Phi}_t(i) a_i \quad \text{and} \quad P_{\widehat{\Phi}_t} = \begin{bmatrix} P_{\widehat{\Phi}_t}(1,1) & P_{\widehat{\Phi}_t}(1,2) \\ P_{\widehat{\Phi}_t}(1,2) & P_{\widehat{\Phi}_t}(2,2) \end{bmatrix}. \tag{4.1}$$

Then the algebraic Riccati equation (3.3) becomes

$$\begin{bmatrix} P_{\widehat{\Phi}_t}(1,1) & P_{\widehat{\Phi}_t}(1,2) \\ P_{\widehat{\Phi}_t}(1,2) & P_{\widehat{\Phi}_t}(2,2) \end{bmatrix} \begin{bmatrix} 0 & 1 \\ 0 & -a_{\widehat{\Phi}_t} \end{bmatrix} + \begin{bmatrix} 0 & 0 \\ 1 & -a_{\widehat{\Phi}_t} \end{bmatrix} \begin{bmatrix} P_{\widehat{\Phi}_t}(1,1) & P_{\widehat{\Phi}_t}(1,2) \\ P_{\widehat{\Phi}_t}(1,2) & P_{\widehat{\Phi}_t}(2,2) \end{bmatrix}$$
$$- \begin{bmatrix} P_{\widehat{\Phi}_t}(1,1) & P_{\widehat{\Phi}_t}(1,2) \\ P_{\widehat{\Phi}_t}(1,2) & P_{\widehat{\Phi}_t}(2,2) \end{bmatrix} \begin{bmatrix} 0 & 0 \\ 0 & S^{-1}b^2 \end{bmatrix} \begin{bmatrix} P_{\widehat{\Phi}_t}(1,1) & P_{\widehat{\Phi}_t}(1,2) \\ P_{\widehat{\Phi}_t}(1,2) & P_{\widehat{\Phi}_t}(2,2) \end{bmatrix} + I = 0, \tag{4.2}$$

which is equivalent to

$$\begin{aligned}
0 &= 1 - S^{-1}b^2 P^2_{\widehat{\Phi}_t}(1,2), \\
0 &= P_{\widehat{\Phi}_t}(1,1) - S^{-1}b^2 P_{\widehat{\Phi}_t}(1,2) P_{\widehat{\Phi}_t}(2,2) - a_{\widehat{\Phi}_t} P_{\widehat{\Phi}_t}(1,2), \\
0 &= P^2_{\widehat{\Phi}_t}(2,2) + 2Sb^{-2} a_{\widehat{\Phi}_t} P_{\widehat{\Phi}_t}(2,2) - Sb^{-2}(1 + 2|b|^{-1}\sqrt{S}).
\end{aligned}$$

Solving this set of equations we get

$$P_{\hat{\Phi}_t}(1,1) = |b|^{-1}\sqrt{S}\left[a^2_{\hat{\Phi}_t} + S^{-1}b^2 + 2|b|S^{-1/2}\right]^{1/2}, \tag{4.3}$$

$$P_{\hat{\Phi}_t}(1,2) = |b|^{-1}\sqrt{S}, \tag{4.4}$$

$$P_{\hat{\Phi}_t}(2,2) = Sb^{-2}\left[a^2_{\hat{\Phi}_t} + S^{-1}b^2 + 2|b|S^{-1/2}\right]^{1/2} - Sb^{-2}a_{\hat{\Phi}_t}. \tag{4.5}$$

Hence, when S is small enough,

$$\begin{aligned}\mathrm{Tr}P_{\hat{\Phi}_t} &= P_{\hat{\Phi}_t}(1,1) + P_{\hat{\Phi}_t}(2,2)\\ &= -Sb^{-2}a_{\hat{\Phi}_t} + (|b|^{-1}\sqrt{S} + Sb^{-2})\left[a^2_{\hat{\Phi}_t} + S^{-1}b^2 + 2|b|S^{-1/2}\right]^{1/2}\\ &= O(S),\end{aligned}$$

where $O(S)$ denotes a function of S satisfying $\limsup_{S\to 0}\frac{|O(S)|}{S} < \infty$.

From this it follows that

$$\limsup_{T\to\infty}\frac{1}{T}E\int_0^T \mathrm{Tr}P_{\hat{\Phi}_t}\,dt < \infty.$$

Let $\mu = S^{-1}b^2 + 2|b|S^{-1/2}$ and $\gamma_{\hat{\Phi}_t} = \left[a^2_{\hat{\Phi}_t} + \mu\right]^{1/2}$. Then it is easy to see that

$$\frac{\partial\left(\gamma_{\hat{\Phi}_t}\right)}{\partial\hat{\Phi}_t(i)} = \left[a^2_{\hat{\Phi}_t} + \mu\right]^{-1/2} a_{\hat{\Phi}_t}\frac{\partial\left(a_{\hat{\Phi}_t}\right)}{\partial\hat{\Phi}_t(i)} = \left[a^2_{\hat{\Phi}_t} + \mu\right]^{-1/2} a_{\hat{\Phi}_t}a_i, \tag{4.6}$$

where $i = 1,\dots,N$.

Furthermore,

$$\begin{aligned}\frac{\partial^2\left(\gamma_{\hat{\Phi}_t}\right)}{\partial\hat{\Phi}_t(i)\partial\hat{\Phi}_t(j)} &= \left[a^2_{\hat{\Phi}_t} + \mu\right]^{-1/2} a_ia_j - a^2_{\hat{\Phi}_t}\left[a^2_{\hat{\Phi}_t} + \mu\right]^{-3/2} a_ia_j\\ &= \mu\left[a^2_{\hat{\Phi}_t} + \mu\right]^{-3/2} a_ia_j, \quad i,\, j = 1,\dots,N.\end{aligned} \tag{4.7}$$

From (4.3)–(4.6) it follows that for $i = 1,\dots,$ or N,

$$\frac{\partial\left(P_{\hat{\Phi}_t}\right)}{\partial\hat{\Phi}_t(i)} = \begin{bmatrix}\dfrac{a_ia_{\hat{\Phi}_t}|b|^{-1}S^{1/2}}{\left[a^2_{\hat{\Phi}_t} + S^{-1}b^2 + 2|b|S^{-1/2}\right]^{1/2}} & 0\\ 0 & \dfrac{a_ia_{\hat{\Phi}_t}Sb^{-2}}{\left[a^2_{\hat{\Phi}_t} + S^{-1}b^2 + 2|b|S^{-1/2}\right]^{1/2}} - Sb^{-2}a_i\end{bmatrix},$$

which implies that for S sufficiently small

$$c_2 \equiv c_2(S) \leq c_3S, \tag{4.8}$$

where c_3 is a constant depending on a_i and b only.

Since

$$\frac{\partial^2\left(a_{\widehat{\Phi}_t}\right)}{\partial\widehat{\Phi}_t(i)\partial\widehat{\Phi}_t(j)}=0,$$

(4.3)–(4.5) and (4.7) yield

$$\frac{\partial^2\left(P_{\widehat{\Phi}_t}\right)}{\partial\widehat{\Phi}_t(i)\partial\widehat{\Phi}_t(j)}=\begin{bmatrix}\dfrac{a_ia_j(2+|b|S^{-1/2})}{\left[a^2_{\widehat{\Phi}_t}+S^{-1}b^2+2|b|S^{-1/2}\right]^{3/2}} & 0\\ 0 & \dfrac{a_ia_j(1+2|b|^{-1}S^{1/2})}{\left[a^2_{\widehat{\Phi}_t}+S^{-1}b^2+2|b|S^{-1/2}\right]^{3/2}}\end{bmatrix}.$$

From this we obtain that as $S\to 0$,

$$\left\|\frac{\partial^2\left(P_{\widehat{\Phi}_t}\right)}{\partial\widehat{\Phi}_t(i)\partial\widehat{\Phi}_t(j)}\right\|=|a_ia_j|Sb^{-2}\left(1+O(S^{1/2})\right),$$

which implies that as $S\to 0$,

$$\varepsilon=S\max_{i,j=1,\dots,N}|a_ia_j|b^{-2}\left(1+O(S^{1/2})\right). \tag{4.9}$$

From (4.3)–(4.5) it follows that

$$\begin{bmatrix}0 & 1\\ 0 & a_{\widehat{\Phi}_t}\end{bmatrix}-S^{-1}\begin{bmatrix}0\\ b\end{bmatrix}\begin{bmatrix}0 & b\end{bmatrix}\begin{bmatrix}P_{\widehat{\Phi}_t}(1,1) & P_{\widehat{\Phi}_t}(1,2)\\ P_{\widehat{\Phi}_t}(1,2) & P_{\widehat{\Phi}_t}(2,2)\end{bmatrix}$$

$$=\begin{bmatrix}0 & 1\\ -|b|S^{-1/2} & 2a_{\widehat{\Phi}_t}-\left[a^2_{\widehat{\Phi}_t}+S^{-1}b^2+2|b|S^{-1/2}\right]^{1/2}\end{bmatrix}.$$

Then we get

$$\|A_{\widehat{\Phi}_t}-S^{-1}B_{\widehat{\Phi}_t}B^{\tau}_{\widehat{\Phi}_t}P_{\widehat{\Phi}_t}\|\le 2S^{-1/2}|b|\left(1+O(S^{1/2})\right),$$

which implies that

$$c_1\le 2S^{-1/2}|b|\left(1+O(S^{1/2})\right).$$

From this and (4.8), (4.9) we see that for some sufficiently small S, condition (3.8), and hence the results of Theorem 3.1, are true. □

5. Maximum likelihood-based adaptive control. Intuitively, if $\widehat{\Phi}_t$ is a good estimate of Φ_t, in some sense, then $A_{\widehat{\Phi}_t}$ and $B_{\widehat{\Phi}_t}$ are good estimates of $A(\theta_t)$ and $B(\theta_t)$. Therefore, in the last two sections, we discuss the stabilization problem of the filtered system (2.5):

$$dx_t=A_{\widehat{\Phi}_t}x_tdt+B_{\widehat{\Phi}_t}u_tdt+d\overline{w}_t\quad\text{by (2.1), (2.2), and (2.12),}$$

rather than that of system (1.1).

Let i_t be defined by

$$i_t = \arg \max_{i=1,\ldots,N} \{\widehat{\Phi}_t(i)\}, \quad t \geq 0. \tag{5.1}$$

Again, if $\widehat{\Phi}_t$ is a good estimate of Φ_t, in some sense, then $A(i_t)$ and $B(i_t)$ should also be good estimates of $A(\theta_t)$ and $B(\theta_t)$. In this case, it is natural to ask whether we could find an adaptive stabilization control law for system (1.1) by only discussing the following system:

$$dx_t = A(i_t)x_t dt + B(i_t)u_t dt + d\overline{w}_t\,.$$

This section, as an application of Corollary 2.1, will answer this problem. By using the notion of a maximum likelihood estimate, we present some sufficient conditions for stabilization control of the system (1.1)–(1.2). These sufficient conditions are different from those used in §3, but similar to those introduced in Ezzine and Haddad (1989).

For simplicity of notation, for a matrix A, let

$$\mu(A) = \lambda_{\max}\left(\frac{A + A^\tau}{2}\right).$$

THEOREM 5.1. *Suppose there is a matrix $K(i)$ $(i = 1, \ldots, N)$ such that*

$$\nu \triangleq -\max_{i=1,\ldots,N} \mu\left(A(i) - B(i)K(i)\right) > 0. \tag{5.2}$$

Then, under the adaptive control law $u_t = -K(i_t)x_t$, the closed-loop system has a solution $\{x_t,\, u_t,\, t \geq 0\}$, and the input and output of the closed-loop system are bounded in the following average sense:

$$\sup_{t\geq 0} \frac{1}{t+1} E \int_0^t (\|x_s\|^2 + \|u_s\|^2) ds < \infty. \tag{5.3}$$

Proof. Similar to the argument of Theorem 3.1, we see that the closed-loop system has a solution $\{x_t,\, \widehat{\Phi}_t,\, t \geq 0\}$. So, here we only need prove (5.3).

From (2.5) and (2.12) it follows that

$$\begin{aligned} dx_t &= H_t\widehat{\Phi}_t dt + d\overline{w}_t = A_{\widehat{\Phi}_t} x_t dt + B_{\widehat{\Phi}_t} u_t dt + d\overline{w}_t \\ &= A(i_t)x_t dt + B(i_t)u_t dt \\ &\quad + [A_{\widehat{\Phi}_t} x_t + B_{\widehat{\Phi}_t} u_t - A(i_t)x_t - B(i_t)u_t]dt + d\overline{w}_t, \end{aligned} \tag{5.4}$$

where i_t is given in (5.1)

Substituting $u_t = -K(i_t)x_t$ into (5.4) we get

$$\begin{aligned} dx_t &= [A(i_t) - B(i_t)K(i_t)]x_t dt \\ &\quad + [A_{\widehat{\Phi}_t} x_t + B_{\widehat{\Phi}_t} u_t - A(i_t)x_t - B(i_t)u_t]dt + d\overline{w}_t, \end{aligned} \tag{5.5}$$

which together with Ito's formula and (5.2) implies that for the ν given by (5.2)

$$\|x_t\|^2 = \|x_0\|^2 + \int_0^t x_s^\tau \left([A(i_s) - B(i_s)K(i_s)]^\tau + [A(i_s) - B(i_s)K(i_s)]\right) x_s ds$$

$$
\begin{aligned}
&+t+2\int_0^t x_s^\tau d\overline{w}_s + 2\int_0^t x_s^\tau[A_{\widehat{\Phi}_s}x_s + B_{\widehat{\Phi}_s}u_s - A(i_s)x_s - B(i_s)u_s]ds \\
&\le \|x_0\|^2 - 2\nu\int_0^t \|x_s\|^2 ds + 2\int_0^t x_s^\tau d\overline{w}_s + t \\
&+2\int_0^t x_s^\tau[A_{\widehat{\Phi}_s}x_s + B_{\widehat{\Phi}_s}u_s - A(i_s)x_s - B(i_s)u_s]ds. \qquad (5.6)
\end{aligned}
$$

Notice that

$$
\begin{aligned}
&2\int_0^t x_s^\tau[A_{\widehat{\Phi}_s}x_s + B_{\widehat{\Phi}_s}u_s - A(i_s)x_s - B(i_s)u_s]ds \\
&\quad\le \nu\int_0^t \|x_s\|^2 ds + \nu^{-1}\int_0^t \|A_{\widehat{\Phi}_s}x_s + B_{\widehat{\Phi}_s}u_s - A(i_s)x_s - B(i_s)u_s\|^2 ds;
\end{aligned}
$$

by (5.6) we get

$$
\begin{aligned}
\|x_t\|^2 \le \|x_0\|^2 - \nu\int_0^t \|x_s\|^2 ds + 2\int_0^t x_s^\tau d\overline{w}_s + t \\
+\nu^{-1}\int_0^t \|A_{\widehat{\Phi}_s}x_s + B_{\widehat{\Phi}_s}u_s - A(i_s)x_s - B(i_s)u_s\|^2 ds,
\end{aligned}
$$

which implies

$$
\begin{aligned}
E\int_0^t \|x_s\|^2 ds \le \nu^{-2}E\int_0^t \|A_{\widehat{\Phi}_s}x_s + B_{\widehat{\Phi}_s}u_s - A(i_s)x_s - B(i_s)u_s\|^2 ds \\
+\nu^{-1}E\|x_0\|^2 + \nu^{-1}t. \qquad (5.7)
\end{aligned}
$$

By (2.1) we see that $\widehat{\Phi}_t(i) \ge 0$ for $i = 1, \dots, N$ and $t \ge 0$; further, since

$$\sum_{i=1}^N \widehat{\Phi}_t(i) = 1,$$

we have $\widehat{\Phi}_t(i_t) \ge \frac{1}{N}$. Thus, by Corollary 2.1 we get

(5.8)

$$
\begin{aligned}
&E\int_0^t \|A_{\widehat{\Phi}_s}x_s + B_{\widehat{\Phi}_s}u_s - A(i_s)x_s - B(i_s)u_s\|^2 ds \\
&\le N^2E\int_0^t \|\widehat{\Phi}_s(i_s)]^2[A_{\widehat{\Phi}_s}x_s + B_{\widehat{\Phi}_s}u_s - A(i_s)x_s - B(i_s)u_s\|^2 ds \\
&\le N^2E\int_0^t \sum_{i=1}^N [\widehat{\Phi}_s(i)]^2\|A_{\widehat{\Phi}_s}x_s + B_{\widehat{\Phi}_s}u_s - A(i)x_s - B(i)u_s\|^2 ds \quad (\text{since } i_s \in \{1, \dots, N\}) \\
&\le N^2(2 + 6\|\Pi\|t).
\end{aligned}
$$

Substituting this into (5.7) leads to the desired result, (5.3). □

From the definition (5.1) of i_t and $u_t = -K(i_t)x_t$ it follows that u_t may jump at any time instant t. In order to get a piecewise continuous control u_t, that is, one that has with probability 1 no accumulation points of switching times on the time axis, one can modify the definition (5.1) of i_t as follows:

$$i_t = i'_{\tau_{k-1}}, \quad \forall t \in [\tau_{k-1},\ \tau_k), \quad \forall k = 1,\ 2, \dots, \qquad (5.9)$$

where

$$i'_t = \arg \max_{i=1,\dots,N} \{\widehat{\Phi}_t(i)\}, \quad \forall t \geq 0, \tag{5.10}$$

$$\tau_k = \inf \left\{ t > \tau_{k-1} : \quad \widehat{\Phi}_t(i_{\tau_{k-1}}) \leq (\gamma N)^{-1} \right\} \tag{5.11}$$

with $\tau_0 = 0$, $\gamma > 1$, and $k = 1, 2, \dots$ being positive integers.

Since the trajectories of $\widehat{\Phi}$ are continuous and $\gamma > 1$ it is evident that $K(\cdot)$ and hence u has the required piecewise continuous property.

THEOREM 5.2. *If $\{i_t;\ t \geq 0\}$ and $\{\tau_k;\ k = 1, 2, \dots\}$ are generated from* (5.9)–(5.11) *and $u_t \in \mathcal{F}_t^x$, then $\lim_{k\to\infty} \tau_k = \infty$ a.s. and i_t is piecewise constant a.s. Furthermore, if condition* (5.2) *of Theorem* 5.1 *is true and the adaptive control law is chosen to be $u_t = -K(i_t)x_t$, then u_t is piecewise continuous and the input and output of the closed-loop system are bounded in the average sense* (5.3).

Proof. First, we show $\lim_{k\to\infty} \tau_k = \infty$ a.s. Noticing that

$$\max_{i=1,\dots,N} \{\widehat{\Phi}_t(i)\} \geq N^{-1}$$

and every component of $\widehat{\Phi}_t$ is a continuous function of t, by $\gamma > 1$ we see that $\tau_k > \tau_{k-1}$. Thus, $\lim_{k\to\infty} \tau_k$ exists a.s.

If the sample set $\mathcal{S} \triangleq \{\omega : \lim_{k\to\infty} \tau_k < \infty\}$ had positive probability, i.e., $P(\mathcal{S}) > 0$, then there would exist a deterministic constant $T < \infty$ such that $\mathcal{S}_1 \triangleq \{\omega : \lim_{k\to\infty} \tau_k \leq T\}$ with positive probability, i.e., $P(\mathcal{S}_1) > 0$.

Notice that for any constant $t \geq 0$,

$$\{\omega :\ 0 < \tau_k I_{\mathcal{S}_1} \leq t\} = \{\omega :\ 0 < \tau_k \leq t\} \cap \{\omega :\ I_{\mathcal{S}_1} = 1\} \in \mathcal{F}_t^x,$$

where

$$I_{\mathcal{S}_1} = \begin{cases} 1, & \text{if } \omega \in \mathcal{S}_1, \\ 0, & \text{if } \omega \notin \mathcal{S}_1. \end{cases}$$

By $\|\widehat{\Phi}_{\tau_{k+1}} - \widehat{\Phi}_{\tau_k}\|^2 \geq N^{-2}(1-\gamma^{-1})^2 > 0$, (2.4), and (2.14) we have

$$\begin{aligned}
\infty = \sum_{k=0}^{\infty} N^{-2}(1-\gamma^{-1})^2 P(\mathcal{S}_1) &\leq E I_{\mathcal{S}_1} \sum_{k=0}^{\infty} \|\widehat{\Phi}_{\tau_{k+1}} - \widehat{\Phi}_{\tau_k}\|^2 \\
&\leq 2\|\Pi\|^2 E I_{\mathcal{S}_1} \sum_{k=0}^{\infty} (\tau_{k+1} - \tau_k)^2 + 2 E I_{\mathcal{S}_1} \sum_{k=0}^{\infty} \left(\int_{\tau_k}^{\tau_{k+1}} (\mathrm{Diag}\widehat{\Phi}_s - \widehat{\Phi}_s \widehat{\Phi}_s^\tau) H_s^\tau d\overline{w}_s \right)^2 \\
&\leq 2\|\Pi\|^2 T E I_{\mathcal{S}_1} \sum_{k=0}^{\infty} (\tau_{k+1} - \tau_k) + 2E \sum_{k=0}^{\infty} \int_{\tau_k I_{\mathcal{S}_1}}^{\tau_{k+1} I_{\mathcal{S}_1}} \left\| (\mathrm{Diag}\widehat{\Phi}_s - \widehat{\Phi}_s \widehat{\Phi}_s^\tau) H_s^\tau \right\|^2 ds \\
&\leq 2\|\Pi\|^2 T^2 P(\mathcal{S}_1) + 2E \int_0^T \left\| (\mathrm{Diag}\widehat{\Phi}_s - \widehat{\Phi}_s \widehat{\Phi}_s^\tau) H_s^\tau \right\|^2 ds \\
&\leq 2\|\Pi\|^2 T^2 P(\mathcal{S}_1) + 2(2 + 6\|\Pi\| T) < \infty.
\end{aligned}$$

This contradiction means that $\lim_{k\to\infty} \tau_k = \infty$ a.s. Thus, from $\tau_k > \tau_{k-1}$ and (5.9) it follows that i_t is piecewise constant a.s.

As in Theorem 5.1, with condition (5.2) we can prove that the under-control law $u_t = -K(i_t)x_t$, with i_t given by (5.9)–(5.11), the closed-loop system has a solution

$\{x_t,\ t \geq 0\}$ a.s. and is stabilized in the average sense of (5.3); this is because the only difference between the proofs is due to (5.8), which now becomes

$$E\int_0^t \|A_{\widehat{\Phi}_s} x_s + B_{\widehat{\Phi}_s} u_s - A(i_s)x_s - B(i_s)u_s\|^2 ds \leq (\gamma N)^2(2 + 6\|\Pi\|t),$$

since, in this case, $\widehat{\Phi}_t(i_t) \geq (\gamma N)^{-1}$ for all $t \geq 0$.

Noticing that $\lim_{k\to\infty} \tau_k = \infty$ a.s. and that i_t is piecewise constant a.s., we see that the control $u_t = -K(i_t)x_t$ is almost surely defined for all $t \geq 0$ and is a piecewise continuous function of t. □

Remark 5.1. Although it is hard to say whether or not condition (5.2) is true in general cases, there exist specific situations where it is readily verified; for instance, (i) the case where $A(i)$ and $B(i)$ are scalar and $(A(i), B(i))$ is stabilizable for every $i = 1, \ldots, N$, and (ii) that where $B(i)$ is invertible for $i = 1, \ldots, N$, and there exists $K(i)$ such that (5.2) holds.

In fact, for case (i), $K(i)$ can be chosen as

$$K(i) = \begin{cases} [B(i)]^{-1}[1 + A(i)], & \text{if } B(i) \neq 0, \\ 0, & \text{if } B(i) = 0, \end{cases}$$

and the constant ν in (5.2) may be taken equal to the following positive quantity:

$$-\max\{\nu_i, \quad i = 1, \ldots, N\},$$

where

$$\nu_i = \begin{cases} -1, & \text{if } B(i) \neq 0, \\ A(i), & \text{if } B(i) = 0. \end{cases}$$

For case (ii), $K(i)$ can be chosen as $K(i) = [B(i)]^{-1}[I + A(i)]$, which results in $\nu = 1$.

Remark 5.2. We now revisit the example given by Dufour and Bertrand (1993). In (1.1), they set $n = 2$, $m = 1$,

$$B_1 = \begin{bmatrix} 0 \\ 1 \end{bmatrix}, B_2 = \begin{bmatrix} 0 \\ -1 \end{bmatrix}, A_1 = \begin{bmatrix} -1 & -1 \\ 0 & 2 \end{bmatrix} \quad \text{and} \quad A_2 = \begin{bmatrix} -1 & 1 \\ 0 & -1 \end{bmatrix}.$$

In this case the conditions of Theorem 3.1 above do not hold, but the conditions of the theorem of Dufour and Bertrand are valid.

However, for this example, the adaptive control law described in Theorem 5.1 or 5.2 is applicable, and can stabilize the closed-loop system. This is because for $K(1) = [0,\ 4]$ and $K(2) = [0,\ -1]$, we get $\nu = \frac{3-\sqrt{2}}{2} > 0$ by a straightforward manipulation. This implies that condition (5.2), and hence the conclusion of Theorems 5.1 and 5.2, are true.

Acknowledgments. The authors gratefully acknowledge conversations with L. Guo concerning this work.

REFERENCES

P. E. CAINES AND H. F. CHEN, *Optimal adaptive LQG control for systems with finite state process parameters,* IEEE Trans. Automatic Control, 30 (1985), pp. 185–189.

P. E. CAINES AND K. NASSIRI-TOUSSI, *On the adaptive stabilization and ergodic behaviour of stochastic systems with jump-Markov parameters via non-linear filtering,* in L. Gerencser and P. E. Caines, eds., Topics in Stochastic Systems: Modelling, Estimation and Adaptive Control, Lecture Notes in Control and Inform. Sci. 161, Springer-Verlag, Heidelberg, 1991.

P. E. CAINES AND J. F. ZHANG, *Adaptive Control for Jump Parameter Systems via Non-Linear Filtering,* Proc. 31st IEEE Conference on Decision and Control, Tuscon, Arizona, December, 16–18, 1992, pp. 699–704.

H. F. CHEN AND P. E. CAINES, *On the adaptive stabilization of linear stochastic systems with jump process parameters,* Proc. IEEE Conference on Decision and Control, Tampa, FL, Dec. 1989.

N. CHRISTOPEIT, *Quasi-least-squares estimation in semimartingale regression models,* Stochastics, 16(1986), pp. 255–278.

C. A. DESOER AND M. VIDYASAGAR, *Feedback Systems: Input-Output Properties,* Academic Press, New York, 1975.

F. DUFOUR AND P. BERTRAND, *Stabilizing control law for hybrid models,* Research Report, Laboratoire des Signaux et Systèmes, Ecole Supérieure D'Electricité, Gif-sur-Yvette, France, 1993.

J. EZZINE AND A. H. HADDAD, *On the largest-Lyapunov exponent of assignment and almost sure stabilization of hybrid systems,* Proc. American Control Conference, 1989, pp. 805–809.

X. FENG, K. A. LOPARO, Y. JI, AND H. J. CHIZECK, *Stochastic stability properties of jump linear systems,* IEEE Trans. Automat. Control, 37 (1992), pp. 38–52.

L. GUO, *A note on continuous-time ELS,* Systems Control Lett., 22 (1994), pp. 111–121.

K. HELMES AND R. RISHEL, *The solution of partially observed stochastic optimal control problem in terms of predicted miss,* Univ. of Kentucky Reprint, Lexington, KY, 1990.

Y. JI AND H. J. CHIZECK, *Controllability, stabilizability and continuous-time Markovian jump linear quadratic control,* IEEE Trans. Automat. Control, 35 (1990), pp. 777–788.

R. S. LIPTSER AND A. N. SHIRYAEV, *Statistics of Random Processes, Vol.* I, Springer-Verlag, New York, 1977, pp. 329–340.

M. MARITON, *On controllability of linear systems with stochastic jump parameters,* IEEE Trans. Automat. Control, 31 (1986), pp. 680–683.

M. MARITON AND P. BERTRAND, *Output feedback for a class of linear systems with stochastic jumping parameters,* IEEE Trans. Automat. Control, 30 (1985), pp. 898–900.

K. MARTENSSON, *On the matrix Riccati equation,* Inform. Sci., 3 (1971), pp. 17–49.

K. NASSIRI-TOUSSI AND P. E. CAINES, *On the adaptive stabilization and ergodic behaviour of stochastic jump parameter systems via non-linear filtering,* IEEE Conference on Decision and Control, Brighton, England, December, 1991, pp. 1784–1785.

R. RISHEL, *A comment on a dual control problem,* IEEE Trans. Automat. Control, 26 (1981), pp. 606–609.

L. SCHWARTZ, *Semimartingales and their Stochastic Calculus on Manifolds,* Les Presses de l'Université de Montréal, Canada, 1984, pp. 99–104.

D. D. SWORDER AND D. S. CHOU, *A survey of design methods for random parameter systems,* Proc. IEEE Conference on Decision and Control, 1985, pp. 894–899.

D. D. SWORDER, *Hybrid adaptive control,* Appl. Math. Comp., 45 (1991), pp. 173–192.

W. M. WONHAM, *Some applications of stochastic differential equations to optimal non-linear filtering,* SIAM J. Control, 2 (1965), pp. 347–369.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 33, No. 6, pp. 1778–1800, November 1995

© 1995 Society for Industrial and Applied Mathematics

008

PERIODIC STABILITY OF NONLINEAR FLEXIBLE SYSTEMS WITH DAMPING*

KOICHIRO NAITO†

Abstract. We study periodic stability of solutions of nonlinear elastic systems with damping under periodic perturbations. Investigating the sufficient conditions for the stability, we find some inequality relations between the system parameters of its linear term and those of its nonlinear term.

Key words. stability, periodic perturbation, nonlinear beam equation, damping

AMS subject classifications. 35B10, 73D35, 73K05

1. Introduction.

Let Ω be a bounded domain in a finite-dimensional Euclidean space. We consider the class of flexible systems that can be described by the following second-order damped evolution equation in $X := L^2(\Omega)$ with a nonlinear forcing term under a periodic perturbation:

$$\frac{d^2u(t)}{dt^2} + 2\alpha A\frac{du(t)}{dt} + Au(t) = F(u(t)) + w(t), \quad t > 0, \tag{1.1}$$

$$u(0) = u_0, \quad u_t(0) = u_1, \quad w(t+T) = w(t). \tag{1.2}$$

We assume that A is a self-adjoint positive definite operator with dense domain $D(A)$ in $L^2(\Omega)$, and that A^{-1} exists and is compact. It is well known that there exist eigenvalues λ_i and corresponding eigenfunctions $\varphi_{i,j}(x)$ of the operator A satisfying the following conditions:

$$\begin{aligned}&0 < \lambda_1 < \lambda_2 < \cdots < \lambda_i < \cdots, \quad \lim_{i\to\infty} \lambda_i = \infty,\\ &A\varphi_{ij} = \lambda_i\varphi_{ij}, \quad j = 1, \ldots, m_i, \quad i = 1, 2, \ldots,\\ &\{\varphi_{ij}(\cdot)\} \text{ forms a complete orthonormal system in } L^2(\Omega).\end{aligned}$$

For each constant $0 \le \sigma \le 1$, the domain $D(A^\sigma)$ of the fractional power A^σ, denoted by X_σ, is topologized by the norm

$$|x|_\sigma^2 := |A^\sigma x|_0^2 = \sum_{i=1}^{\infty}\sum_{j=1}^{m_i} \lambda_i^{2\sigma}|(x, \varphi_{ij})|^2, \quad x \in X_\sigma, \tag{1.3}$$

where $|\cdot|_0$ denotes the norm of X. We also assume that the perturbation function $w(t) : R^+ \to X$ is locally Hölder continuous and uniformly bounded and we denote its usual supremum norm by

$$|w|_\infty := \sup\{|w(t)|_X : t \in R^+\}.$$

We consider the following conditions on the nonlinear function F for a given fixed constant $\beta : 0 < \beta < 1$. F is locally Lipschitz continuous from X_β to X and there exists a constant $k(c) > 0$ such that

$$|Fx - Fy|_0 \le k(c)|x - y|_\beta \quad \text{for } |x|_\beta, |y|_\beta \le c. \tag{1.4}$$

* Received by the editors August 9, 1993; accepted for publication (in revised form) June 1, 1994.

† Department of Mathematics, Faculty of Engineering, Kumamoto University, Kurokami 2-39-1, Kumamoto, 860 Japan (knaito@gpo.kumamoto-u.ac.jp).

There exists a positive constant K_0 such that

$$|F(x)|_0 \le K_0(1+|x|_\beta), \quad x \in X_\beta. \tag{1.5}$$

The formulation (1.1) includes vibrations in mechanically flexible systems, e.g., flexible arms of industrial robots or flexible structures such as antennas of spacecrafts (cf. [1], [10]–[12] in linear systems: $F \equiv 0$). In this paper we treat the case with nonlinear forcing, which is determined not only by the displacement $u(t,x)$, but also by the bending force $u_{xx}(t,x)$. Our main objective is to show sufficient conditions for periodicity and stability of solutions under periodic perturbations $w(t)$. We introduce some inequality relations between system parameters, the eigenvalues of the linear term, and the growth rate or the (locally) Lipschitz constant of the nonlinear term. While it should be considered that the first eigenvalue of the linear operator A essentially determines these relations, we find that the eigenvalues λ_h, λ_{h+1} which satisfy

$$0 < \lambda_1 < \cdots < \lambda_h < \frac{1}{\alpha^2} < \lambda_{h+1} < \cdots$$

have some significant properties for the stability of this system. If the values λ_1, $\lambda_{k+1} - (1/\alpha^2), (1/\alpha^2) - \lambda_h$ are sufficiently large, we can show the asymptotic behavior of solutions, the existence of a global attractor, and periodicity or asymptotic periodicity of solutions under periodic perturbations. Also, we estimate some essential relations among the system parameters $\lambda_1, \lambda_h, \lambda_{h+1}, \alpha$, and $K_0, k(\cdot)$, considering the equation of motion of a one-dimensional nonlinear flexible beam.

Our formulation depends on the method by Sakawa [10] in linear flexible systems and we use spectral properties of analytic semigroups. To analyze nonlinear systems we apply a variation of the Gronwall inequality, which was introduced in [5] (see also [8]). As for the other methods to show periodic stability of nonlinear systems, we can refer to [4], [6], and [7], which mainly depend on the monotone operator theory.

In §2 we give the formulation and state the main theorems: stability, periodicity, and asymptotic periodicity of solutions. We prepare some lemmas on analytic semigroups in §3 and prove the main results in §4. In §5 we investigate a one-dimensional nonlinear flexible beam system to derive sufficient conditions, described by its system parameters, for stability and periodic stability.

2. Formulation and main theorems.

Following the formulation by Sakawa [10] in the linear part, we assume that

$$\alpha^2\lambda_i^2 - \lambda_i \neq 0, \quad i = 1, 2, \ldots$$

and that $\alpha > 0$ is so small that

$$\alpha\lambda_1 < \frac{1}{2\alpha}. \tag{2.1}$$

Define a complex-valued function g by

$$g(\lambda) = \sqrt{\alpha^2\lambda^2 - \lambda}.$$

Then, since A is self-adjoint, we can define an operator $g(A)$ by

$$g(A)u = \sum_{i=1}^{\infty}\sum_{j=1}^{m_i} g(\lambda_i)(u, \varphi_{ij})\varphi_{ij},$$

$$D(g(A)) = \left\{ u \in L^2(\Omega) : \sum_{i=1}^{\infty} \sum_{j=1}^{m_i} |g(\lambda_i)(u, \varphi_{ij})|^2 < \infty \right\}.$$

Note that $D(g(A)) = D(A)$ and define the following two operators by

$$A^+ := \alpha A - g(A), \quad A^- := \alpha A + g(A).$$

Then for each $u \in D(A)$,

$$A^{\pm} u = \sum_{i=1}^{\infty} \sum_{j=1}^{m_i} (\alpha \lambda_i \mp g(\lambda_i))(u, \varphi_{ij})\varphi_{ij}$$

and the eigenvalues and the eigenfunctions of $A^{\pm}$ are given by

$$\begin{aligned} &\nu_i = \alpha\lambda_i - g(\lambda_i), \quad \mu_i = \alpha\lambda_i + g(\lambda_i), \\ &A^+ u = \nu_i \varphi_{ij}, \quad A^- u = \mu_i \varphi_{ij}, \quad j = 1, \dots, m_i, \quad i = 1, 2, \dots. \end{aligned}$$

It follows from (2.1) that there is an integer $h \geq 1$ such that

$$\alpha^2 \lambda_h^2 - \lambda_h < 0, \quad \alpha^2 \lambda_{h+1}^2 - \lambda_{h+1} > 0.$$

In this paper we can show that the three eigenvalues $\lambda_1, \lambda_h, \lambda_{h+1}$ are the most essential parameters in the sufficient conditions for periodic stability.

Since the operators $-A^+, -A^-$ generate analytic semigroups $S_1(t), S_2(t)$, respectively (cf. Lemma 3.1 in [10]), and especially since A^+ is a bounded operator, we can consider the following system of the semilinear equations:

$$\dot{\xi}(t) + A^+ \xi(t) = g^{-1}(A) \left[F\left(\frac{\xi + \eta}{2}\right) + w(t) \right], \tag{2.2}$$

$$\dot{\eta}(t) + A^- \eta(t) = -g^{-1}(A) \left[F\left(\frac{\xi + \eta}{2}\right) + w(t) \right], \tag{2.3}$$

which can be described by

$$\dot{\zeta}(t) + \mathcal{A}\zeta(t) = \mathcal{F}(\zeta(t)) + \mathbf{w}(t), \tag{2.4}$$

where

$$\zeta(t) = \begin{bmatrix} \xi(t) \\ \eta(t) \end{bmatrix}, \quad \mathcal{A} = \begin{bmatrix} A^+ & 0 \\ 0 & A^- \end{bmatrix},$$

$$\mathcal{F}(\zeta(t)) = \begin{bmatrix} g^{-1}(A) F\left(\dfrac{\xi+\eta}{2}\right) \\ -g^{-1}(A) F\left(\dfrac{\xi+\eta}{2}\right) \end{bmatrix}, \quad \mathbf{w}(t) = \begin{bmatrix} g^{-1}(A) w(t) \\ -g^{-1}(A) w(t) \end{bmatrix},$$

and $g^{-1}(A)$ is the inverse operator of $g(A)$, that is,

$$g^{-1}(A) u = \sum_{i=1}^{\infty} \sum_{j=1}^{m_i} (g(\lambda_i))^{-1} (u, \varphi_{ij}) \varphi_{ij}.$$

Also, their mild forms are described as follows:

$$\xi(t) = S_1(t)\xi_0 + \int_0^t S_1(t-s)g^{-1}(A)\left[F\left(\frac{\xi+\eta}{2}(s)\right) + w(s)\right] ds, \tag{2.5}$$

$$\eta(t) = S_2(t)\eta_0 - \int_0^t S_2(t-s)g^{-1}(A)\left[F\left(\frac{\xi+\eta}{2}(s)\right) + w(s)\right] ds. \tag{2.6}$$

Under the conditions (1.4), (1.5) for a fixed constant $0 < \beta < 1$, we can admit the (classical) solution (cf. Appendix 2)

$$[\xi(t), \eta(t)] \in C(0,T : D(A^\beta) \times D(A^\beta)) \cap C^1(0,T : L^2(\Omega) \times L^2(\Omega)) \tag{2.7}$$

for each initial condition $[\xi_0, \eta_0] \in D(A^\beta) \times D(A^\beta)$ and for an arbitrarily fixed constant $T > 0$.

Furthermore, we can estimate the regularity of the solutions as follows: if $[\xi_0, \eta_0] \in D(A) \times D(A^{1+\sigma})$ for some constant $\sigma : 0 < \beta \le \sigma < 1$, then by multiplying λ_i^2 and $\lambda_i^{2(1+\sigma)}$ to the spectral expansions of (2.5) and (2.6), respectively, and applying the direct estimation, such as (1.3), we have

$$\xi \in C(0,T : D(A)), \quad \eta \in C(0,T : D(A^{1+\sigma})). \tag{2.8}$$

Then it follows from (2.2) and (2.3) that

$$\dot{\xi} \in C(0,T : D(A)), \quad \dot{\eta} \in C(0,T : D(A^\sigma)). \tag{2.9}$$

Now, define the functions u, v by

$$u := \frac{\xi+\eta}{2}, \quad v := \frac{\xi-\eta}{2}. \tag{2.10}$$

Then from (2.8) and (2.9) it follows that

$$u, v \in C(0,T : D(A)) \cap C^1(0,T : D(A^\sigma)). \tag{2.11}$$

Hereafter, we consider the case $\sigma = \beta$. From (2.2) and (2.3) we have

$$\dot{u} + \dot{v} + (\alpha A - g(A))(u+v) = g^{-1}(A)(F(u) + w),$$
$$\dot{u} - \dot{v} + (\alpha A + g(A))(u-v) = -g^{-1}(A)(F(u) + w),$$

and then the difference and the sum of the above equations give

$$\dot{u} = -\alpha A u + g(A)v, \tag{2.12}$$
$$\dot{v} - g(A)u + \alpha A v = g^{-1}(A)(F(u) + w). \tag{2.13}$$

By modifying the argument in [10] without the assumption $\dot{u} \in D(A)$, we have

$$\begin{aligned} \dot{v} &= g(A)u - \alpha A v + g^{-1}(A)(F(u) + w) \\ &= g^2(A)g(A)^{-1}u - \alpha A g^{-1}(A)g(A)v + g^{-1}(A)(F(u) + w) \\ &= (\alpha^2 A^2 - A)g^{-1}(A)u - \alpha A g^{-1}(A)g(A)v + g^{-1}(A)(F(u) + w) \\ &= \alpha A(g^{-1}(A)\alpha A u - g^{-1}(A)g(A)v) - A g^{-1}(A)u + g^{-1}(A)(F(u) + w). \end{aligned} \tag{2.14}$$

Thus, using (2.12), we have

$$\dot{v} = -\alpha A g^{-1}(A)\dot{u} - A g^{-1}(A)u + g^{-1}(A)(F(u) + w). \tag{2.15}$$

Also, we note that (2.13) and (2.15) give

$$\alpha A v - g(A)u = \alpha A g^{-1}(A)\dot{u} + A g^{-1}(A)u. \tag{2.16}$$

Obviously, by differentiating (2.12) under the assumption that $\dot{u} \in D(A)$ and using (2.15), we can obtain the evolution equation (1.1). On the other hand, without the assumption $\dot{u} \in D(A)$, consider an initial condition

$$[u(0), \dot{u}(0)] = [u_0, u_1] \in D(A) \times D(A^{\beta}).$$

Then, since (2.12) yields

$$v(0) = g^{-1}(A)(u_1 + \alpha A u_0) \in D(A),$$

we have

$$\xi(0) = u(0) + v(0) \in D(A). \tag{2.17}$$

And also, since (2.12) yields

$$\begin{aligned} g(A)u - g(A)v &= -\dot{u} - (\alpha A u - g(A)u) \\ &= -\dot{u} - A^{+}u, \end{aligned}$$

we have

$$\begin{aligned} \eta(0) &= u(0) - v(0) \\ &= g^{-1}(A)(-u_1 - A^{+}u_0) \in D(A^{1+\beta}). \end{aligned} \tag{2.18}$$

Thus, by applying the previous argument with (2.17) and (2.18), we can admit the solution $u = (\xi + \eta)/2$ in the mild sense such that

$$[u, \dot{u}] \in C(0, T : D(A)) \times C(0, T : D(A^{\beta})).$$

Before stating the stability of solutions, we introduce some notations. Define

$$\begin{aligned} \lambda(\beta) &= \min\{\sqrt{\lambda_1}, \lambda_1^{1-\beta}\}, \\ M_h &= \max\left\{\frac{1}{\sqrt{1-\alpha^2\lambda_h}}, \sqrt{\frac{\alpha^2\lambda_{h+1}}{\alpha^2\lambda_{h+1}-1}+1}\right\}, \\ C_h &= \max\left\{\sqrt{\frac{\lambda_h}{1-\alpha^2\lambda_h}}, \sqrt{\frac{\lambda_{h+1}}{\alpha^2\lambda_{h+1}-1}}\right\}. \end{aligned} \tag{2.19}$$

Furthermore, for a given constant $\delta : 0 < \delta < \alpha\lambda_1$, define

$$M_\beta = M_h\left(\lambda_1^{\beta} + \frac{1}{\alpha}\right)\left(\frac{\beta}{\alpha\lambda_1 - \delta}\right)^{\beta} e^{-\beta}. \tag{2.20}$$

THEOREM 2.1. *Under hypotheses* (1.4), (1.5), *let* $[\xi_0, \eta_0] \in D(A) \times D(A^{1+\beta})$ *and assume that system parameters,* $\delta, \alpha, \beta, \lambda_1, \lambda_h, \lambda_{h+1}, K_0$, *satisfy the following inequality conditions*: $0 < \delta < \alpha\lambda_1$, $0 < \beta \le 1/2$, *and*

$$\delta > \vartheta := \left(\frac{M_\beta K_0 \Gamma(\bar{\beta})}{2\lambda(\beta)}\right)^{1/\bar{\beta}}, \tag{2.21}$$

where $\bar{\beta} = 1 - \beta$. Then the estimate

$$|A\xi(t)|_0 + |A^-\eta(t)|_\beta \le K_1(t)(|A\xi_0|_0 + |A^-\eta_0|_\beta) + K_2|w|_\infty + K_3 \tag{2.22}$$

holds for some positive constants K_2, K_3, and

$$K_1(t) = \frac{e^{-(\delta-\vartheta)t}}{\bar{\beta}} + e^{-\alpha\lambda_1 t}\Gamma(\bar{\beta}). \tag{2.23}$$

Consequently, the solution $[u(t), \dot{u}(t)]$, given by $u = (\xi + \eta)/2$, has a global attractor in $X_1 \times X_\beta$: $\{[x, y] \in X_1 \times X_\beta : |x|_1 + |y|_\beta \le K_p(K_2|w|_\infty + K_3)\}$ for some $K_p > 0$.

Remark 2.1. In case $1/2 < \beta < 1$ the assertion of Theorem 2.1 holds if one substitutes the constant $\Gamma(\bar{\beta})$ by $\Gamma'(\bar{\beta}) := \Gamma(\bar{\beta})/(\sin\bar{\beta}\pi)^{\bar{\beta}}$ (cf. Appendix 1).

Remark 2.2. Obviously, if the constant K_0 is sufficiently small, then (2.21) must be satisfied. In §5, investigating a flexible beam model, we will derive an essential relation, described by its system parameters, from (2.21).

For convenience we give the following estimation of the constants:

$$\begin{aligned} K_p &= \frac{1}{2}\left(1 + \max\left\{\frac{1}{\alpha\lambda_1^\beta}, \frac{1}{\alpha\lambda_1^{1-\beta}}\right\}\right), \\ K_2 &= M_\beta \times \left[\delta\Gamma_1\left\{\frac{1}{\bar{\beta}(\delta-\vartheta)} + \frac{\Gamma(\bar{\beta})}{\delta}\right\} + (\delta e)^{-(1-\beta)}(1-\beta)^{-\beta}\Gamma(\bar{\beta}) + \frac{\Gamma_2}{\bar{\beta}}\right], \\ K_3 &= K_2K_0, \end{aligned} \tag{2.24}$$

where

$$\Gamma_1 = \int_0^\infty s^{-\beta}e^{-\delta s}ds, \quad \Gamma_2 = \int_0^\infty s^{-\beta}e^{-\vartheta s}ds. \tag{2.25}$$

Next, let $w(t)$ be a periodic function. Then we consider periodicity of solution $[u(t), \dot{u}(t)]$. As in Theorem 2.1, we also use the pair of functions $[\xi(t), \eta(t)] \in X_1 \times X_{1+\beta}$. Define the norm $|||[x, y]|||_{1,\beta}$ by

$$|||[x, y]|||_{1,\beta} := |Ax|_0 + |A^-y|_\beta,$$

which is equivalent to the $X_1 \times X_{1+\beta}$-norm. Then we can show the periodicity or the asymptotic periodicity of $[u(t), \dot{u}(t)]$ in $X_1 \times X_\beta$ by estimating $|||[\xi(t), \eta(t)]|||_{1,\beta}$.

THEOREM 2.2. *Let $w(t)$ be periodic: $w(t) = w(t+T)$ and assume the same hypotheses as Theorem 2.1. For a given constant $d > 0$ which satisfies*

$$d > K_2 r + K_3, \tag{2.26}$$

where $r := |w|_\infty$ and K_2, K_3 are the constants in (2.24), assume that

$$\delta > \vartheta' := \left(\frac{M_\beta k(d)\Gamma(\bar{\beta})}{2\lambda(\beta)}\right)^{1/\bar{\beta}}, \tag{2.27}$$

where $k(\cdot)$ is the locally Lipschitz coefficient in (1.4). Then there exists a unique T-periodic solution $[\xi_\infty(t), \eta_\infty(t)]$ such that

$$|||[\xi_\infty(t), \eta_\infty(t)]|||_{1,\beta} \le d, \quad t \ge 0, \tag{2.28}$$

and consequently, there exists a unique T-periodic solution $[u_\infty, \dot{u}_\infty]$:

$$|u_\infty(t)|_1 + |\dot{u}_\infty(t)|_\beta \le K_p d, \quad t \ge 0.$$

THEOREM 2.3. *Assume the same hypotheses as Theorem 2.2 and let $w(t)$ be asymptotically periodic:*

$$|w(t) - w_\infty(t)|_0 \to 0 \quad \textit{as } t \to \infty$$

for some T-periodic function $w_\infty : |w_\infty|_\infty \le r$. Then the solution $[u(t:w), \dot{u}(t:w)]$, starting with any initial state $[u_0, u_1] \in X_1 \times X_\beta$, converges to the T-periodic solution $[u_\infty(t:w_\infty), \dot{u}_\infty(t:w_\infty)]$ under the periodic perturbation w_∞:

$$(2.29) \quad |u(t:w) - u_\infty(t:w_\infty)|_1 + |\dot{u}(t:w) - \dot{u}_\infty(t:w_\infty)|_\beta \to 0 \quad \textit{as } t \to \infty.$$

3. Fundamental lemmas. In order to prove the main results we need the estimate of the norm of $[u(t), \dot{u}(t)]$ using the norm of $[\xi(t), \eta(t)]$. First, we show the equivalence of the norms $|u(t)|_\beta + |\dot{u}(t)|_\beta$ and $|A^+\xi|_\beta + |A^-\eta|_\beta$.

LEMMA 3.1. *For the solutions of* (2.5)–(2.6),

$$\xi \in C^1(0,T : D(A)) \cap C(0,T : D(A)),$$
$$\eta \in C^1(0,T : D(A^\beta)) \cap C(0,T : D(A^{1+\beta})),$$

and $u = (\xi + \eta)/2$, there exist constants $N_1, N_h > 0$ such that

$$(3.1) \quad N_1(|u(t)|_\beta + |\dot{u}(t)|_\beta) \le |A^+\xi(t)|_\beta + |A^-\eta(t)|_\beta \le N_h(|u(t)|_\beta + |\dot{u}(t)|_\beta).$$

Proof. From the definition of the operators A^+, A^- and (2.12), we have

$$(3.2) \quad \begin{aligned} |A^+\xi|_\beta + |A^-\eta|_\beta &\ge |(\alpha A - g(A))(u+v) + (\alpha A + g(A))(u-v)|_\beta \\ &= 2|\alpha A u - g(A)v|_\beta = 2|\dot{u}(t)|_\beta. \end{aligned}$$

Also, (2.16) yields

$$(3.3) \quad \begin{aligned} |A^+\xi|_\beta + |A^-\eta|_\beta &\ge |(\alpha A - g(A))(u+v) - (\alpha A + g(A))(u-v)|_\beta \\ &= 2|\alpha A v - g(A)u|_\beta \\ &= 2|\alpha A g^{-1}(A)\dot{u} + A g^{-1}(A) u|_\beta. \end{aligned}$$

Now we estimate the operator norm of $Ag^{-1}(A)$ by using the eigenvalues.
If $i \le h$, we have

$$|g^{-1}(\lambda_i)|\lambda_i = \sqrt{\frac{\lambda_i}{1-\alpha^2\lambda_i}} \ge \sqrt{\frac{\lambda_1}{1-\alpha^2\lambda_1}},$$

and if $i \ge h+1$, we have

$$|g^{-1}(\lambda_i)|\lambda_i = \sqrt{\frac{\lambda_i}{\alpha^2\lambda_i - 1}} \ge \frac{1}{\alpha}.$$

Put

$$N(\alpha, \lambda_1) := \min\left\{ \sqrt{\frac{\lambda_1}{1-\alpha^2\lambda_1}}, \frac{1}{\alpha} \right\}.$$

Then we can estimate

$$\begin{aligned}(3.4)\qquad |Ag^{-1}(A)(a\dot{u}+u)|_\beta^2 &= \sum_{i=1}^{\infty}\sum_{j=1}^{m_i}\lambda_i^{2\beta}\lambda_i^2|g^{-1}(\lambda_i)|^2|(\alpha\dot{u}+u,\varphi_{ij})|^2\\ &\geq N(\alpha,\lambda_1)^2|a\dot{u}+u|_\beta^2.\end{aligned}$$

On the other hand, it follows from (3.2) that

$$(3.5)\qquad |A^+\xi|_\beta+|A^-\eta|_\beta \geq \frac{2}{\alpha}|\alpha\dot{u}|_\beta \geq 2N(\alpha,\lambda_1)|\alpha\dot{u}|_\beta.$$

Combining (3.3), (3.4), and (3.5), and taking the sum with (3.2), we have

$$(3.6)\qquad |A^+\xi|_\beta+|A^-\eta|_\beta \geq \min\left\{1,\frac{N(\alpha,\lambda_1)}{2}\right\}(|u(t)|_\beta+|\dot{u}(t)|_\beta).$$

Since it follows from (2.1) that

$$N(\alpha,\lambda_1)\geq\sqrt{\lambda_1},$$

we obtain the left-hand side estimate in (3.1) by putting

$$N_1 := \min\left\{1,\frac{\sqrt{\lambda_1}}{2}\right\}.$$

To show the converse relation we also need to estimate $\|Ag^{-1}(A)\|$. If $i\leq h$,

$$\lambda_i^2|g^{-1}(\lambda_i)|^2=\frac{\lambda_i^2}{\lambda_i-\alpha^2\lambda_i^2}\leq\frac{1}{\frac{1}{\lambda_h}-\alpha^2},$$

and if $i\geq h+1$,

$$\lambda_i^2|g^{-1}(\lambda_i)|^2=\frac{\lambda_i^2}{\alpha^2\lambda_i^2-\lambda_i}\leq\frac{1}{\alpha^2-\frac{1}{\lambda_{h+1}}}.$$

Thus we have

$$(3.7)\qquad \lambda_i|g^{-1}(\lambda_i)|\leq\max\left\{\sqrt{\frac{\lambda_h}{1-\alpha^2\lambda_h}},\sqrt{\frac{\lambda_{h+1}}{\alpha^2\lambda_h-1}}\right\}=C_h.$$

Since the relations (2.12) and (2.16) yield

$$\begin{aligned}&|A^+\xi|_\beta+|A^-\eta|_\beta\\ &\quad\leq|(\alpha Au-g(A)v)+(\alpha Av-g(A)u)|_\beta+|(\alpha Au-g(A)v)+(-\alpha Av+g(A)u)|_\beta\\ &\quad=|-\dot{u}+(\alpha Ag^{-1}(A)\dot{u}+Ag^{-1}(A)u)|_\beta+|-\dot{u}-(\alpha Ag^{-1}(A)\dot{u}+Ag^{-1}(A)u)|_\beta,\end{aligned}$$

we can estimate

$$|A^+\xi|_\beta+|A^-\eta|_\beta\leq2|\dot{u}|_\beta+2C_h|\alpha\dot{u}+u|_\beta.$$

Thus we can complete the proof by putting

$$(3.8)\qquad N_h:=2\max\{1+\alpha C_h,C_h\}.\qquad\square$$

We prepare some inequality relations.

LEMMA 3.2. *For the operators A, A^+, A^-, the following inequalities hold:*

$$|Ax|_0 \geq \lambda_1^{1-\beta}|x|_\beta, \quad x \in X_1, \tag{3.9}$$

$$|A^+y|_\beta \geq \sqrt{\lambda_1}|y|_\beta, \quad y \in X_\beta, \tag{3.10}$$

$$|A^-z|_\beta \geq \sqrt{\lambda_1}|z|_\beta, \quad z \in X_{1+\beta}, \tag{3.11}$$

$$|A^-w|_\beta \geq \alpha\lambda_1^\beta|w|_1, \quad w \in X_{1+\beta}, \tag{3.12}$$

$$|A\xi|_0 \geq \alpha\lambda_1^{1-\beta}|A^+\xi|_\beta, \quad \xi \in X_1. \tag{3.13}$$

Proof. By the following spectral expansions:

$$|Ax|_0^2 = \sum_{i=1}^\infty \sum_{j=1}^{m_i} \lambda_i^2 |(x, \varphi_{ij})|^2,$$

$$|A^+y|_\beta^2 = \sum_{i=1}^\infty \sum_{j=1}^{m_i} |\nu_i|^2 \lambda_i^{2\beta} |(y, \varphi_{ij})|^2,$$

$$|A^-z|_\beta^2 = \sum_{i=1}^\infty \sum_{j=1}^{m_i} |\mu_i|^2 \lambda_i^{2\beta} |(z, \varphi_{ij})|^2,$$

we have (3.9), since

$$|Ax|_0^2 = \sum_{i=1}^\infty \sum_{j=1}^{m_i} \lambda_i^{2(1-\beta)} \lambda_i^\beta |(x, \varphi_{ij})|^2$$
$$\geq \lambda_1^{2(1-\beta)} \sum_{i=1}^\infty \sum_{j=1}^{m_i} \lambda_i^\beta |(x, \varphi_{ij})|^2,$$

and, for (3.10), (3.11), it is sufficient to estimate the eigenvalues ν_i, μ_i as follows. For ν_i, if $i \leq h$,

$$|\nu_i|^2 = |\alpha\lambda_i - g(\lambda_i)|^2 = \alpha^2\lambda_i^2 + (\lambda_i - \alpha^2\lambda_i^2)$$
$$= \lambda_i \geq \lambda_1,$$

and, if $i \geq h+1$,

$$|\nu_i|^2 = \left(\alpha\lambda_i - \sqrt{\alpha^2\lambda_i^2 - \lambda_i}\right)^2 = \left(\frac{\lambda_i}{\alpha\lambda_i + \sqrt{\alpha^2\lambda_i^2 - \lambda_i}}\right)^2$$
$$\geq \left(\frac{1}{2\alpha}\right)^2.$$

It follows from (2.1) that

$$|\nu_i| \geq \min\left\{\sqrt{\lambda_1}, \frac{1}{2\alpha}\right\} \geq \sqrt{\lambda_1}. \tag{3.14}$$

On the other hand, if $i \leq h$, we have

$$|\mu_i|^2 = |\alpha\lambda_i + g(\lambda_i)|^2 = \lambda_i \geq \lambda_1,$$

and, if $i \geq h+1$,

$$\begin{aligned}|\mu_i|^2 &= |\alpha\lambda_i + \sqrt{\alpha^2\lambda_i^2 - \lambda_i}|^2 \\ &= \left(\frac{\lambda_i}{\alpha\lambda_i - \sqrt{\alpha^2\lambda_i^2 - \lambda_i}}\right)^2 \\ &\geq \left(\frac{\lambda_i}{\alpha\lambda_i + \alpha\lambda_i}\right)^2 = \left(\frac{1}{2\alpha}\right)^2.\end{aligned}$$

For (3.12), consider the estimate

$$|\mu_i|^2 = \left|\alpha\lambda_i + \sqrt{\alpha^2\lambda_i^2 - \lambda_i}\right|^2 \geq \alpha^2\lambda_i^2.$$

Then (3.13) is an easy consequence of the estimate $|\nu_i|^2 \leq (1/\alpha)^2$. □

Given Lemma 3.1 and Lemma 3.2, we can estimate $|u(t)|_1 + |\dot{u}(t)|_\beta$ by the norm $|A\xi(t)|_0 + |A^-\eta(t)|_\beta$.

LEMMA 3.3. *There exists a constant $K_p > 0$ such that*

$$|u(t)|_1 + |\dot{u}(t)|_\beta \leq K_p(|A\xi(t)|_0 + |A^-\eta(t)|_\beta), \tag{3.15}$$

where K_p is given in (2.24).

Proof. From Lemma 3.2 we have

$$\begin{aligned}|u(t)|_1 &\leq \frac{1}{2}(|\xi(t)|_1 + |\eta(t)|_1) \\ &\leq \frac{1}{2}|A\xi(t)|_0 + \frac{1}{2\alpha\lambda_1^\beta}|A^-\eta(t)|_\beta, \\ |A^+\xi(t)|_\beta &\leq \frac{1}{\alpha\lambda_1^{1-\beta}}|A\xi(t)|_0.\end{aligned}$$

The estimate (3.2) and the above inequality relations yield (3.15). □

Since the definitions (2.19) and (2.20) imply that $\alpha C_h \leq M_h$, the estimate

$$(M_h\lambda_1^\beta + C_h)\left(\frac{\beta}{\alpha\lambda_1 - \delta}\right)^\beta e^{-\beta} \leq M_\beta \tag{3.16}$$

holds. Now we prepare the estimate, which corresponds to the well-known estimate of the operator norm of an analytic semigroup and its generator.

LEMMA 3.4. *For a given constant δ : $0 < \delta < \alpha\lambda_1$, we have the following estimate*:

$$|AS_1(t)g^{-1}(A)y|_0 + |A^-A^\beta S_2(t)g^{-1}(A)y|_0 \leq M_\beta e^{-\delta t}t^{-\beta}|y|_0, \quad y \in L^2(\Omega). \tag{3.17}$$

Proof. By using spectral expansion we have

$$\begin{aligned}|A^-A^\beta S_2(t)g^{-1}(A)y|_0^2 = &\left(\sum_{i=1}^{\infty}\sum_{j=1}^{m_i}\lambda_i^{2\beta}|\mu_i|^2|g^{-1}(\lambda_i)|^2 e^{-2(\mathrm{Re}[\mu_i]-\delta)t}t^{2\beta}|(y,\varphi_{ij})|^2\right) \\ &\times e^{-2\delta t}t^{-2\beta}.\end{aligned}$$

If $i \le h$,

$$|(\alpha\lambda_i + g(\lambda_i))g^{-1}(\lambda_i)|^2 = \left(\frac{\alpha^2\lambda_i^2}{\lambda_i - \alpha^2\lambda_i^2} + 1\right) \le \frac{1}{1-\alpha^2\lambda_h},$$

and, if $i \ge h+1$,

$$|(\alpha\lambda_i + g(\lambda_i))g^{-1}(\lambda_i)|^2 = \left(\frac{\alpha\lambda_i}{\sqrt{\alpha^2\lambda_i^2 - \lambda_i}} + 1\right)^2 \le \left(\frac{\alpha\lambda_{h+1}}{\sqrt{\alpha^2\lambda_{h+1}^2 - \lambda_{h+1}}} + 1\right)^2 .$$

Furthermore, since we can easily see from elementary calculations that each term $\lambda_i^{2\beta} e^{-2(\mathrm{Re}[\mu_i]-\delta)t} t^{2\beta}$ takes its maximal value at $t = \beta/(\mathrm{Re}[\mu_i] - \delta)$, we have

$$|A^- A^\beta S_2(t) g^{-1}(A) y|_0^2 \le M_h^2 \left(\sum_{i=1}^{\infty}\sum_{j=1}^{m_i} e^{-2\beta}\left(\frac{\lambda_i\beta}{\mathrm{Re}[\mu_i]-\delta}\right)^{2\beta} |(y,\varphi_{ij})|^2\right) e^{-2\delta t} t^{-2\beta}.$$

Since

$$\begin{aligned}\frac{\lambda_i\beta}{\mathrm{Re}[\mu_i]-\delta} &= \begin{cases} \lambda_i\beta/(\alpha\lambda_i - \delta), & \text{if } i \le h, \\ \lambda_i\beta/(\alpha\lambda_i + g(\lambda_i) - \delta), & \text{if } i \ge h+1 \end{cases} \\ &\le \frac{\lambda_1\beta}{\alpha\lambda_1 - \delta},\end{aligned}$$

we obtain

$$\begin{aligned}(3.18)\qquad &|A^- A^\beta S_2(t) g^{-1}(A) y|_0^2 \\ &\le M_h^2 e^{-2\beta}\left(\frac{\lambda_1\beta}{\alpha\lambda_1-\delta}\right)^{2\beta}\left(\sum_{i=1}^{\infty}\sum_{j=1}^{m_i} |(y,\varphi_{ij})|^2\right) e^{-2\delta t} t^{-2\beta}.\end{aligned}$$

On the other hand, applying the same argument as above, we have

$$\begin{aligned}(3.19)\ &|AS_1(t) g^{-1}(A) y|_0^2 \\ &= \left(\sum_{i=1}^{\infty}\sum_{j=1}^{m_i} \lambda_i^2 |g^{-1}(\lambda_i)|^2 e^{-2(\mathrm{Re}[\nu_i]-\delta)t} t^{2\beta} |(y,\varphi_{ij})|^2\right) e^{-2\delta t} t^{-2\beta} \\ &\le \left(\sum_{i=1}^{\infty}\sum_{j=1}^{m_i} \lambda_i^2 e^{-2\beta} |g^{-1}(\lambda_i)|^2 \left(\frac{\beta}{\mathrm{Re}[\nu_i]-\delta}\right)^{2\beta} |(y,\varphi_{ij})|^2\right) e^{-2\delta t} t^{-2\beta}.\end{aligned}$$

Since, if $i \ge h+1$,

$$\begin{aligned}\mathrm{Re}[\nu_i] = \alpha\lambda_i - g(\lambda_i) &= \frac{\lambda_i}{\alpha\lambda_i + \sqrt{\alpha^2\lambda_i^2 - \lambda_i}} \\ &\ge \frac{1}{2\alpha} \ge \alpha\lambda_1,\end{aligned}$$

we have

$$\frac{\beta}{\mathrm{Re}[\nu_i]-\delta} \le \frac{\beta}{\alpha\lambda_1 - \delta}$$

for every i. It follows from (3.7) that

$$(3.20)\quad |AS_1(t)g^{-1}(A)y|^2 \le C_h^2 e^{-2\beta}\left(\frac{\beta}{\alpha\lambda_1 - \delta}\right)^{2\beta}\left(\sum_{i=1}^{\infty}\sum_{j=1}^{m_i}|(y,\varphi_{ij})|^2\right)e^{-2\delta t}t^{-2\beta}.$$

Thus, combining (3.18) with (3.20) and using (3.16), we obtain the conclusion. □

4. Proofs of main theorems. In this section we give the proofs of our main theorems.

Proof of Theorem 2.1. From (2.5) and (2.6) we have

$$|A\xi(t)|_0 \le |S_1(t)A\xi_0|_0 + \int_0^t \left|AS_1(t-s)g^{-1}(A)\left\{F\left(\frac{\xi+\eta}{2}\right) + w(s)\right|_0\right\}ds,$$
$$|A^-\eta(t)|_\beta \le |S_2(t)A^-\eta_0|_\beta + \int_0^t \left|A^-A^\beta S_2(t-s)g^{-1}(A)\left\{F\left(\frac{\xi+\eta}{2}\right) + w(s)\right\}\right|_0 ds.$$

Summing up and using Lemma 3.2, Lemma 3.4, and (1.5), we obtain

$$\begin{aligned}
&|A\xi(t)|_0 + |A^-\eta(t)|_\beta \\
&\quad\le e^{-\alpha\lambda_1 t}(|A\xi_0|_0 + |A^-\eta_0|_\beta) + \int_0^t M_\beta e^{-\delta(t-s)}(t-s)^{-\beta}(K_0 + |w(s)|_0)ds \\
&\qquad + \int_0^t M_\beta e^{-\delta(t-s)}(t-s)^{-\beta}\frac{K_0}{2\lambda(\beta)}(|A\xi(s)|_0 + |A^-\eta(s)|_\beta)ds.
\end{aligned}$$

Multiplying each term by $e^{\delta t}$ and putting

$$\begin{aligned}
a(t) &:= e^{-(\alpha\lambda_1-\delta)t}(|A\xi_0|_0 + |A^-\eta_0|_\beta) + M_\beta(K_0 + |w|_\infty)\int_0^t e^{\delta\sigma}(t-\sigma)^{-\beta}d\sigma, \\
y(t) &:= e^{\delta t}(|A\xi(t)|_0 + |A^-\eta(t)|_\beta), \\
b &:= \frac{M_\beta K_0}{2\lambda(\beta)}
\end{aligned}$$

in the Gronwall inequality, introduced in Appendix 1, we obtain the following estimate:

$$\begin{aligned}
&|A\xi(t)|_0 + |A^-\eta(t)|_\beta \\
&\quad\le E(\vartheta t)e^{-\delta t}(|A\xi_0|_0 + |A^-\eta_0|_\beta) \\
&\qquad + \int_0^t E(\vartheta(t-s))\left\{M_\beta(K_0 + |w|_\infty)e^{-\delta t}\left[\delta e^{\delta s}\int_0^s \sigma^{-\beta}e^{-\delta\sigma}d\sigma + s^{-\beta}\right]\right. \\
&\qquad\qquad\left. - e^{-\delta t}(\alpha\lambda_1 - \delta)(|A\xi_0|_0 + |A^-\eta_0|_\beta)e^{-(\alpha\lambda_1-\delta)s}\right\}ds
\end{aligned}$$

where, as we can see in Appendix 1 also

$$\vartheta = [b\Gamma(\bar\beta)]^{1/\bar\beta}, \quad E(z) := \sum_{n=0}^{\infty}\frac{z^{n\bar\beta}}{\Gamma(n\bar\beta+1)}, \quad E(z) \le \frac{e^z}{\bar\beta} + \Gamma(\bar\beta), \quad z \ge 0.$$

Thus, we have the following sequence of estimation:

$$\begin{aligned}
&|A\xi(t)|_0 + |A^- \eta(t)|_\beta \\
&\quad \le (|A\xi_0|_0 + |A^-\eta_0|_\beta) e^{-\delta t} \left\{ \frac{e^{\vartheta t}}{\bar{\beta}} + \Gamma(\bar{\beta}) \right\} \\
&\quad\quad + M_\beta (K_0 + |w|_\infty) e^{-\delta t} \int_0^t \left(\frac{e^{\vartheta(t-s)}}{\bar{\beta}} + \Gamma(\bar{\beta}) \right) (\delta e^{\delta s}\Gamma_1 + s^{-\beta}) ds \\
&\quad\quad - e^{-\delta t}(\alpha\lambda_1 - \delta)(|A\xi_0|_0 + |A^-\eta_0|_\beta) \int_0^t \left(\frac{e^{\vartheta(t-s)}}{\bar{\beta}} + \Gamma(\bar{\beta}) \right) e^{-(\alpha\lambda_1-\delta)s} ds \\
&\quad \le M_\beta(K_0 + |w|_\infty) \\
&\quad\quad \times \left[\delta\Gamma_1 \left\{ \frac{1 - e^{-(\delta-\vartheta)t}}{\bar{\beta}(\delta-\vartheta)} + \frac{\Gamma(\bar{\beta})(1-e^{-\delta t})}{\delta} \right\} + e^{-\delta t} \frac{t^{1-\beta}}{1-\beta}\Gamma(\bar{\beta}) + e^{-(\delta-\vartheta)t}\frac{\Gamma_2}{\bar{\beta}} \right] \\
&\quad\quad + (|A\xi_0|_0 + |A^-\eta_0|_\beta) e^{-\delta t} \left\{ \frac{e^{\vartheta t}}{\bar{\beta}} + \Gamma(\bar{\beta}) \right\} - (\alpha\lambda_1 - \delta)(|A\xi_0|_0 + |A^-\eta_0|_\beta) \\
&\quad\quad \times \left[\frac{e^{-(\delta-\vartheta)t}}{\bar{\beta}} \int_0^t e^{-(\vartheta+\alpha\lambda_1-\delta)s} ds + e^{-\delta t}\Gamma(\bar{\beta}) \int_0^t e^{-(\alpha\lambda_1-\delta)s} ds \right] \\
&\quad \le M_\beta(K_0 + |w|_\infty) \\
&\quad\quad \times \left[\delta\Gamma_1 \left\{ \frac{1 - e^{-(\delta-\vartheta)t}}{\bar{\beta}(\delta-\vartheta)} + \frac{\Gamma(\bar{\beta})(1-e^{-\delta t})}{\delta} \right\} + e^{-\delta t} \frac{t^{1-\beta}}{1-\beta}\Gamma(\bar{\beta}) + e^{-(\delta-\vartheta)t}\frac{\Gamma_2}{\bar{\beta}} \right] \\
&\quad\quad + (|A\xi_0|_0 + |A^-\eta_0|_\beta) \left\{ \frac{e^{-(\delta-\vartheta)t}}{\bar{\beta}} \left(\frac{\vartheta + (\alpha\lambda_1 - \delta)e^{-(\vartheta+\alpha\lambda_1-\delta)t}}{\vartheta + \alpha\lambda_1 - \delta} \right) + \Gamma(\bar{\beta})e^{-\alpha\lambda_1 t} \right\},
\end{aligned}$$

where Γ_1, Γ_2 are the constants, given in (2.25). Hence, we can estimate

$$|A\xi(t)| + |A^-\eta(t)|_\beta \le K_1(t)(|A\xi_0|_0 + |A^-\eta_0|_\beta) + K_2|w|_\infty + K_3 \tag{4.1}$$

using the constants K_2, K_3, defined in (2.24). Thus we can complete the proof by using Lemma 3.3. □

Proof of Theorem 2.2. From Theorem 2.1 we can assume that there exists a solution $[\xi, \eta]$ with an initial condition $[\xi_0, \eta_0]$ in $X_1 \times X_{1+\beta}$:

$$\|[\xi(t), \eta(t)]\|_{1,\beta} \le d, \quad t \ge 0.$$

Then we show that $[\xi_n(t), \eta_n(t)] := [\xi(t+nT), \eta(t+nT)]$ converges to a T-periodic solution $[\xi_\infty(t), \eta_\infty(t)]$ as $n \to \infty$.

By using (2.5) and (2.6), we have

$$\begin{aligned}
&|A(\xi_n - \xi_{n+m})(t)|_0 \\
&\quad \le |S_1(nT)A(\xi - \xi_m)(t)|_0 \\
&\quad\quad + \int_0^{nT} \left| AS_1(nT-s)g^{-1}(A) \left[F\left(\frac{\xi+\eta}{2}(s+t) \right) - F\left(\frac{\xi_m + \eta_m}{2}(s+t) \right) \right] \right|_0 ds, \\
&|A^-(\eta_n - \eta_{n+m})(t)|_\beta \\
&\quad \le |S_2(nT)A^-(\eta - \eta_m)(t)|_\beta \\
&\quad\quad + \int_0^{nT} \left| A^- A^\beta S_2(nT-s)g^{-1}(A) \left[F\left(\frac{\xi+\eta}{2}(s+t) \right) - F\left(\frac{\xi_m + \eta_m}{2}(s+t) \right) \right] \right|_0 ds.
\end{aligned}$$

It follows from Lemma 3.4, (1.4), and Lemma 3.2 that

$$\begin{aligned}&|A(\xi_n-\xi_{n+m})(t)|_0+|A^-(\eta_n-\eta_{n+m})(t)|_\beta\\&\quad\le e^{-\alpha\lambda_1 nT}(|A(\xi-\xi_m)(t)|_0+|A^-(\eta-\eta_m)(t)|_\beta)\\&\quad+\int_0^{nT}M_\beta\frac{k(d)}{2\lambda(\beta)}e^{-\delta(nT-s)}(nT-s)^{-\beta}\\&\quad\times\{|A(\xi-\xi_m)(s+t)|_0+|A^-(\eta-\eta_m)(s+t)|_\beta\}ds.\end{aligned}$$

Thus the Gronwall inequality and the same argument as in Theorem 2.1 give

$$\begin{aligned}&|A(\xi_n-\xi_{n+m})(t)|_0+|A^-(\eta_n-\eta_{n+m})(t)|_\beta\\&\le\left[\frac{e^{-(\delta-\vartheta')nT}}{\bar\beta}+\Gamma(\bar\beta)e^{-\alpha\lambda_1 nT}\right](|A(\xi-\xi_m)(t)|_0+|A^-(\eta-\eta_m)(t)|_\beta)\\&\le 2d\left[\frac{e^{-(\delta-\vartheta')nT}}{\bar\beta}+\Gamma(\bar\beta)e^{-\alpha\lambda_1 nT}\right].\end{aligned}$$

Hence the sequence $[\xi_n(t),\eta_n(t)]$ is a Cauchy sequence in $BC(R^+:X_1\times X_{1+\beta})$, the space of uniformly bounded, continuous, and $X_1\times X_{1+\beta}$-valued functions. So there exists $[\xi_\infty(t),\eta_\infty(t)]$ such that

$$[\xi_n(t),\eta_n(t)]\to[\xi_\infty,\eta_\infty]\ \text{ in }\ BC(R^+:X_1\times X_{1+\beta}).$$

By taking the limit $n\to\infty$ of the mild formulas:

$$\begin{aligned}\xi(t+nT)&=S_1(t)\xi(nT)+\int_0^t S_1(t-s)g^{-1}(A)\left[F\left(\frac{\xi+\eta}{2}(s+nT)\right)+w(s)\right]ds,\\\eta(t+nT)&=S_2(t)\eta(nT)-\int_0^t S_2(t-s)g^{-1}(A)\left[F\left(\frac{\xi+\eta}{2}(s+nT)\right)+w(s)\right]ds,\end{aligned}$$

we can show that $[\xi_\infty(t),\eta_\infty(t)]$ satisfies the mild formulas (2.5), (2.6) with the initial state $[\xi_\infty(0),\eta_\infty(0)]$.

Furthermore, T-periodicity of $[\xi_\infty(t),\eta_\infty(t)]$ holds, since

$$\begin{aligned}[\xi_\infty(t+T),\eta_\infty(t+T)]&=\lim_{n\to\infty}[\xi_\infty(t+T+nT),\eta_\infty(t+T+nT)]\\&=\lim_{n\to\infty}[\xi_\infty(t+(n+1)T),\eta_\infty(t+(n+1)T)]\\&=[\xi_\infty(t),\eta_\infty(t)].\end{aligned}$$

As for the uniqueness of the T-periodic solution, since we have the following estimate for a sufficiently large N:

$$\begin{aligned}&\|[\xi_\infty(t+NT),\eta_\infty(t+NT)]-[\xi'_\infty(t+NT),\eta'_\infty(t+NT)]\|_{1,\beta}\\&=\|[\xi_\infty(t),\eta_\infty(t)]-[\xi'_\infty(t),\eta'_\infty(t)]\|_{1,\beta}\\&\le\left[\frac{e^{-(\delta-\vartheta')NT}}{\bar\beta}+\Gamma(\bar\beta)e^{-\alpha\lambda_1 NT}]\|[\xi_\infty(t),\eta_\infty(t)\right]-[\xi'_\infty(t),\eta'_\infty(t)]\|_{1,\beta},\end{aligned}$$

we can complete the proof. □

Proof of Theorem 2.3. Instead of estimating the solution $u(\cdot : w)$, it is sufficient to show the convergence of the pair of functions $[\xi(t), \eta(t)]$ to a pair of T-periodic functions $[\xi_\infty, \eta_\infty]$ in $X_1 \times X_{1+\beta}$-norm. Here we can also assume that

$$\| [\xi(t), \eta(t)] \|_{1,\beta} \le d, \quad t \ge 0.$$

Let N be a large integer which satisfies

$$\frac{e^{-(\delta-\vartheta')NT}}{\bar{\beta}} + \Gamma(\bar{\beta}) e^{-\alpha\lambda_1 NT} < 1$$

and define the following sequences:

$$\begin{aligned} w_m(t) &= w(t + mNT), \\ \xi_m(t) &= \xi(t + mNT), \\ \eta_m(t) &= \eta(t + mNT), \quad m = 0, 1, 2, \ldots. \end{aligned}$$

Then, applying the same argument as that in the proof of Theorem 2.2 to the difference of the solutions

$$[\xi(t + (m+1)NT), \eta(t + (m+1)NT)] - [\xi_\infty(t + (m+1)NT), \eta_\infty(t + (m+1)NT)],$$

which start with the initial values

$$[\xi(t + mNT), \eta(t + mNT)], \quad [\xi_\infty(t + mNT), \eta_\infty(t + mNT)],$$

respectively, we have

$$\begin{aligned} &\| [(\xi_{m+1} - \xi_\infty)(t), (\eta_{m+1} - \eta_\infty)(t)] \|_{1,\beta} \\ &\quad = \| [(\xi - \xi_\infty)(t + (m+1)NT), (\eta - \eta_\infty)(t + (m+1)NT)] \|_{1,\beta} \\ &\quad \le \left\{ \frac{e^{-(\delta-\vartheta)NT}}{\bar{\beta}} + \Gamma(\bar{\beta}) e^{-\alpha\lambda_1 NT} \right\} \| [(\xi - \xi_\infty)(t + mNT), (\eta - \eta_\infty)(t + mNT)] \|_{1,\beta} \\ &\qquad + \left[\delta\Gamma_1 \left\{ \frac{1}{\bar{\beta}(\delta - \vartheta)} + \frac{\Gamma(\bar{\beta})}{\delta} \right\} + (\delta e)^{-(1-\beta)} (1-\beta)^{-\beta} \Gamma(\bar{\beta}) + \frac{\Gamma_2}{\bar{\beta}} \right] \\ &\qquad \times M_\beta \sup\{ |w(s) - w_\infty(s)|_0 : t + mNT \le s \le t + (m+1)NT \}. \end{aligned}$$

Put

$$\begin{aligned} \varphi_m &= \| [(\xi_m - \xi_\infty)(t), (\eta_m - \eta_\infty)(t)] \|_{1,\beta}, \\ K &= \left[\frac{e^{-(\delta-\vartheta)NT}}{\bar{\beta}} + \Gamma(\bar{\beta}) e^{-\alpha\lambda_1 NT} \right] < 1 \end{aligned}$$

and for a small constant $\varepsilon > 0$, take a large number $m_0 : m \ge m_0 \Longrightarrow$

$$\begin{aligned} &\left[\delta\Gamma_1 \left\{ \frac{1}{\bar{\beta}(\delta - \vartheta)} + \frac{\Gamma(\bar{\beta})}{\delta} \right\} + (\delta e)^{-(1-\beta)} (1-\beta)^{-\beta} \Gamma(\bar{\beta}) + \frac{\Gamma_2}{\bar{\beta}} \right] \\ &\qquad \times M_\beta \sup\{ |w(s) - w_\infty(s)| : t + mNT \le s \le t + (m+1)NT \} < \varepsilon. \end{aligned}$$

Then we have

$$\varphi_m \le K\varphi_{m-1} + \varepsilon \le \cdots \le K^m \varphi_0 + \varepsilon \frac{1 - K^m}{1 - K}.$$

Since $\varphi_0 = |||[(\xi - \xi_\infty)(t), (\eta - \eta_\infty)(t)]|||_{1,\beta} \leq 2d$, we can show that for every small $\varepsilon > 0$, there exists a large number $m_1 : m \geq m_1 \Longrightarrow$

$$|||[\xi(t + mNT), \eta(t + mNT)] - [\xi_\infty(t + mNT), \eta_\infty(t + mNT)]|||_{1,\beta} < \varepsilon$$

for every $t \in [0, NT]$, that is,

$$|||[\xi(t), \eta(t)] - [\xi_\infty(t), \eta_\infty(t)]|||_{1,\beta} < \varepsilon$$

for every $t \geq m_1 NT$. Estimating $|u(t) - u_\infty(t)|_1$ and $|\dot{u}(t) - \dot{u}_\infty(t)|_\beta$ according to the argument in (3.2) and Lemma 3.3, we have

$$|u(t) - u_\infty(t)|_1 + |\dot{u}(t) - \dot{u}_\infty(t)|_\beta \leq K_p |||[\xi(t), \eta(t)] - [\xi_\infty(t), \eta_\infty(t)]|||_{1,\beta},$$

which completes the proof. □

5. Flexible beam. We consider the equation of motion of slender and flexible structures with internal viscous damping and nonlinear forcing, determined by displacement $u(t,x)$ and bending force $u_{xx}(t,x)$, under a periodic perturbation:

$$\text{(5.1)} \quad \frac{\partial^2 u(t,x)}{\partial t^2} + 2\alpha \frac{\partial^5 u(t,x)}{\partial t \partial x^4} + \frac{\partial^4 u(t,x)}{\partial x^4} = f\left(x, u(t,x), \frac{\partial^2 u(t,x)}{\partial x^2}\right) + w(t,x),$$

where $0 < x < L$. The beam is clamped at one end, $x = 0$, and at the free end, $x = L$, the bending moment and the shearing force vanish. Then the boundary conditions and the initial conditions are given by

$$\text{(5.2)} \quad \begin{aligned} &u(t,0) = \frac{\partial u(t,0)}{\partial x} = 0, \\ &\frac{\partial^2 u(t,L)}{\partial x^2} + 2\alpha \frac{\partial^3 u(t,L)}{\partial x^2 \partial t} = 0, \\ &\frac{\partial^3 u(t,L)}{\partial x^3} + 2\alpha \frac{\partial^4 u(t,L)}{\partial x^3 \partial t} = 0, \end{aligned}$$

$$\text{(5.3)} \quad u(0,x) = u_0(x), \quad \frac{\partial u(0,x)}{\partial t} = u_1(x),$$

and the periodic perturbation satisfies $w(t+T) = w(t)$. We define an operator A in $L^2(0,L)$ by

$$\begin{aligned} &D(A) = \{u \in H^4(0,L) : u(0) = u_x(0) = 0, \quad u_{xx}(L) = u_{xxx}(L) = 0\}, \\ &Au = \frac{\partial^4 u}{\partial x^4}. \end{aligned}$$

Let γ_i be the solutions of

$$\cosh \gamma \cos \gamma + 1 = 0$$

such that $0 < \gamma_1 < \gamma_2 < \ldots$. Then the eigenvalues of A are given by

$$\lambda_i = \left(\frac{\gamma_i}{L}\right)^4, \quad i = 1, 2, \ldots$$

(cf. [10]).

We assume that the nonlinear function $f(x,u,v) : R \times R \times R \to R$ satisfies the growth condition

$$|f(x,u,v)| \leq k_0(|u|+|v|) \quad \text{for some } k_0 > 0$$

and the following Lipschitz and locally Lipschitz continuity: there exists positive constants $k_0(c), k$ such that

$$|f(x,u,v) - f(x,u',v')| \leq k_0(c)|u-u'| + k|v-v'| \quad \text{for } |u|,|u'| \leq c, \quad v,v' \in R.$$

Define a nonlinear mapping $F : D(A^{\frac{1}{2}}) \to L^2(0,L)$ by

$$F(u)(x) = f(x, u(x), u_{xx}(x)).$$

Then, since the injections

$$D(A^{\frac{1}{2}}) \hookrightarrow H^2(0,L) \hookrightarrow C(0,L)$$

are continuous, the conditions (1.4) and (1.5) hold for the constant $\beta = 1/2$ and some constants $K_0, k(c)$.

Now we investigate the inequality conditions in the case where $\beta = 1/2$. Our purpose is to find some relations among the constants $\lambda_1, \lambda_h, \lambda_{h+1}, \alpha, K_0, k(d)$, where $d > K_2 r + K_3$ and K_2, K_3 are given in (2.24). Let $\delta := \alpha\lambda_1/2$. Then we can describe (2.19) by

$$M_{\frac{1}{2}} = M_h \left(\sqrt{\lambda_1} + \frac{1}{\alpha}\right)\left(\frac{1}{\alpha\lambda_1}\right)^{\frac{1}{2}} e^{-\frac{1}{2}} \tag{5.4}$$

and it follows that

$$\begin{aligned} \vartheta &= \left(M_{\frac{1}{2}} K_0 \frac{\sqrt{\pi}}{2\sqrt{\lambda_1}}\right)^2 \\ &= M_h^2 \left(\sqrt{\lambda_1} + \frac{1}{\alpha}\right)^2 \frac{1}{\alpha\lambda_1^2} \frac{K_0^2\pi}{4e}. \end{aligned}$$

Thus the condition $\alpha\lambda_1/2 > \vartheta$ can be described by

$$\alpha^2\lambda_1^2 > 2M_h^2 \left(1 + \frac{1}{\alpha\sqrt{\lambda_1}}\right)^2 \frac{K_0^2\pi}{4e}.$$

Taking the square root yields

$$\alpha(\sqrt{\lambda_1})^3 - K_0\sqrt{\frac{\pi}{2e}} M_h \sqrt{\lambda_1} - K_0\sqrt{\frac{\pi}{2e}} M_h \frac{1}{\alpha} > 0. \tag{5.5}$$

Hence we can admit the first eigenvalue $\lambda_1 : 0 < \lambda_1 < 1/(2\alpha^2)$, which satisfies the condition $\delta > \vartheta$, if

$$\alpha\left(\frac{1}{\sqrt{2}\alpha}\right)^3 - K_0\sqrt{\frac{\pi}{2e}} M_h \frac{1}{\sqrt{2}\alpha} - K_0\sqrt{\frac{\pi}{2e}} M_h \frac{1}{\alpha} > 0. \tag{5.6}$$

It follows that

$$\alpha M_h < \frac{1}{(\sqrt{2}+1)K_0\sqrt{\frac{2\pi}{e}}}. \tag{5.7}$$

Hereafter, we use the notations

$$\kappa = (\sqrt{2}+1)K_0\sqrt{\frac{2\pi}{e}},$$
$$\kappa' = (\sqrt{2}+1)k(d)\sqrt{\frac{2\pi}{e}}.$$

Considering the definition of M_h and (5.7), we have

$$\frac{\alpha}{\sqrt{1-\alpha^2\lambda_h}} < \frac{1}{\kappa}, \tag{5.8}$$

$$\alpha\left(\sqrt{\frac{\alpha^2\lambda_{h+1}}{\alpha^2\lambda_{h+1}-1}}+1\right) < \frac{1}{\kappa}. \tag{5.9}$$

By (5.8) we can derive the conditions on $\lambda_h, \alpha, \kappa$:

$$\lambda_h < \frac{1}{\alpha^2} - \kappa^2, \quad \alpha\kappa < 1 \tag{5.10}$$

and, assuming $\alpha\kappa < 1/2$, from (5.9) we obtain

$$\lambda_{h+1} > \frac{1}{\alpha^2} + \frac{\kappa^2}{1-2\alpha\kappa}. \tag{5.11}$$

We note that as the values

$$\frac{1}{\alpha^2} - \lambda_h, \;\; \lambda_{h+1} - \frac{1}{\alpha^2}$$

become sufficiently large, $M_h \downarrow 2$.

When (5.10) and (5.11) are satisfied, the first eigenvalue λ_1 can be estimated as follows. The third-order algebraic inequality

$$ax^3 - bx - \frac{b}{a} > 0, \;\; a, b > 0 \tag{5.12}$$

admits a solution

$$x > \frac{b}{3a\left(\frac{b}{2a^2}+\sqrt{-\frac{b^3}{27a^3}+\frac{b^2}{4a^4}}\right)^{\frac{1}{3}}} + \left(\frac{b}{2a^2}+\sqrt{-\frac{b^3}{27a^3}+\frac{b^2}{4a^4}}\right)^{\frac{1}{3}}$$

in the positive real $x > 0$. For sufficiently small $a > 0$, a rough estimation,

$$\begin{aligned}\sqrt{-\frac{b^3}{27a^3}+\frac{b^2}{4a^4}} &= \frac{b}{2a^2}\sqrt{1-\frac{4ab}{27}} \\ &\simeq \frac{b}{2a^2},\end{aligned}$$

gives a sufficient condition for (5.12)

$$x > \frac{b^{\frac{2}{3}}}{3a^{\frac{1}{3}}} + \frac{b^{\frac{1}{3}}}{a^{\frac{2}{3}}},$$

where we take a positive real value of each fractional power 1/3. Considering the case

$$a = \alpha, \quad b = \sqrt{\frac{\pi}{2e}} K_0 M_h$$

with α sufficiently small, we can then estimate λ_1, which satisfies our inequality conditions, as follows:

$$\left(1 + \frac{1}{3}(K_0 M_h \alpha)^{\frac{1}{3}} \left(\frac{\pi}{2e}\right)^{\frac{1}{6}}\right)^2 (K_0 M_h)^{\frac{2}{3}} \left(\frac{\pi}{2e}\right)^{\frac{1}{3}} \left(\frac{1}{\alpha}\right)^{\frac{4}{3}} < \lambda_1 < \frac{1}{2\alpha^2}. \tag{5.13}$$

For sufficiently small $\alpha > 0$, it follows that

$$C_1 \left(\frac{K_0 M_h}{\alpha^2}\right)^{\frac{2}{3}} < \lambda_1 < \frac{1}{2\alpha^2}$$

for some constant $C_1 > 0$. Furthermore, if the values $1/\alpha^2 - \lambda_h$ and $\lambda_{h+1} - 1/\alpha^2$ are sufficiently large, for instance,

$$\frac{1}{\alpha^2} - \lambda_h \simeq \frac{K}{\alpha^2},$$
$$\lambda_{h+1} - \frac{1}{\alpha^2} \simeq \frac{K}{\alpha^2},$$

for some large $K > 0$, we have

$$M_h \simeq 1 + \sqrt{1 + \frac{1}{K}}.$$

Then we can estimate the stability condition for the first eigenvalue:

$$C_2 \left(\frac{K_0}{\alpha^2}\right)^{\frac{2}{3}} < \lambda_1 < \frac{1}{2\alpha^2} \tag{5.14}$$

for some constant $C_2 > 0$.

For the condition $\delta > \vartheta'$, we can derive a similar estimate as (5.14), substituting K_0 with $k(d)$ and κ with κ'. Using a sufficiently small constant α, we can roughly estimate the first eigenvalue λ_1 as follows. Assume that $\delta - \vartheta \simeq \delta \simeq \alpha\lambda_1$. Then it follows from (2.24) that we can estimate the order of the constants $K_2, K_3 \simeq M_{\frac{1}{2}} \delta^{-1/2}$. Thus, when the perturbation of w is comparatively small, $r << \alpha^{-1}$, we may assume that $d \simeq M_{\frac{1}{2}} \delta^{-1/2}$. Since we can roughly estimate $M_{\frac{1}{2}} \simeq \alpha^{-3/2} \lambda_1^{-1/2}$ from (5.4), it follows that $d \simeq \alpha^{-2} \lambda_1^{-1}$. For instance, assume that $k(d) = kd$. Then from (5.14), substituting K_0 with $k(d)$, we have

$$\lambda_1 > C_3 \alpha^{-\frac{4}{3}} [\alpha^{-2} \lambda_1^{-1}]^{\frac{2}{3}}, \quad C_3 > 0.$$

It follows that the periodic stability condition is possibly satisfied if the value λ_1 has the order between $\alpha^{-\frac{8}{5}}$ and α^{-2}.

Remark 5.1. As another main example for (1.1) we can consider the following a strongly damped wave equation:

$$u_{tt} - 2\alpha \Delta u_t - \Delta u = f(x, u, \text{grad} u) + w(t), \tag{5.15}$$

where Δ denotes the Laplacian on a bounded domain Ω in a finite-dimensional Euclidean space with a sufficiently smooth boundary Γ. For a homogeneous Dirichlet boundary condition we can define a operator A

$$\begin{aligned} &D(A) = H^2(\Omega) \cap H_0^1(\Omega), \\ &Au = -\Delta u, \end{aligned}$$

which has the positive discrete eigenvalues:

$$0 < \lambda_1 < \lambda_2 < \cdots < \cdots \to \infty.$$

Under suitable conditions on the nonlinear function f we can treat the case as $\beta = 1/2$, using the well-known fact

$$D(A^{\frac{1}{2}}) = H_0^1(\Omega)$$

(cf. [3]), and then we can follow the same estimate as above.

Remark 5.2. In Theorem 2.2 and Theorem 2.3, instead of hypothesis (1.5) it is sufficient to assume the uniform boundedness of solutions

$$[u(t), \dot{u}(t)] \in X_{1+\beta} \times X_\beta, \quad t > 0,$$

since this boundedness assumption yields the uniform boundedness $[\xi(t), \eta(t)]$ in $X_{1+\beta} \times X_{1+\beta}$ (cf. (2.10) and (2.12)) and, consequently, in $X_1 \times X_{1+\beta}$.

6. Appendix 1. In [5] the following Gronwall's inequality was proved: Suppose $b \geq 0, \bar{\beta} > 0$ and $a(t)$ is a nonnegative function locally integrable on $0 \leq t < +\infty$, and suppose that $y(t)$ is nonnegative and locally integrable on $0 \leq t < +\infty$ with

$$y(t) \leq a(t) + b\int_0^t (t-s)^{\bar{\beta}-1} y(s)ds$$

on this interval. Then

$$y(t) \leq a(t) + \vartheta \int_0^t E'(\vartheta(t-s))a(s)ds,$$

where

$$\vartheta = [b\Gamma(\bar{\beta})]^{1/\bar{\beta}}, \quad E(z) = \sum_{n=0}^{\infty} \frac{z^{n\bar{\beta}}}{\Gamma(n\bar{\beta}+1)}, \quad E'(z) = \frac{dE(z)}{dz}.$$

If $a(t)$ is differentiable, we note that, since

$$\frac{dE(\vartheta(t-s))}{ds} = E'(\vartheta(t-s)) \cdot (-\vartheta),$$

integration by parts gives

$$\begin{aligned} y(t) &\leq a(t) - \int_0^t \frac{dE(\vartheta(t-s))}{ds} a(s)ds \\ &= a(t) - [E(\vartheta(t-s))a(s)]_0^t + \int_0^t E(\vartheta(t-s))a'(s)ds \\ &= E(\vartheta t)a(0) + \int_0^t E(\vartheta(t-s))a'(s)ds. \end{aligned} \tag{6.1}$$

Here we consider the estimate of the entire function $E(z)$. If $0 < z < 1$, we can estimate

$$E(z) \le 1 + \frac{z^{\bar\beta}}{\Gamma_0(1 - z^{\bar\beta})}, \tag{6.2}$$

where $\Gamma_0 := \inf_{1<x<2} \Gamma(x) \approx 0.8$.

On the other hand, for a constant $\alpha : 0 < \alpha < 1$, it is known [2] that, if $z \ge \alpha$,

$$E(z) \le \frac{e^z}{\bar\beta} + \left| \frac{1}{2\pi i} \int_l \frac{u^{\bar\beta - 1} e^u}{u^{\bar\beta} - z^{\bar\beta}} du \right|,$$

where the contour $l : (-\infty - 0i, 0, -\infty + 0i)$ is the negative real axis described twice. Since elementary calculations give

$$\inf\{|u^{\bar\beta} - z^{\bar\beta}| : \; u \in (-\infty, 0), \; z > \alpha\} \ge \begin{cases} \alpha^{\bar\beta} & \text{if } \frac{1}{2} \le \bar\beta < 1, \\ (\alpha \sin \bar\beta\pi)^{\bar\beta} & \text{if } 0 < \bar\beta < \frac{1}{2}, \end{cases}$$

we have

$$E(z) \le \frac{e^z}{\bar\beta} + \frac{\Gamma(\bar\beta)}{\pi\alpha^{\bar\beta}}$$

if $1/2 \le \bar\beta < 1$ and

$$E(z) \le \frac{e^z}{\bar\beta} + \frac{\Gamma(\bar\beta)}{\pi(\alpha \sin \bar\beta\pi)^{\bar\beta}}$$

if $0 < \bar\beta < 1/2$. It follows from (6.2) that for every constant $\alpha_0 : 0 < \alpha_0 < 1$, which satisfies

$$1 + \frac{\alpha_0^{\bar\beta}}{\Gamma_0(1 - \alpha_0^{\bar\beta})} \le \frac{e^{\alpha_0}}{\bar\beta} + \frac{\Gamma(\bar\beta)}{\pi\alpha_0^{\bar\beta}} \quad \left[1 + \frac{\alpha_0^{\bar\beta}}{\Gamma_0(1 - \alpha_0^{\bar\beta})} \le \frac{e^{\alpha_0}}{\bar\beta} + \frac{\Gamma(\bar\beta)}{\pi(\alpha_0 \sin \bar\beta\pi)^{\bar\beta}} \right],$$

the following estimate holds:

$$E(z) \le \frac{e^z}{\bar\beta} + \frac{\Gamma(\bar\beta)}{\pi\alpha_0^{\bar\beta}} \quad \left[E(z) \le \frac{e^z}{\bar\beta} + \frac{\Gamma(\bar\beta)}{\pi(\alpha_0 \sin \bar\beta\pi)^{\bar\beta}} \right]$$

for every $z \ge 0$ if $\frac{1}{2} \le \bar\beta < 1$ $[0 < \bar\beta < \frac{1}{2}]$. For instance, taking $\alpha_0 = \pi^{-1/\bar\beta}$, we have

$$E(z) \le \frac{e^z}{\bar\beta} + \Gamma(\bar\beta) \quad \left[E(z) \le \frac{e^z}{\bar\beta} + \frac{\Gamma(\bar\beta)}{(\sin \bar\beta\pi)^{\bar\beta}} \right]$$

for every $z \ge 0$ if $\frac{1}{2} \le \bar\beta < 1$ $[0 < \bar\beta < \frac{1}{2}]$, since

$$1 + \frac{\alpha_0^{\bar\beta}}{\Gamma_0(1 - \alpha_0^{\bar\beta})} \approx 1.58,$$
$$\frac{e^{\alpha_0}}{\bar\beta} + \frac{\Gamma(\bar\beta)}{(\sin \bar\beta\pi)^{\bar\beta}} \ge \frac{e^{\alpha_0}}{\bar\beta} + \frac{\Gamma(\bar\beta)}{\pi\alpha_0^{\bar\beta}} > 1.8.$$

7. Appendix 2. Following a routine argument (for instance, see Theorems 3.1 and 3.3 in §6 of [9]), we can show the existence of the (classical) solution of (2.4). In fact, note that

$$D(\mathcal{A}^\beta) = L^2(\Omega) \times D(A^\beta)$$

and put

$$\begin{aligned} Y_\beta &:= D(A^\beta) \times D(A^\beta), \\ Y &:= D(A^\beta) \times L^2(\Omega). \end{aligned}$$

Then it follows from (1.4) that the nonlinear function $\mathcal{F}(t,\zeta) := \mathcal{F}(\zeta) + \mathbf{w}(t)$ is locally Hölder continuous and locally Lipschitz continuous from $R^+ \times Y_\beta$ to Y such that there exist constants $L(c), 0 < \gamma \leq 1$:

$$\|\mathcal{F}(t_1,\zeta_1) - \mathcal{F}(t_2,\zeta_2)\|_Y \leq L(c)(\|\zeta_1 - \zeta_2\|_{Y_\beta} + |t_1 - t_2|^\gamma)$$

for $(t_i,\zeta_i) \in R^+ \times Y_\beta : |t_i| + \|\zeta_i\|_{Y_\beta} \leq c,\ i = 1,2$. Since the function $w(t)$ is uniformly bounded, from (1.5) we can show that $\mathcal{F}$ satisfies the linear growth condition: there exists a constant K_0' such that

$$\|\mathcal{F}(t,\zeta)\|_Y \leq K_0'(1 + \|\zeta\|_{Y_\beta}), \quad \zeta \in Y_\beta. \tag{7.1}$$

On the other hand, denote the analytic semigroup, generated by $\mathcal{A}$, by $T(t) := [S_1(t), S_2(t)]$. Then, since the operator A^+ is bounded, we can easily show that $T(t) : Y \to Y, \quad t \geq 0$. Thus, to apply the contraction principle as in the proof of Theorems 3.1 and 3.3 in [9], it is sufficient to define the mapping $\mathcal{G}$ on the space $\mathcal{Y} := C(0,t_0 : Y)$ by

$$\mathcal{G}y(t) = T(t)\mathcal{A}^\beta y_0 + \int_0^t \mathcal{A}^\beta T(t-s)\mathcal{F}(s, \mathcal{A}^{-\beta}y(s))ds,$$

where $y_0 \in Y_\beta$ and t_0 are given by a routine estimate, which yields the contraction property of $\mathcal{G}$, to obtain a local solution. Furthermore, following the conventional argument with (7.1), we obtain the global solution $\zeta(t) = \mathcal{A}^{-\beta}y(t)$ in $C(0,T : Y_\beta) \cap C^1(0,T : L^2(\Omega) \times L^2(\Omega))$ for every $T > 0$.

Acknowledgments. The author would like to thank the referees for many suggestions that improved the presentation of this paper.

REFERENCES

[1] M. J. Balas, *Modal control of certain flexible dynamic systems*, SIAM J. Control Optim., 16 (1978), pp. 450–462.

[2] M. A. Evgrafov, *Asymptotic Estimates and Entire Functions*, Gordon and Breach, New York, 1961.

[3] D. Fujiwara, *Concrete characterization of the domains of fractional powers of some elliptic differential operators of the second type*, Proc. Japan Acad., 43 (1967), pp. 82–86.

[4] A. Haraux, *Nonlinear Evolution Equations: Global Behavior of Solutions*, Lecture Notes in Mathematics 841, Springer-Verlag, Berlin, New York, 1981.

[5] D. Henry, *Geometric Theoy of Semilinear Parabolic Equations*, Lecture Notes in Mathematics 840, Springer-Verlag, Berlin, New York, 1981.

[6] H. Ishii, *On the existence of almost periodic complete trajectories for contractive almost periodic processes*, J. Differential Equations, 43 (1982), pp. 66–72.

[7] K. Naito, *On the almost periodicity of solutions of a reaction diffusion system*, J. Differential Equations, 44 (1982), pp. 9–20.

[8] ———, *Periodically reachable sets of nonlinear parabolic systems under periodic forcing*, Yokohama Mathematical Journal, to appear.

[9] A. Pazy, *Semigroups of Linear Operators and its Applications to Partial Differential Equations*, Applied Mathematical Sciences 44, Springer-Verlag, Berlin, New York, 1983.

[10] Y. Sakawa, *Feedback control of second order evolution equations with damping*, SIAM J. Control Optim., 22 (1984), pp. 343–361.

[11] Y. Sakawa and Z. H. Luo, *Modeling and control of coupled bending and torsional vibrations of flexible beams*, IEEE Trans. Automat. Control, 34 (1989), pp. 970–977.

[12] K. Tsujioka, *On a hyperbolic equation in robotics with a singular boundary condition*, Saitama Math. J., 7 (1989), pp. 23–47.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 33, No. 6, pp. 1801–1815, November 1995

© 1995 Society for Industrial and Applied Mathematics

009

IDENTIFICATION FOR PARABOLIC DISTRIBUTED PARAMETER SYSTEMS WITH CONSTRAINTS ON THE PARAMETERS AND THE STATE*

WENHUAN YU†

Abstract. We consider the problems for identifying the parameters $a_{11}(x,t),\ldots, a_{mm}(x,t)$ and $c(x,t)$ involved in a second-order, linear, uniformly parabolic equation

$$\begin{cases} \partial_t u - \partial_i(a_{ij}(x,t)\partial_j u) + b_i(x,t)\partial_i u + c(x,t)u = f(x,t) & \text{in } \Omega \times (0,T), \\ u \mid_{\partial\Omega} = g, \quad u \mid_{t=0} = u_0(x), \quad x \in \Omega \end{cases}$$

on the basis of noisy measurement data

$$z(x) = u(x,T) + w(x), \quad x \in \Omega$$

with equality and inequality constraints on the parameters and the state variable. The cost functionals are (one-sided) Gâteaux-differentiable with respect to the state variables and the parameters. Using the Duboviskii–Miljutin lemma we get the two maximum principles for the two identification problems, respectively, i.e., the necessary conditions for the existence of optimal parameters.

Key words. system identification, constraints on the parameters and state, Gâteaux-differentiable functional, distributed parameter systems, maximum principle

AMS subject classifications. 35R30, 49K20, 93E12, 49K35

1. Introduction.

In this paper we use a general optimization theory, the Duboviskii–Miljutin theory, to study identification problems for a parabolic distributed parameter system with constraints on the parameters and the state variable.

These problems are motivated by a number of practical, industrial applications. For example, one hopes to identify the parameter $q(x)$ in the parabolic differential equation:

$$\frac{\partial p}{\partial t} = \nabla \cdot (q(x)\nabla p) + f(x,t), \tag{1}$$

based on noisy measurements of $p(x,t)$, $p^{\text{obs}}(x_i,t)$, at a set of discrete spatial locations, $i = 1,\ldots,m$.

It is well known that the equation (1) can be used to describe the motion of a fluid flow in the so-called reservoir region, where the dependent variable $p(x,t)$ represents the pressure distribution, $f(x,t)$ accounts for the withdrawal or injection of the fluid, and $q(x)$ is the transmissibility in the reservoir. In this instance one can get not only the observation of $p(x)$, $p^{\text{obs}}(x_i,t)$, but also the observations of $q(x)$, $q(x_i)$, at the measurement locations, $x_1,\ldots,x_m$. Furthermore, the pressure $p(x,t)$ is always nonnegative. So, this is a typical identification problem for a parabolic distributed parameter system with constraints on the parameter and the state variable.

There are many papers dealing with various identification problems. We only list a few references [1]–[18] for the reader's convenience. Most of the research workers only considered that the parameters submit certain limits, such as smoothness properties or upper and lower bounds. But only a few papers are concerned with equality constraints on parameters and/or inequality constraints on state variables;

* Received by the editors October 8, 1993; accepted for publication (in revised form) June 1, 1994. This research was partially supported by National Science Foundation of China grant 19271040.

† Department of Mathematics, Tianjin University, Tianjin 300072, People's Republic of China.

for example, cf. the references [19]–[22], some of which discuss optimal control problems with constraints on the controls and the state variables. In many cases one can get the direct measurement on a subset S of the parameters to be determined and the properties of the state variables. Of course, this kind of information is very helpful to estimate the parameters. Therefore, it is natural for us to present identification problems with inequality and/or equality constraints on the parameters. For these problems we could not directly use any method that needs the computation of Fréchet derivatives with respect to (w.r.t.) unknown parameter functions because the admissible parameter sets in these cases, perhaps, are those with empty interior.

Next, the outline of this paper is given as follows. In §2 we present a description of the identification problem we discuss and of the main results of the paper. That is, we consider identification problems for the distributed parameter system governed by a uniformly parabolic partial differential equation (PDE) of order two, of which we want to determine the coefficients of the lowest-order derivative and the highest-order derivatives, which satisfy certain inequality constraints and have given values at a set, by means of measurement data of the state variable on the final time.

In §3 we state the above-mentioned identification problems as optimization problems in appropriate function spaces and then state the Duboviskii–Miljutin lemma for general optimization problems. In §4 we, first, calculate a descent direction cone, cones of admissible directions, a tangent direction cone, and their dual cones according to the requirement of the Duboviskii–Miljutin lemma. Next, from the dual cones we get functionals that satisfy the Euler–Lagrange equation. Finally, from the Euler–Lagrange equation we derive the useful necessary conditions that are satisfied by the optimal parameters minimizing the cost functionals, which may be only one-sided Gâteaux differentiable.

By the way, it should be pointed out that the study of uniqueness of the solution to an inverse problem remains a great challenge for applied mathematics, so there are only a few papers concerning this topic; see, e.g., [9]–[14], [16], and [18]. But it is a most important topic for practical researchers. We do not discuss this problem in the paper.

Finally, we also mention the relationship between identification of distributed parameter systems governed by PDEs or inverse problems in PDEs and the optimal control theory or optimization theory. Usually, one recognizes an identification problem as a special optimal control problem and the parameter as a special control. Therefore, the optimization theory offers a research approach for the study of identification. On the other hand, because of the specific characteristics of identification problems, they provide new research topics for optimization theory.

2. Description of the problem.

Consider the problem of identifying the unknown coefficient functions $a_{11}(x,t), \ldots, a_{mm}(x,t)$ and $c(x,t)$, which are regarded as a parameter vector $q \equiv (a_{11}, \ldots, a_{mm}, c)$, involved in the following second-order, linear, uniformly parabolic system:

$$\begin{aligned} &Lu \equiv \partial_t u - \partial_i(a_{ij}(x,t)\partial_j u) + b_i(x,t)\partial_i u + c(x,t)u = f(x,t), \quad (x,t) \in D, \\ &u\mid_{\partial\Omega} = g(s,t), \quad (s,t) \in \partial\Omega \times [0,T], \quad u\mid_{t=0} = u^0(x), x \in \bar{\Omega}, \end{aligned} \tag{2}$$

on the basis of the measurement

$$u\mid_{t=T} = z(x) + w(x), \quad x \in \Omega, \tag{3}$$

where $D \equiv \Omega \times (0, T)$, z is a given function defined on Ω, w is a white measurement noise with zero mean, and

$$\begin{array}{ll} x \equiv (x_1, \ldots, x_m), & a_i b_i \equiv \sum_i a_i b_i, \\ \partial_t \equiv \partial / \partial t, & \partial_i \equiv \partial / \partial x_i. \end{array}$$

In addition, some value of q is given:

$$q \mid_S = q_o, \tag{4}$$

where S is a subset of $\bar{\Omega}$ and q_o is an $(m^2 + 1)$-valued given function defined on S.

Furthermore, the solution of (2) is subject to a constraint:

$$\beta(x, t, u(x, t)) \leq 0, \qquad \forall (x, t) \in D. \tag{5}$$

Remark. If we take $\beta(x, t, u) = -u$, then (5) means that the solution of (2) is nonnegative on D.

In order to let the problem (2) be well posed we always suppose in the paper that the following assumptions are true.

- H1. $\Omega \subset \mathbb{R}^m$ is a bounded open set and its boundary $\partial\Omega \in C^{2+\alpha}$.
- H2. $b_i \in \bar{C}^{1+\alpha}(D)$, $f \in \bar{C}^{\alpha}(D)$, $u^0 \in C^{2+\alpha}$, and $g \in \bar{C}^{2+\alpha}$ are given.
- H3. u^0 and g satisfy the compatibility conditions, i.e., $\zeta \in C^{2+\alpha}$, where

$$\zeta(x, t) = \begin{cases} g(x, t), & (x, t) \in \partial\Omega \times [0, T], \\ u^0(x), & (x, t) \in \bar{\Omega} \times \{0\}. \end{cases}$$

Then by [24] for any q in

$$Q_{\mathrm{ad},1} \equiv \{q \in Q; \quad \nu|\xi|^2 \leq a_{ij}(x, t)\xi_i\xi_j \leq \mu|\xi|^2, \ \forall \xi \in \mathbb{R}^m, \ \forall (x, t) \in D\}, \tag{6}$$

there exists a unique solution $u \in V \equiv \bar{C}^{2+\alpha}(D)$ to the problem (2), which is denoted by $u = u(q) = u(x, t; q)$ to show the dependence of u on q, where $Q \equiv [\bar{C}^{1+\alpha}(D)]^{m^2} \times \bar{C}^{\alpha}(D)$, and the definitions of $\bar{C}^{\alpha}(D)$, $\bar{C}^{1+\alpha}(D)$, and $\bar{C}^{2+\alpha}(D)$ can be found in [24].

One can recognize the above-mentioned parameter identification problem as a minimum problem exactly as most researchers have done. That is, seek a minimal parameter $\hat{q} \in Q_{\mathrm{ad}}$ such that the cost functional

$$F_o(q) \equiv f_o(u(q), q; z) \tag{7}$$

reaches its minimum over the admissible parameter set Q_{ad} at $\hat{q}$, i.e.,

$$F_o(\hat{q}) = \min_{q \in Q_{\mathrm{ad}}} F_o(q), \tag{8}$$

where the admissible set Q_{ad} is defined by

$$Q_{\mathrm{ad}} \equiv Q_{\mathrm{ad},1} \cap Q_{\mathrm{ad},2}, \tag{9}$$

$$Q_{\mathrm{ad},2} \equiv \{q \in Q; \ q \mid_S = q_o\}, \tag{10}$$

and f_o is a (one-sided) Gâteaux-differentiable functional.

Under the same constraint conditions (2), (6), and (10), depending upon the choice of f_o, one can consider the following extremum problems.

PROBLEM A. The cost functional is

$$F_a(q) \equiv \int_\Omega \phi(x, u(x,T;q), q(x,T); z(x))\, dx. \tag{11}$$

PROBLEM B. The cost functional is

$$F_b(q) \equiv \max_\Omega |u(x,T;q) - z(x)|. \tag{12}$$

Adding to the assumptions H1–H3, we also suppose that the function ϕ satisfies the following assumption, H4.

$$H4 \left\{ \begin{array}{l} \phi \text{ is continuous w.r.t. all their arguments. Moreover, } \phi(x,\cdot,\cdot;z) \text{ satisfies} \\ \text{the Lipschitz condition} \\ \qquad |\phi(x,u_1,q_1;z) - \phi(x,u_2,q_2;z)| \le L\{|u_1 - u_2| + |q_1 - q_2|\} \\ \text{for any } \ x \in \Omega \text{ and for any bounded } u \text{ and } q, \text{ and is differentiable w.r.t.} \\ u \text{ and } q \text{ in the directions } \dot{u} \text{ and } \dot{q}. \text{ That is, there exist the following limits :} \end{array} \right.$$

$$\begin{array}{l} \lim_{t\to+0}[\phi(x,u+t\dot{u},q;z) - \phi(x,u,q;z)]/t \equiv \phi_u(x,u,q;z)\dot{u}, \\ \lim_{t\to+0}[\phi(x,u,q+t\dot{q};z) - \phi(x,u,q;z)]/t \equiv \langle\phi_q(x,u,q;z),\dot{q}\rangle. \end{array} \tag{13}$$

Now, we state the main results as follows.

THEOREM 2.1. *Suppose that the hypotheses* H1–H4 *are valid and that there exists a solution* $q^o \in Q_{\text{ad}} \equiv Q_{\text{ad},1} \cap Q_{\text{ad},2}$ *to Problem A, where* $q^o \equiv (a^o_{11}, \ldots, a^o_{mm}, c^o)$, *and* $Q_{\text{ad},1}$ *and* $Q_{\text{ad},2}$ *are defined by* (6) *and* (10), *respectively, such that* (u^o, q^o) *with* $u^o = u(q^o)$ *minimizes the functional* (11).

Then there is a nonnegative measure $\nu \in rca(D)$ *with compact support contained in the set*

$$\mathfrak{N} \equiv \{(x,t) \in D; \quad \beta(x,t,u^o(x,t)) = 0\}, \tag{14}$$

such that the following maximum principle,

$$\begin{array}{l} \sup_{q\in Q_{\text{ad}}}\{\int_D[a_{ij}\partial_j u^o\partial_i v + cu^o v] - \int_\Omega\langle\phi_q(x,u^o(x,T),q^o(x,T);z(x)),q\rangle\} \\ \quad = \int_\Omega[a^o_{ij}\partial_j u^o\partial_i v + c^o u^o v] - \int_\Omega\langle\phi_q(x,u^o(x,T),q^o(x,T);z(x)),q^o(x,T)\rangle, \end{array} \tag{15}$$

holds, where v is the generalized solution in Ladyzhenskaya's sense to the adjoint equation of (2), *i.e., v is the solution of the following variation equation:*

$$\begin{array}{l} \int_D[(\partial_t w + b_i\partial_i w + c^o w)v + a^o_{ji}\partial_i w\partial_j v] \\ \quad = \int_\Omega \phi_u(x,u^o(x,T),q^o(x,T);z(x))w(x,T)dx \\ \quad + \int_D \chi_{\mathfrak{N}}\beta_u(x,t,u^o;z)w\,\nu(dxdt), \qquad \forall w \in \mathfrak{H}, \end{array} \tag{16}$$

where $\beta_u \equiv \partial\beta/\partial u$, $\chi_{\mathfrak{N}} = 0$ *if* $\mathfrak{N} = \emptyset$; *otherwise* $\chi_{\mathfrak{N}}$ *is the characteristic function of* $\mathfrak{N}$,

$$\mathfrak{H} \equiv \{w \in H^1(0,T;H^1_0(\Omega)); \quad w\,|_{t=0} = 0\},$$

and $rca(\mathfrak{S})$ *is the space of all regular countably additive scalar-valued functions on the* σ*-field of all Borel sets in* $\mathfrak{S}$ *and the norm in* $rca(\mathfrak{S})$, $|\mu|$, *is the total variation* $v(\mu,\mathfrak{S})$.

Similarly, for Problem B, we have the following results.

THEOREM 2.2. *Suppose that the hypotheses* H1–H3 *are valid and that there exists an optimal solution* $q^o \in Q_{\text{ad}}$ *to Problem B, such that* $u^o = u(q^o)$ *minimizes the functional* (12).

Then there is a nonnegative measure $\mu \in rca(\Omega)$, *of which support is contained in the set*

$$\mathfrak{C} \equiv \left\{ x \in \Omega; \quad |u^o(x,T) - z(x)| = \max_{y \in \bar{\Omega}} |u^o(y,T) - z(y)| \right\}, \tag{17}$$

and a nonnegative measure $\nu \in rca(D)$ *with compact support contained in the set* $\mathfrak{N}$ *such that the following maximum principle,*

$$\sup_{q \in Q_{ad}} \int_D [a_{ij}\partial_j u^o \partial_i v + cu^o v]\, dxdt = \int_D [a^o_{ij}\partial_j u^o \partial_i v + c^o u^o v]\, dxdt, \tag{18}$$

holds, where v *is defined by the following variation equation:*

$$\begin{aligned} &\textstyle\int_D [(\partial_t w + b_i \partial_i w + c^o w)v + a^o_{ij}\partial_j w \partial_i v]\, dxdt \\ &= \textstyle\int_\Omega \chi_{\mathfrak{C}} \operatorname{sign}[u^o(s,T) - z(s)] w(s,T)\, \mu(ds) + \int_D \chi_{\mathfrak{N}} \beta_u(x,t,u^o;z) w\, \nu(dxdt), \\ &\qquad \forall w \in \mathfrak{H}, \end{aligned} \tag{19}$$

and the definitions of Q_{ad}, $rca(D)$, $rca(\Omega)$, $\mathfrak{N}$, $\chi_{\mathfrak{N}}$, β_u, *and* $\mathfrak{H}$ *are the same as those in Theorem* 2.1, *and* $\chi_{\mathfrak{C}} = 0$ *if* $u^o(\cdot,T) = z(\cdot)$, *otherwise* $\chi_{\mathfrak{C}}$ *is the characteristic function of* $\mathfrak{C}$.

3. Preliminary.

In order to solve the minimum problems we use the Duboviskii–Miljutin lemma [25]. Therefore, we consider a general extremum problem in a locally convex topological space E, i.e., minimize a cost functional

$$F_o(Y) \tag{20}$$

with inequality constraints

$$Y \in E_j \quad (j = 1, \ldots, n) \tag{21}$$

and an equality constraint

$$Y \in E_{n+1}, \tag{22}$$

where E_j $(j = 1, \ldots, n)$ are sets with a nonempty interior in E.

From [25] or [28] we get the following lemma.

LEMMA 3.1 (Duboviskii–Miljutin). *Assume that the following hypotheses hold.*

- $F_o(Y)$ *is a regular decrease functional at* $Y_o \in \cap_{j=1}^{n+1} E_j$, *and its decrease direction cone at* Y_o *is* K_0.
- *The inequality constraints* E_j $(j = 1, \ldots, n)$ *are regular at* Y_o *and their admissible direction cones at* Y_o *are* K_j.
- *The equality constraint* E_{n+1} *is regular at* Y_o *and its tangential direction cone at* Y_o *is* K_{n+1}.
- *The functional* $F_o(Y)$ *reaches its minimum at* Y_o *on* $\cap_{j=1}^{n+1} E_j$.

Then $\psi_i \in K_i^+$, K_i^+ *are the dual cones of* K_i, $i = 0, \ldots, n+1$, *such that* $\psi_0, \ldots, \psi_{n+1}$, *which are not all simultaneously equal to zero, satisfy the Euler–Lagrange equation*

$$\sum_{i=0}^{n+1} \psi_i = 0. \tag{23}$$

From now on we suppose

$$Y \equiv (u, q), \quad E \equiv V \times Q, \tag{24}$$

where $V = \bar{C}^{2+\alpha}(D)$ and $Q = [\bar{C}^{1+\alpha}(D)]^{m^2} \times \bar{C}^{\alpha}(D)$. Moreover, it is obvious that

$$E^* = \{f = (f_1, f_2); \quad f_1 \in V^*, \quad f_2 \in Q^*\},$$

i.e., $E^* = V^* \otimes Q^*$, where E^*, V^*, and Q^* are the adjoint spaces of E, V, and Q, respectively.

Moreover, suppose

$$E_1 \equiv \{Y = (u, q); \quad u \in V, \ q \in Q_{\mathrm{ad},1}\} = V \times Q_{\mathrm{ad},1}, \tag{25}$$

$$E_2 \equiv \{Y = (u, q) \in E; \quad P(Y) = 0\}, \tag{26}$$

$$E_3 \equiv \{Y \in E; \quad \beta(x, t, u(x, t)) \leq 0, \quad \forall\, (x, t) \in D\}, \tag{27}$$

where $Q_{\mathrm{ad},1}$ is defined by (6) and $P : E \longrightarrow F$ is defined by

$$P(Y) \equiv \{L(q)u - f, L_b(q)u - g, L_s q - q_o\}, \tag{28}$$

$$L(q)u \equiv \partial_t u - \partial_i(a_{ij}\partial_j u) + b_i \partial_i u + cu, \tag{29}$$

$$L_b(q)u \equiv u \mid_{\partial\Omega}, \tag{30}$$

$$L_s q \equiv q \mid_S, \tag{31}$$

where

$$F \equiv \bar{C}^{\alpha}(D) \times \bar{C}^{2+\alpha}(\partial\Omega \times [0, T]) \times \Psi, \tag{32}$$

$$\Psi \equiv \{p(x, \cdot) \in (\bar{C}^{1+\alpha}([0, T]))^{m^2} \times \bar{C}^{\alpha}([0, T]); \quad x \in S, \quad p \text{ is continuous on } S \times [0, T]\};$$

moreover, if we set $\|w\|_F \equiv \|w_1\|_\alpha + \|w_2\|_{2+\alpha} + \|w_3\|_\Psi$ and $\|w_3\|_\Psi \equiv \max_{x \in S} \|w_3(x, \cdot)\|$, then F is a Banach space.

Furthermore, set

$$Z = E_1 \cap E_2 \cap E_3, \tag{33}$$

$$J_a(Y) \equiv \int_\Omega \phi(x, u(x; T), q(x); z(x))\, dx, \tag{34}$$

$$J_b(Y) \equiv \max_{\bar{\Omega}} |u(x, T) - z(x)|. \tag{35}$$

Therefore, the parameter estimation problems A and B will be deduced to special extremum problems.

PROBLEM A′. Minimize $J_a(Y)$ on the same constraints (33) with (25)–(27).

PROBLEM B′. Minimize $J_b(Y)$ on the same constraints (33) with (25)–(27).

4. Proof of the main results.

In order to use the Duboviskii–Miljutin lemma to prove Theorems 2.1 and 2.2 we need several lemmas.

From [25] one can get the following.

LEMMA 4.1. *Let $f \in E^*$ and*

$$I_1 \equiv \{x \in E; \quad f(x) = 0\}, \qquad I_2 \equiv \{x \in E; \quad f(x) \geq 0\},$$
$$I_3 \equiv \{x \in E; \quad f(x) > 0\}.$$

Then

$$I_1^+ = \{\lambda f;\ \lambda \in \mathbf{R}^1\}, \quad I_2^+ = \{\lambda f;\ \lambda \in \mathbf{R}_+^1\},$$

and $I_3^+ = E^*$ *if* $f = 0$; *otherwise* $I_3^+ = I_2^+$.

From [25] one also gets the following.

LEMMA 4.2. *Suppose that* $F(Y)$ *satisfies the Lipschitz condition in a neighborhood of* $Y_0 \in E$ *and is (one-sided) Gâteaux differentiable at* Y_0 *in any direction* $\dot{Y}$, *and that the Gâteaux variation,* $F'(Y_0, \dot{Y})$, *of* $F(Y)$ *at* Y_0 *is convex w.r.t.* $\dot{Y}$. *Then* $F(Y)$ *regularly descends at* Y_0 *and the descent direction cone,* K, *of* F *at* Y_0 *is*

$$K = \{\dot{Y};\quad F'(Y_0, \dot{Y}) < 0\}.$$

LEMMA 4.3. *The functional* $J_a(Y)$ *is one-sided Gâteaux differentiable, and its Gâteaux differential is determined by*

$$\begin{aligned} J_a'(Y)\dot{Y} = {} & \textstyle\int_\Omega [\phi_u(s, u(s,T), q(s,T); z(s))\dot{u}(s,T) \\ & + \langle \phi_q(s, u(s,T), q(s,T); z(s)), \dot{q}(s,T)\rangle]\, ds, \qquad \forall \dot{Y} = (\dot{u}, \dot{q}) \in E, \end{aligned} \tag{36}$$

where ϕ_u *and* ϕ_q *are defined by* (13). *Moreover,* $J_a(Y)$ *is a regular descent functional at* $Y = (u, q)$, *and its descent direction cone,* K_0, *at* Y *is*

$$K_0 = \{\dot{Y} \in E;\ \ J_a'(Y)\dot{Y} < 0\}, \tag{37}$$

and the dual cone of K_0 *is*

$$K_0^+ = \{-\lambda_0 J_a'(Y);\quad 0 < \lambda_0 < +\infty\}. \tag{38}$$

Proof. By the assumption H4 one immediately gets

$$\begin{aligned} \lim\nolimits_{t\to+0} [J_a(Y + t\dot{Y}) - J_a(Y)]/t = {} & \lim\nolimits_{t\to+0} \textstyle\int_\Omega \{[\phi(s, u + t\dot{u}, q + t\dot{q}; z) \\ & -\phi(s, u, q + t\dot{q}; z)]/t + [\phi(s, u, q + t\dot{q}; z) - \phi(s, u, q; z)]/t\}\, ds \\ = {} & \textstyle\int_\Omega [\phi_u(s, u, q; z)\dot{u} + \langle \phi_q(s, u, q; z), \dot{q}\rangle]\, ds, \end{aligned}$$

and then (36) is true.

It is obvious that $J_a'(Y)\dot{Y} < 0$ if $\dot{Y} \in K_0$.

On the other hand, if $\dot{Y}$ satisfies $J_a'(Y)\dot{Y} = -\delta < 0$, then there exists $\epsilon_1 > 0$ such that $J_a(Y + \epsilon\dot{Y}) \leq J_a(Y) - \epsilon\delta/2$, $\forall \epsilon \in (0, \epsilon_1)$, by the definition of Gâteaux differential. Let $h \in B_1(\dot{Y}) \equiv \{h \in E;\ \|h - \dot{Y}\| < \delta/(4L)\}$. Then, by the assumption H4,

$$\begin{aligned} J_a(Y + \epsilon h) = {} & J_a(Y + \epsilon\dot{Y}) + [J_a(Y + \epsilon h) - J_a(Y + \epsilon\dot{Y})] \leq J_a(Y + \epsilon\dot{Y}) \\ & +\epsilon\delta/4 \leq J_a(Y) - \epsilon\delta/2 + \epsilon\delta/4 = J_a(Y) - \epsilon\delta/4. \end{aligned}$$

So, if we take $\alpha = -\delta/4$, then $\dot{Y} \in K_0$. Therefore, (37) is true. In addition, it follows by (36) that the set K_0 is convex. Hence, $J_a(Y)$ is a regular descent functional at Y.

Finally, (38) can be deduced by Lemma 4.1. □

Furthermore, we also have the following lemma.

LEMMA 4.4. *The functional* $J_b(Y)$ *is one-sided Gâteaux differentiable, and its Gâteaux variation is determined by*

$$J_b'(Y, \dot{Y}) = \max_{s \in \mathfrak{C}} \{\dot{u}(s,T)\, \mathrm{sign}[u(s,T) - z(s)]\}, \quad \forall \dot{Y} = (\dot{u}, \dot{q}) \in E, \quad u\,|_{t=T} \neq z, \tag{39}$$

$$\textit{or} \quad J_b'(\tilde{Y}, \dot{Y}) = \max_{s \in \Omega} |\dot{u}(s,T)|, \quad \forall q \in Q, \tilde{Y} = (\tilde{u}, q) \in E \ \ \textit{with}\ \ \tilde{u}\,|_{t=T} = z. \tag{40}$$

Moreover, $J_b(Y)$ is a regular descent functional at $Y = (u,q)$; its descent direction cone at Y, K_0, and the dual cone of K_0, K_0^+, are

(i) *if $Y = (u,q)$ with $u \,|_{t=T} \neq z$,*

$$K_0 = \{\dot{Y} \in E;\quad \dot{u}(x,T)\operatorname{sign}[u(x,T) - z(x)] < 0,\ \ x \in \mathfrak{C}\}, \tag{41}$$

$$K_0^+ = \left\{\mu \in \Phi;\quad -\int_{\mathfrak{C}} \dot{u}(x,T)\operatorname{sign}[u(x,T) - z(x)]\,\mu(dx) > 0,\ \ \forall \dot{Y} \in K_0\right\}, \tag{42}$$

(ii) *or if $\tilde{Y} = (\tilde{u},q)$ with $\tilde{u} \,|_{t=T} = z$,*

$$K_0 = \emptyset, \qquad\qquad K_0^+ = 0, \tag{43}$$

where

$$\Phi \equiv \{\mu \in rca(\mathfrak{C});\quad \operatorname{supp}(\mu) \subseteq \mathfrak{C},\ \ \mu(dx) \geq 0\}. \tag{44}$$

Proof. For all $u \,|_{t=T} \neq z$ we have

$$\begin{aligned} J_b'(Y,\dot{Y}) &= \lim\nolimits_{t\to+0}\{\max|u(x,T) + t\dot{u}(x,T) - z(x)| - \max|u(x,T) - z(x)|\}/t \\ &= \lim\{\max|\,|u(x_o,T) - z(x_o)| \\ &\qquad + t\dot{u}(x,T)\operatorname{sign}[u(x,T) - z(x)]| - \max|u(x,T) - z(x)|\}/t. \end{aligned}$$

Obviously, for t small enough we have

$$\begin{aligned} &\max\nolimits_{x\in\bar{\Omega}}|\,|u(x,T) - z(x)| + t\dot{u}(x,T)\operatorname{sign}[u(x,T) - z(x)]| \\ &\quad = \max\nolimits_{x\in\mathfrak{C}}|u(x,T) - z(x)| + t\max\nolimits_{x\in\mathfrak{C}}\dot{u}(x,T)\operatorname{sign}[u(x,T) - z(x)] + o(t). \end{aligned}$$

Thus,

$$J_b'(Y,\dot{Y}) = \max_{x\in\mathfrak{C}}\{\dot{u}(x,T)\operatorname{sign}[u(x,T) - z(x)]\}.$$

Obviously, (40) and (43) are true.

Next, one easily obtains that the Gâteaux variation, $J_b'(Y,\dot{Y})$, defined by (39) or (40) is convex and that the functional $J_b(Y)$ satisfies the Lipschitz condition and is one-sided differentiable in any direction $\dot{Y}$. So, by Lemma 4.2 $J_b(Y)$ is a regular descent functional. Moreover, (41) is true.

From the definition of a dual cone and the Riesz representation theorem, (e.g., cf. [23, p. 265]), it immediately follows that the equality (42) is also true. □

Regarding the set E_1 we have the following lemma.

LEMMA 4.5. *Suppose $Y_o = (u^o, q^o) \in E_1 = V \times Q_{\text{ad},1}$, where $Q_{\text{ad},1}$ is defined by (6). Then E_1 is regular at Y_o and the dual cone of E_1 at Y_o is*

$$K_1^+ = \{(0,f_2) \in E^*;\ \ f_2(q - q^o) \geq 0,\ \forall q \in Q^o_{\text{ad},1}\}, \tag{45}$$

where $Q^o_{\text{ad},1}$ is the interior of $Q_{\text{ad},1}$, i.e., f_2 is a support functional of $Q_{\text{ad},1}$ at q^o.

Proof. It is obvious that the cone of admissible directions of E_1 at $Y_o = (u^o, q^o)$ is

$$K_1 = V \times \{k \in Q;\ \ k = \lambda(q - q^o),\ q \in Q^o_{\text{ad},1},\ \lambda \geq 0\}. \tag{46}$$

Thus, K_1 is convex and E_1 is regular at Y_o.

Next, by $E^* = V^* \otimes Q^*$ and (46) we have

$$K_1^+ = \{(f_1, f_2);\ f_1(u) \geq 0,\ \forall u \in V,\ f_2(k) \geq 0,\ k = \lambda(q - q^o),\ \lambda \geq 0\}. \tag{47}$$

Considering that V is the whole space, we have $f_1 = 0$. Owing to $\lambda \geq 0$, we get $f_2(q - q^o) \geq 0$, $\forall q \in Q^o_{\text{ad},1}$. □

From [27] we can get the following.

LEMMA 4.6 (Lusternik). *Let $W(Y)$ be an operator mapping E into F, differentiable in a neighborhood of a point Y_o with $W(Y_o) = 0$. Let $W'(Y)$ be continuous in a neighborhood of Y_o, and suppose $W'(Y_o)$ maps E onto F (i.e., the linear equation $W'(Y_o)\dot{Y} = h$ has a solution $\dot{Y} \in E$ for any $h \in F$).*

Then the set of tangential directions to the set

$$\tilde{E}_2 \equiv \{Y \in E;\quad W(Y) = 0\}$$

at the point Y_o is the subspace

$$\tilde{K}_2 \equiv \{\dot{Y} \in E;\quad W'(Y_o)\dot{Y} = 0\}.$$

LEMMA 4.7. *The operator $P(Y)$, which is defined by (28), is Fréchet differentiable with respect to $Y = (u, q)$ and $\forall \dot{Y} = (\dot{u}, \dot{q}) \in E$ the Fréchet differential of $P(Y)$ at $Y \in E_1$ is determined by*

$$P'(Y)\dot{Y} = \{L(q)\dot{u} - \partial_i(\dot{a}_{ij}\partial_j u) + \dot{c}u, L_b(q)\dot{u}, L_s\dot{q}\}, \tag{48}$$

where $\dot{q} \equiv (\dot{a}_{11}, \ldots, \dot{a}_{mm}, \dot{c})$, i.e.,

$$P'(Y) \in \mathcal{L}(E; F), \tag{49}$$

and F is defined by (32).

In addition, the tangential direction cone K_2 to E_2, where E_2 is defined by (26), at $Y \in E_1$, is

$$K_2 = \{\dot{Y} \in E;\ P'(Y)\dot{Y} = 0\}; \tag{50}$$

i.e., $\forall \dot{q} \in Q$ satisfying the condition

$$L_s\dot{q} = 0, \tag{51}$$

$\dot{u}$ is the solution to the problem

$$\left\{ \begin{array}{ll} \partial_t\dot{u} - \partial_i(a_{ij}\partial_j\dot{u}) + b_i\partial_i\dot{u} + c\dot{u} = \partial_i(\dot{a}_{ij}\partial_j u) - \dot{c}u, & (x, t) \in D, \\ \dot{u}\,|_{\partial\Omega} = 0, \qquad \dot{u}\,|_{t=0} = 0, & \end{array} \right. \tag{52}$$

where $(u, q) \equiv (u, a_{11}, \ldots, a_{mm}, c) \in E_1$ also locates on the manifold E_2, or $u = u(q)$ is the solution to the problem (2) with the constraints on q.

Proof. The results (48) and (49) are evident. So, we only need to prove (50) using Lemma 4.6.

By [26] we know that the mapping is surjective, in fact, $\forall (f, g, q_o) \in F$, the operator equation

$$P'(Y_o)\dot{Y} = \{f, g, q_o\} \tag{53}$$

has a unique solution $\dot{Y} = (\dot{u}, \dot{q}) \in E$. And then it follows from Lemma 4.6 that (50), and therefore (51) and (52), is true. □

LEMMA 4.8. *The dual cone of* K_2, *which is defined by* (50), *is determined by*

$$\begin{aligned} K_2^+ \equiv &\{f = (f_1, f_2) \in E^*;\ f_1 \in V^*,\ f_2 = -[u'(q)]^* f_1\} \\ &+ \cup_{x \in S} \{(0, \lambda_x) L_x;\ \lambda_x \in \mathbb{R}^1\}, \\ L_x q \equiv &\, q(x), \quad \forall x \in S, \end{aligned} \tag{54}$$

where $[u'(q)]^*$ *is the adjoint operator of* $u'(q)$ *and* $u'(q)\dot{q} = \dot{u}$ *is determined by* (52).

Proof. We first denote the cone of tangent directions K_2 as

$$K_2 = L_1 \cap L_2, \tag{55}$$

where

$$L_1 \equiv \{(\dot{u}, \dot{q}) \in E;\ L(q)\dot{u} - \partial_i(\dot{a}_{ij}\partial_j u) + \dot{c}u = 0,\ L_b(q)\dot{u} = 0\} \tag{56}$$

and

$$L_2 \equiv \{(\dot{u}, \dot{q}) \in E;\ L_s \dot{q} = 0\}. \tag{57}$$

Obviously, we have

$$K_2^+ = L_1^* + L_2^*, \tag{58}$$

where L_1^* and L_2^* are dual spaces of L_1 and L_2, respectively, which are subspaces of E^*.

For any $(\dot{u}, \dot{q}) \in L_1$ we have

$$\left\{ \begin{array}{ll} L(q)\dot{u} = \partial_i(\dot{a}_{ij}\partial_j u) - \dot{c}u, & (x,t) \in D, \\ \dot{u}\,|_{\partial\Omega} = 0, & \dot{u}\,|_{t=0} = 0. \end{array} \right. \tag{59}$$

It is easy to see $\dot{u} = u'(q)\dot{q}$, where $\dot{q} = (\dot{a}_{11}, \ldots, \dot{a}_{mm}, \dot{c})$, i.e., $\dot{u}$ is the Fréchet differential of u at q. Therefore,

$$L_1 = \{(\dot{u}, \dot{q}) \in E;\ \dot{u} = u'(q)\dot{q}\}.$$

Moreover, L_1 is a subspace of E, so

$$L_1^* = \{f = (f_1, f_2) \in E^*;\ 0 = f_1(\dot{u}) + f_2(\dot{q}),\ \forall(\dot{u}, \dot{q}) \in L_1\}.$$

Considering $\dot{u} = u'(q)\dot{q}$ on L_1 we have

$$\begin{aligned} 0 &= f_1(\dot{u}) + f_2(\dot{q}) = \langle f_1, u'(q)\dot{q}\rangle + \langle f_2, \dot{q}\rangle \\ &= \langle [u'(q)]^* f_1 + f_2, \dot{q}\rangle, \quad \forall \dot{q} \in Q. \end{aligned}$$

Thus, $[u'(q)]^* f_1 + f_2 = 0$, i.e.,

$$f_2 = -[u'(q)]^* f_1.$$

Next, $L_2 = \cap_{x \in S} A_x$, where $A_x \equiv \{(\dot{u}, \dot{q}) \in E;\ L_x \dot{q} = 0\}$.

Obviously, $L_2^* = \cup_{x \in S} A_x^*$ and by Lemma 4.1 we have $f_1 = 0$ and $f_2 = \lambda_x L_x$, hence

$$A_x^* = \{(0, \lambda_x) L_x;\quad \lambda_x \in \mathbb{R}^1\}.$$

So, we get (54). □

LEMMA 4.9. *The inequality constraint, E_3, which is defined by (27), is regular at $Y_o = (u^o, q^o) \in E_3$, and its admissible direction cone, K_3, at Y_o is determined by*

$$K_3 = E, \qquad \textit{if} \quad \mathfrak{N} = \emptyset; \tag{60}$$

otherwise

$$K_3 = \{\dot{Y} = (\dot{u}, \dot{q}) \in E; \quad \beta_u(x,t,u(x,t)) < 0, \ \forall\,(x,t) \in \mathfrak{N}\}. \tag{61}$$

Furthermore, the dual cone of K_3 is

$$K_3^+ = \{0\}, \qquad \textit{if}\, \mathfrak{N} = \emptyset, \tag{62}$$

or, $\forall\, w \in K_3^+$ there is $\tilde{\nu} \in rca(\bar{D})$ with supp$\tilde{\nu}$ $\subseteq \mathfrak{N}$ *and $\tilde{\nu}(dxdt) \geq 0$ such that*

$$\begin{aligned} & w = (\tilde{w}, 0) \in E^*, \\ & w(Y) = \tilde{w}(\dot{u}) = -\int_{\mathfrak{N}} \beta_u(x,t,u^o(x,t))\dot{u}(x,t)\tilde{\nu}(dxdt), \quad \forall\, \dot{Y} \in E. \end{aligned} \tag{63}$$

Proof. Obviously, (60) and (62) are valid.

Next, consider the following nonlinear functional

$$G(u) = \max_{(x,t)\in\bar{D}} \beta(x,t,u(x,t)). \tag{64}$$

Then $G(u^o) = 0$ if $\mathfrak{N} \neq \emptyset$. Moreover, following [25], we can prove that the mapping $G : \ C(\bar{D}) \to \mathbb{R}^1$ is Gâteaux differentiable and its Gâteaux differential at u^o in direction h is

$$G'(u^o)h = \max_{(x,t)\in\mathfrak{N}} [\beta_u(x,t,u^o(x,t))h(x,t)]. \tag{65}$$

By the definition of admissible direction cone, $\dot{Y} = (\dot{u}, \dot{q}) \in K_3$ means that there exist an $\epsilon_0 > 0$ and a neighborhood of zero, $U(0)$, such that

$$Y_o + \epsilon(\dot{Y} + Y) \in E_3, \quad \forall\,\epsilon \in [0, \epsilon_0], \quad \forall\, Y = (u,q) \in U(0).$$

So, by the definition of E_3

$$G(u^o + \epsilon(\dot{u} + u)) \leq 0,$$

and owing to $G(u^o) = 0$,

$$\epsilon G'(u^o)(\dot{u} + u) + o(\epsilon) \leq 0.$$

Consider $(u,q) \in U(0)$; therefore,

$$G'(u^o)\dot{u} < 0. \tag{66}$$

Hence, $\forall\, \dot{Y} \in K_3$, we have

$$\beta_u(x,t,u^o(x,t))\dot{u}(x,t) < 0, \qquad \forall\,(x,t) \in \mathfrak{N}. \tag{67}$$

On the contrary, if $\dot{Y} = (\dot{u}, \dot{q})$ satisfies (67), then (66) is true, and it is easy to prove $\dot{Y} \in K_3$.

Therefore, (61) is valid.

Because the space V is a dense subset of $C(\bar{D})$, $\forall w = (\tilde{w}, 0) \in K_3^+$ there is $\tilde{\nu} \in rca(\bar{D})$ by the Riesz representation theorem [23] and the Minkowskii–Farkas theorem [25] such that (62) is valid. □

Now, we return to the proof of Theorem 2.1.

Proof of Theorem 2.1. If (u^o, q^o) $[u^o = u(q^o)]$ is the solution to Problem A′ (i.e., Problem A), then the cost functional $J_a(Y)$ defined by (11) attains the minimum at $Y_o = (u^o, q^o)$.

By (Duboviskii–Miljutin) Lemma 3.1 there exist $\psi_i \in K_i^+$ $(i = 0, 1, \dots, 3)$, which are not all simultaneously equal to zero and satisfy the Euler–Lagrange equation

$$\sum_{i=0}^{3} \psi_i(\dot{Y}) = 0, \quad \forall \dot{Y} \in \bigcap_{i=0}^{3} K_i \equiv \tilde{Z}. \tag{68}$$

By Lemma 4.3 there is a $\lambda_0 \geq 0$ such that

$$\psi_0(\dot{Y}) = -\lambda_0 J_a'(Y_o)(\dot{Y}), \tag{69}$$

and by Lemma 4.5 we have

$$\begin{aligned} &\psi_1(\dot{Y}) = (0, f_2)(\dot{Y}) = f_2(\dot{q}), \\ &f_2 \text{ is a support functional of } Q_{\text{ad},1} \text{ at } q^o, \end{aligned} \tag{70}$$

and by Lemma 4.8 we have

$$\psi_2(\dot{Y}) = \psi_{2,1}(\dot{Y}) + \psi_{2,2}(\dot{Y}), \tag{71}$$

where

$$\psi_{2,1} = (f_1, f_2) \in E^*, \quad f_1 \in V^*, \quad f_2 = -[u'(q^o)]^* f_1.$$

Therefore, we get

$$\psi_{2,1}(\dot{Y}) = f_1(\dot{u}) + f_2(\dot{q}) = f_1(\dot{u}) - [u'(q^o)]^* f_1(\dot{q}) \tag{72}$$

and

$$\psi_{2,2}(\dot{Y}) = \sum_{x \in S} \lambda_x \dot{q}(x). \tag{73}$$

Considering $\dot{Y} = (\dot{u}, \dot{q}) \in \tilde{Z}$ we have

$$\dot{q}(x) = 0, \quad x \in S, \tag{74}$$

and by Lemma 4.7, $\dot{u} = u'(q^o)\dot{q}$, so it follows from (72) and (73) that

$$\psi_{2,1}(\dot{Y}) = 0, \quad \forall \dot{Y} \in \tilde{Z}, \tag{75}$$

$$\psi_{2,2}(\dot{Y}) = 0, \quad \forall \dot{Y} \in \tilde{Z}. \tag{76}$$

If we take $\dot{q} \equiv (q - q^o) \in Q$ and $q \in Q_{\text{ad},1}^o$, then by (70)

$$\psi_1(\dot{Y}) = f_2(\dot{q}) = f_2(q - q^o) \geq 0, \quad \forall q \in Q_{\text{ad},1}^o.$$

And then by the continuity of f_2 and the convexity of $Q_{\mathrm{ad},1}$ we have

$$f_2(q - q^o) \geq 0, \quad \forall q \in Q_{\mathrm{ad},1}. \tag{77}$$

Finally, by Lemma 4.9,

$$\begin{aligned}\psi_3(\dot{Y}) = \lambda_0 J'_a(Y_o)(\dot{Y}) + \textstyle\int_D \chi_{\mathfrak{N}}(x,t)\beta_u(x,t,u^o(x,t))\dot{u}(x,t)\,\tilde{\nu}(dxdt),\\ \forall \dot{Y} \in \tilde{Z}.\end{aligned} \tag{78}$$

Substituting (69), (70), (71), and (78) into (68) and considering (75)–(77) one gets

$$\begin{aligned}0 \leq f_2(\dot{q}) = \psi_1(\dot{Y}) = \lambda_0 J'_a(Y_o)(\dot{Y})\\ + \textstyle\int_D \chi_{\mathfrak{N}}(x,t)\beta_u(x,t,u^o(x,t))\dot{u}(x,t)\,\tilde{\nu}(dxdt), \quad \forall \dot{Y} \in \tilde{Z}.\end{aligned} \tag{79}$$

Obviously, $\lambda_0 \neq 0$, otherwise by (69) and (79) one can get $\psi_0 = \psi_1 = \psi_3 = 0$; besides, by (75) and (76), $\psi_2 = 0$, but by Lemma 3.1 this is impossible.

Therefore,

$$J'_a(Y_o)(\dot{Y}) + \int_D \chi_{\mathfrak{N}}(x,t)\beta_u(x,t,u^o(x,t))\dot{u}(x,t)\,\nu(dxdt) \geq 0, \quad \forall \dot{Y} \in \tilde{Z}, \tag{80}$$

where $\nu = \tilde{\nu}/\lambda_0$. Furthermore, by Lemma 4.3 and (16)

$$\begin{aligned}0 \leq &\textstyle\int_\Omega [\phi_u(s,u^o(s,T),q^o(s,T);z(s))\dot{u}(s,T) + \langle \phi_q(s,u^o(s,T),q^o(s,T);z),\dot{q}(s,T)\rangle]\,ds\\ &+ \textstyle\int_D \chi_{\mathfrak{N}}(x,t)\beta_u(x,t,u^o(x,t))\dot{u}(x,t)\,\nu(dxdt)\\ = &\textstyle\int_D [\partial_t \dot{u} - \partial_i(a^o_{ij}\partial_j \dot{u}) + b_i\partial_i\dot{u} + c^o\dot{u}]v\,dxdt\\ &+ \textstyle\int_\Omega \langle \phi_q(s,u^o(s,T),q^o(s,T);z(s)),\dot{q}(s,T)\rangle\,ds.\end{aligned} \tag{81}$$

Because $(\dot{u},\dot{q}) \in \tilde{Z} \subset K_2$, by (52) in Lemma 4.7

$$\begin{aligned}0 \leq &\textstyle\int_\Omega \langle \phi_q(x,u^o(x,T),q^o(x,T);z(x)),\dot{q}(x,T)\rangle\,dx + \int_D [\partial_i(\dot{a}_{ij}\partial_j u^o) - \dot{c}u^o]v\,dxdt\\ = &\textstyle\int_\Omega \langle \phi_q(s,u^o(x,T),q^o(x,T);z(x)),\dot{q}(x,T)\rangle\,dx\\ &- \textstyle\int_D [\dot{a}_{ij}\partial_j u^o\partial_i v + \dot{c}u^o v]\,dxdt.\end{aligned}$$

Taking $\dot{q} = q - q^o$ we immediately have

$$\begin{aligned}&\textstyle\int_D [a_{ij}\partial_j u^o \partial_i v + cu^o v]\,dxdt - \int_\Omega \langle \phi_q(x,u^o,q^o;z),q\rangle\,dx\\ &\quad \leq \textstyle\int_D [a^o_{ij}\partial_j u^o\partial_i v + c^o u^o v]\,dxdt - \int_\Omega \langle \phi_q(x,u^o,q^o;z),q^o\rangle\,dx\\ &\qquad \forall q = (a_{11},\dots,a_{mm},c) \in Q_{\mathrm{ad}},\end{aligned}$$

which is just (15). □

Proof of Theorem 2.2. Obviously, most of the proof is the same as that in Theorem 2.1 except the proof of the inequalities (18) with (19). We have, especially,

$$\psi_0(\dot{Y}) + \psi_3(\dot{Y}) = -\psi_1(\dot{Y}) = -f_2(\dot{q}) \leq 0.$$

By Lemma 4.4 there exists $\mu \in \Phi$ such that

$$\psi_0(\dot{Y}) = -\int_\Omega \chi_{\mathfrak{C}}(x)\dot{u}(x,T)\mathrm{sign}[u^o(x,T) - z(x)]\,\mu(dx),$$

and by Lemma 4.9 there exists $\nu \in rca(D)$ with supp$\nu \subseteq \mathfrak{N}$ and $\nu(dxdt) \geq 0$ such that

$$\psi_3(\dot{Y}) = -\int_D \chi_{\mathfrak{N}}(x,t)\beta_u(x,t,u^o(x,t))\dot{u}(x,t)\,\nu(dxdt).$$

From the above two equalities and the inequality one can get

$$\begin{aligned} &\textstyle\int_D \chi_{\mathfrak{N}}(x,t)\beta_u(x,t,u^o(x,t))\dot{u}(x,t)\,\nu(dxdt) \\ &\quad + \textstyle\int_\Omega \chi_{\mathfrak{C}}(x)\dot{u}(x,T)\,\mathrm{sign}[u^o(x,T) - z(x)]\,\mu(dx) \geq 0. \end{aligned} \tag{82}$$

Owing to (19), one has

$$\int_D [(\partial_t \dot{u} + b_i\partial_i\dot{u} + c^o\dot{u})v + a^o_{ij}\partial_i v\partial_j\dot{u}]\,dxdt \geq 0. \tag{83}$$

Considering $\dot{Y} \in \tilde{Z} \subset K_2$, it follows by (52) in Lemma 4.7 that

$$0 \leq \int_D [\partial_i(\dot{a}_{ij}\partial_j u^o) - \dot{c}u^o]v\,dxdt = -\int_D (\dot{a}_{ij}\partial_j u^o\partial_i v + \dot{c}u^o v)\,dxdt. \tag{84}$$

In particular, take $\dot{q} = q - q^o$ with $q = (a_{11}, \ldots, a_{mm}, c)$ and $q^o = (a^o_{11}, \ldots, a^o_{mm}, c^o)$, and then we can immediately get (18). □

REFERENCES

[1] Yu. E. Anikonov, *Generating functions, evolution equations and inverse problems*, Dokl. Akad. Nauk, 323 (1992), pp. 172–174. (In Russian.)

[2] H. T. Banks and K. Kunisch, *Estimation Techniques for Distributed Parameter Systems*, Birkhäuser, Boston, Basel, Berlin, 1989.

[3] H. T. Banks, S. Reich, and I. G. Rosen, *An approximation theory for the identification of nonlinear distributed parameter systems*, SIAM J. Control Optim., 28 (1990), pp. 523–569.

[4] Yu. Ya. Belov, *On an inverse problem for a semilinear parabolic equation*, Soviet Math. Dokl., 43 (1991), pp. 216–220.

[5] G. Chavent, *Identification of distributed parameter systems: About the output least square method, its implementation and identifiability*, in Identification and System Parameter Estimation, R. Isermann, ed., Pergamon, New York, 1982, pp. 85–97.

[6] Y. M. Chen and F. G. Zhang, *Hierarchical multigrid strategy for efficiency improvement of the* GPST *inversion algorithm*, Appl. Numer. Math., 6 (1990), pp. 431–446.

[7] F. Guyon and J. P. Yvon, *On a weighting method improving identifiability of distributed parameter systems*, in Control of Partial Differential Equations, Lecture Notes in Control and Inform. Sci., 149, Springer-Verlag, Berlin, 1991, pp. 104–119.

[8] S. J. Hu and W. H. Yu, *Identification of floated surface temperature in floated gyroscope*, Proc. 3rd IFAC Symp. Control of Distributed Parameter Systems, Toulouse, France, June 29–July 2, 1982.

[9] V. Isakov, *Inverse Source Problems*, American Mathematical Society, Providence, RI, 1990.

[10] V. M. Isakov, *Inverse parabolic problems with the final overdetermination*, Comm. Pure Appl. Math., 45 (1991), pp. 185–209.

[11] S. Kitamura and S. Nakagiri, *Identifiability of spatially-varying and constant parameters in distributed systems of parabolic type*, SIAM J. Control Optim., 15 (1977), pp. 785–802.

[12] A. Pierce, *Unique identification of eigenvalues and coefficients in a parabolic problem*, SIAM J. Control Optim., 17 (1979), pp. 494–499.

[13] T. Suzuki, *Gel'fand-Levitan's theory and related inverse problems*, in Inverse Problems, Proc. conference held at the Mathematical Research Institute at Oberwolfach, Black Forest, May 18–24, 1986, J. R. Cannon and U. Hornung, eds., Birkhäuser-Verlag, Basel, Boston, Stuttgart, 1986.

[14] W. H. Yu, *On well-posedness of inverse problems for a class of hyperbolic equation*, Chinese Math. Ann., 10 (1988), pp. 1-9.

[15] W. H. YU AND J. H. SEINFELD, *Identification of parabolic distributed parameter systems by regularization with differential operator*, J. Math. Anal. Appl., 132 (1988), pp. 365–398.

[16] W. H. YU, *Well-posedness of inverse problems for a class of parabolic systems*, Res. Comm. Math., 8 (1988), pp. 95–100. (In Chinese.)

[17] W. H. YU AND J. H. SEINFELD, *Identification of distributed parameter systems with pointwise constrains on the parameters*, J. Math. Anal. Appl., 136 (1988), pp. 497–520.

[18] W. H. YU, *Well-posedness of determining the source term of an elliptic equation*, Bull. Austral. Math. Soc., to appear.

[19] J. MOSSINO, *An application of duality to distributed optimal control problems with constraints on the control and the state*, J. Math. Anal. Appl., 50 (1975), pp. 223–242.

[20] A. PARAGEORGIOU AND N. S. PARAGEORGIOU, *Necessary and sufficient conditions for optimality in nonlinear distributed parameter systems with variable initial state*, J. Math. Soc. Japan, 42 (1990), pp. 387–396.

[21] D. TIBA AND M. TIBA, *Approximation for control problems with pointwise state constraints*, Internat. Ser. Numer. Math., 91 (1989), pp. 379–390.

[22] W. H. YU, *On maximum principle for optimality of quasi-linear parabolic systems*, Proc. 4th IFAC Symp. Control of Distributed Parameter Systems, Los Angeles, 1986.

[23] N. DUNFORD AND J. T. SCHWARTZ, *Linear Operators*, Vol. I, Interscience Publishers, Inc., New York, 1958.

[24] A. FRIEDMAN, *Partial Differential Equations of Parabolic Type*, Prentice-Hall, Inc., Englewood Cliffs, NJ, 1964.

[25] I. V. GIRSANOV, *Lectures on Mathematical Theory of Extremum Problems*, Lecture Notes in Economics and Mathematical Systems, Vol. 67, Springer-Verlag, Berlin, Heidelberg, New York, 1972.

[26] J. L. LIONS AND E. MAGENES, *Non-Homogeneous Boundary Value Problems and Applications*, Vols. I and II, Springer-Verlag, New York, Berlin, 1972.

[27] L. A. LUSTERNIK AND V. J. SOBOLEV, *Elements of Functional Analysis*, Hindustan Publishing Corp., Delhi, India, 1961.

[28] E. ZEIDLER, *Nonlinear Functional Analysis and Its Applications, Vol.* 3: *Variational Methods and Optimization* (Transl. from German by L. F. Boroll), Springer-Verlag, Berlin, New York, 1985.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 33, No. 6, pp. 1816–1833, November 1995

© 1995 Society for Industrial and Applied Mathematics
010

A CONVEX APPROACH TO THE MIXED $\mathcal{H}_2/\mathcal{H}_\infty$ CONTROL PROBLEM FOR DISCRETE-TIME UNCERTAIN SYSTEMS*

J. C. GEROMEL†, P. L. D. PERES†, AND S. R. SOUZA‡

Abstract. This paper considers $\mathcal{H}_2/\mathcal{H}_\infty$ control problems involving discrete-time uncertain linear systems. The uncertain domain is supposed to be convex bounded, which naturally covers, as a particular case, the important class of interval matrices. The $\mathcal{H}_\infty$ guaranteed-cost control problem, solved for this class of uncertain systems, under no matching conditions, can be stated as follows: determine a state feedback gain (if one exists) such that the $\mathcal{H}_\infty$ norm of a given transfer function remains bounded by a prespecified level for all possible models. In the same context, problems on the determination of the smallest $\mathcal{H}_\infty$ upper bound and the minimization of an $\mathcal{H}_2$ cost upper bound subject to $\mathcal{H}_\infty$ constraints are also addressed. The results follow from the fact that those problems are convex in the particular parametric space under consideration. Some examples illustrate the theory.

Key words. discrete-time system, uncertain systems, mixed $\mathcal{H}_2/\mathcal{H}_\infty$ control, convex analysis

AMS subject classifications. 93C15

1. Introduction.

In the last decade, $\mathcal{H}_\infty$ control theory has been largely developed in both frequency and state-space approaches. Concerning continuous-time systems, [4] states elegantly the optimal solutions of $\mathcal{H}_\infty$ and $\mathcal{H}_2$ control problems in terms of algebraic Riccati equations where both state feedback and observer-based output feedback are considered (see [4] and [21] and the references therein). However, few papers have proposed methods that take into account uncertain models. In [13], $\mathcal{H}_\infty$ control by state feedback is related to quadratic stabilizability in norm-bounded uncertain domains, but no control synthesis is proposed. On the other hand, methods such as the one introduced in [23] do not consider any uncertainty acting on the input matrix or assume some kind of matching conditions. This obviously restricts the class of uncertain systems to be dealt with. Quadratic stabilizability of continuous-time systems under convex-bounded uncertainties, by means of a state feedback control and with prescribed $\mathcal{H}_\infty$ norm upper bound, is addressed in [7].

The development is not quite the same when one regards discrete-time systems. Necessary and sufficient conditions have appeared in [5], [12], [17], and [24] for known models (i.e., all parameters are precisely known), relating discrete-time Riccati-like equations and $\mathcal{H}_\infty$ norm bounds. An interesting work [15], dealing with mixed $\mathcal{H}_2/\mathcal{H}_\infty$ control via convex optimization, focuses on precisely known systems. Using an appropriate matrix transformation, the state feedback control problem with mixed $\mathcal{H}_2/\mathcal{H}_\infty$ performance measure, the same used in [11], can be reduced to the minimization of a nonlinear—but convex—function over a bounded convex set of real matrices.

However, the extension of the above necessary and sufficient conditions to uncertain systems is not straightforward and, if possible, may lead to highly nonlinear problems to be solved. In [22], discrete-time norm-bounded uncertain systems are

* Received by the editors October 8, 1992; accepted for publication (in revised form) June 10, 1994. This research was supported in part by grants from Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP) and Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq).

† Laboratory of Convex Analysis, Faculty of Electrical Engineering, University of Campinas, CP 6101, 13081-970, Campinas, SP, Brazil (geromel@dt.fee.unicamp.br).

‡ Department of Electronic and Systems, School of Electrical Engineering, Federal University of Goiás, Praça Universitária s/n, 74605-220, Goiânia, GO, Brazil.

investigated under $\mathcal{H}_\infty$ constraints. Sufficient conditions for quadratic stability with disturbance attenuation are provided, which circumvents the above-cited difficulty. The results presented in the control synthesis problem suppose the existence of a dynamic output controller, which remains to be determined.

This paper is devoted to analyze the convexity and solve several problems involving discrete-time uncertain systems. First we consider the $\mathcal{H}_\infty$ guaranteed cost control problem, for which preliminary results have already been presented in [19]. It consists of the determination of a constant state feedback control gain such that a certain closed-loop transfer function remains bounded by a prespecified $\mathcal{H}_\infty$ level for all uncertainties varying in a convex-bounded domain. Further, related control problems are also analyzed and solved. For instance, the above $\mathcal{H}_\infty$ norm bound may be included in the optimization procedure in order to determine the smallest feasible one [1], [14]. Moreover, an approximate version of the the mixed $\mathcal{H}_2/\mathcal{H}_\infty$ control problem is considered. It is important to emphasize that all results provided here follow from the definition of a special parameter space [6] on which the above problems are shown to be convex.

This paper is organized as follows. The next section introduces the system to be dealt with and basic assumptions needed throughout. Section 3 is devoted to the parametrization of $\mathcal{H}_2$ and $\mathcal{H}_\infty$ norms in terms of algebraic Riccati inequalities. In §4, the $\mathcal{H}_\infty$ guaranteed-cost control problem is solved, including the determination of the smallest $\mathcal{H}_\infty$ norm bound. The approximate mixed $\mathcal{H}_2/\mathcal{H}_\infty$ control problem is addressed in §5, where, in addition, decentralized control structures are considered and compared with [10]. Section 6 particularizes the results to the case of precisely known discrete-time systems and compares them with the ones provided in [24]. The examples included in §7 illustrate the theory, and in §8 the conclusion follows.

The notation used throughout is standard. The set of real numbers is denoted by $\Re$, Tr $(\cdot)$ means the trace of $(\cdot)$, $(')$ stands for transpose of matrices and vectors, and $\langle A, B\rangle = \mathrm{Tr}\,(A'B)$ denotes the inner product of matrices with appropriate dimensions. Singular values and eigenvalues are denoted by $\sigma(\cdot)$ and $\lambda(\cdot)$, respectively. Finally, $\mathcal{H}_\infty$ and $\mathcal{H}_2$ norms are defined as follows:

$$\| H\|_\infty \triangleq \max_{\omega\in[-\pi,\pi]} \sigma_{\max}\left[H(e^{j\omega})\right],$$

$$\| H\|_2^2 \triangleq \frac{1}{2\pi}\int_{-\pi}^{\pi} \mathrm{Tr}\left\{H(e^{-j\omega})'H(e^{j\omega})\right\} d\omega.$$

2. Preliminaries. Let us consider the following LTI discrete-time system:

$$(1)\qquad \left\{\begin{array}{rcl} x(k+1) & = & Ax(k) + B_1w(k) + B_2u(k), \\ u(k) & = & -Kx(k), \\ z(k) & = & C_1x(k) + D_1u(k), \end{array}\right.$$

where $x(k) \in \Re^n$ is the state, $u(k) \in \Re^m$ is the control, $z(k) \in \Re^q$ is the controlled output, and $w(k) \in \Re^l$ is the external disturbance. It is assumed that all matrices are of appropriate and known dimensions and, as usual, $C_1'D_1 = 0$ and $D_1'D_1 > 0$. In fact, the above orthogonality hypothesis could be weakened with minor changes in the results that follow. Associated with (1), we define the following extended matrices

$(p = n + m)$:

$$F = \begin{bmatrix} A & -B_2 \\ \mathbf{0} & \mathbf{0} \end{bmatrix} \in \Re^{p\times p}, \quad G = \begin{bmatrix} \mathbf{0} \\ \mathbf{I} \end{bmatrix} \in \Re^{p\times m} \tag{2}$$

and

$$Q = \begin{bmatrix} B_1B_1' & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} \in \Re^{p\times p}, \quad R = \begin{bmatrix} C_1{}'C_1 & \mathbf{0} \\ \mathbf{0} & D_1'D_1 \end{bmatrix} \in \Re^{p\times p}. \tag{3}$$

It is further assumed that matrices Q and R are precisely known. On the contrary, matrix F, which defines the open-loop model, belongs to a convex-bounded uncertain polyhedral domain $\mathcal{D}$, defined as [6]

$$\mathcal{D} \overset{\triangle}{=} \left\{ F \ : \ F = \sum_{i=1}^{N} \xi_i F_i \ , \ \xi_i \geq 0 \ , \ \sum_{i=1}^{N} \xi_i = 1 \right\}. \tag{4}$$

That is, any feasible F can be expressed as an unknown convex combination of the "extreme matrices" $F_i \sim (A, B_2)_i$, $i = 1, \ldots, N$. At this point, it is important to stress that any polyhedral convex and bounded domain can be expressed as (4) by a suitable choice of F_i, $i = 1, \ldots, N$ and N. It is clear that $\mathcal{D}$ generalizes the well known and important case of interval matrices uncertainties.

For each feasible model $F \in \mathcal{D}$ and $K \in \Re^{m\times n}$, the closed-loop transfer function from $w(k)$ to $z(k)$ is given by

$$H_F(\zeta) \overset{\triangle}{=} C_{\text{cl}} \left[\zeta \mathbf{I} - A_{\text{cl}} \right]^{-1} B_1, \tag{5}$$

where $C_{\text{cl}} = C_1 - D_1K$ and $A_{\text{cl}} = A - B_2K$. Now, restricting our attention to those gains such that $K \in \mathcal{K}_F$, with $\mathcal{K}_F$ being the set of all stabilizing state feedback gains for the model defined by $F \in \mathcal{D}$,

$$\mathcal{K}_F \overset{\triangle}{=} \left\{ \ K \in \Re^{m\times n} \ : \ A_{\text{cl}} \text{ asympt. stable} \ \right\}, \tag{6}$$

the norms $\|H_F\|_2$ and $\|H_F\|_\infty$ are easily calculated [12]. Indeed, the $\mathcal{H}_2$ norm is given by

$$\|H_F\|_2^2 = \text{Tr}\left(C_{\text{cl}} L_c C_{\text{cl}}' \right) = \text{Tr}\left(B_1' L_o B_1 \right), \tag{7}$$

where L_o and L_c are the observability and controllability Gramians, respectively. That is, L_o and L_c are the solutions to the linear equations

$$\begin{aligned} A_{\text{cl}} L_c A_{\text{cl}}' - L_c + B_1B_1' &= 0, \\ A_{\text{cl}}' L_o A_{\text{cl}} - L_o + C_{\text{cl}}' C_{\text{cl}} &= 0. \end{aligned} \tag{8}$$

The determination of the $\mathcal{H}_\infty$ norm is numerically more involved. The results presented in [12] can be used to define a unidimensional search procedure for its calculation. As will be clear later, neither $\mathcal{H}_2$ nor $\mathcal{H}_\infty$ norms have to be calculated explicitly. From the above discussion, it is clear that the real-valued functions $g_F(K) = \|H_F\|_2$ and $h_F(K) = \|H_F\|_\infty$ are well defined for all elements of the set $\mathcal{K}_F$.

Let us state the problems to be dealt with in the rest of the paper. The reader is referred to the appendix A for the definition of the epigraph of real-valued functions.

$\mathcal{H}_\infty$ guaranteed-cost control (P1). Given a prespecified $\mathcal{H}_\infty$ norm level γ, find a constant state feedback gain K such that $K \in \mathcal{K}_F$ and $\|H_F\|_\infty \leq \gamma \; \forall F \in \mathcal{D}$. This is equivalent to the determination of K such that $(K, \gamma) \in \text{epi } h_F \; \forall \; F \in \mathcal{D}$. Note that in the case of precisely known systems ($N = 1$), the above problem reduces to the one considered in [12], [24].

Optimal $\mathcal{H}_\infty$ guaranteed-cost control (P2). It is a natural generalization of the previous one, where γ itself is involved in the optimization process, that is,

$$\min \left\{ \gamma \; : \; (K, \gamma) \in \text{epi } h_F \quad \forall \; F \in \mathcal{D} \right\}. \tag{9}$$

Obviously, in the case $N = 1$, (9) reduces to the classical $\mathcal{H}_\infty$ optimal control problem $\min\{\|H_F\|_\infty \; : \; K \in \mathcal{K}_F\}$.

Mixed $\mathcal{H}_2/\mathcal{H}_\infty$ guaranteed-cost control (P3). Given a prespecified $\mathcal{H}_\infty$ norm level γ, find a constant state feedback gain K and the smallest $\mathcal{H}_2$ norm level β such that $K \in \mathcal{K}_F$, $\|H_F\|_\infty \leq \gamma$, and $\|H_F\|_2 \leq \beta \; \forall \; F \in \mathcal{D}$. Equivalently, it can be restated as

$$\min \left\{ \beta \; : \; (K, \beta) \in \text{epi } g_F \; , \; (K, \gamma) \in \text{epi } h_F \quad \forall \; F \in \mathcal{D} \right\}. \tag{10}$$

Once again, in the case ($N = 1$), the above problem amounts to the determination of a constant state feedback gain K such that

$$\min \left\{ \|H_F\|_2 \; : \; \|H_F\|_\infty \leq \gamma \; , \; K \in \mathcal{K}_F \right\}, \tag{11}$$

which, after some approximations, has already been solved in [1] and [14] for continuous-time linear systems.

To the authors' knowledge, there are no available results for the above problems for discrete-time uncertain systems in convex-bounded domains. Even then, the motivation for introducing them is quite clear in the sense that they generalize the most important and well-known $\mathcal{H}_\infty$ control design results available for precisely known models to uncertain systems.

Note that (P1) is a generalization of the state-space $\mathcal{H}_\infty$ control approach proposed in [12] and [24]. The same is true for (P2), on which γ is included as an additional decision variable in the optimization process. In principle, results available to solve (P1) can also be used for solving (P2) iteratively. For instance, consider the following procedure: set γ^k sufficiently large, solve (P1), redefine $\gamma^{k+1} = \max_{F \in \mathcal{D}} \; \|H_F\|_\infty$, and iterate until convergence. Since the sequence $\{\gamma^k\}$ is bounded below and $\gamma^{k+1} \leq \gamma^k$, it converges but not necessarily to the optimal solution of (P2). Furthermore, it may be time consuming due to the fact that its rate of convergence is normally poor and γ^{k+1} is hard to obtain. For this reason, we propose here a new method which solves (P2) directly and provides both the stabilizing state feedback gain and the optimal $\mathcal{H}_\infty$-norm level. Finally, it is important to add that (P3) is a generalization to discrete-time uncertain systems of the mixed $\mathcal{H}_2/\mathcal{H}_\infty$ control problem solved in [1] and [14].

As stated, problems (P1)–(P3) are difficult to solve. The main difficulty stems from the nonconvexity of the set $\mathcal{K}_F$ and functions $g_F(\cdot)$ and $h_F(\cdot)$ in the parameter space generated by the elements of K, for each $F \in \mathcal{D}$. The obvious conclusion is that some kind of approximation or majorization should be adopted in order to render them more tractable. In this way we consider the following:

(i) Instead of $\mathcal{K} = \bigcap_{F\in\mathcal{D}} \mathcal{K}_F$, the set of all state feedback gains which stabilize simultaneously all models $F \in \mathcal{D}$, we restrict our attention to the subset $\mathcal{K}_Q \subseteq \mathcal{K}$ composed by all state feedback gains which stabilize quadratically the uncertain system (1) and impose on it a certain disturbance attenuation level γ (see Definition 4.1).

(ii) The $\mathcal{H}_2$ norm is replaced by a suitable and convex upper bound. On the other hand, the inequality $\|H_F\|_\infty \leq \gamma$ is tested by means of a simple sufficient condition yielding only convex problems to be solved.

Throughout the paper, we assume that matrix $\mathcal{W} \in \Re^{p\times p}$, partitioned as

$$\mathcal{W} = \begin{bmatrix} W_1 & W_2 \\ W_2' & W_3 \end{bmatrix} \geq 0, \tag{12}$$

is symmetric with $W_1 > 0 \in \Re^{n\times n}$, $W_2 \in \Re^{n\times m}$, and $W_3 \in \Re^{m\times m}$. The null space of G' is denoted by $\mathcal{N}$. From (2), it is clear that all $v \in \mathcal{N}$ have the particular structure $v' = [x' \ \ 0]$ with $x \in \Re^n$. In addition, we define the matrix functions $\Theta_{\infty F}(\mathcal{W}, \mu) = F\mathcal{W}F' - \mathcal{W} + \mathcal{W}R\mathcal{W} + \mu Q$ and $\Theta_{\infty i}(\cdot)$, which is the same as $\Theta_{\infty F}(\cdot)$, but with F replaced by the extreme matrix F_i, $i = 1, \ldots, N$. Furthermore, we suppose that $\forall\ F \in \mathcal{D}$, the pair (A, B_1) is controllable and range $(B_2) \subseteq$ range (B_1). This condition implies that (A_{cl}, B_1) is controllable $\forall\ K \in \Re^{m\times n}$ and $\forall\ F \in \mathcal{D}$. It is simple to verify that this assumption can be dropped if the inequalities, given in what follows, involving functions $\Theta_{\infty F}(\cdot)$ and $\Theta_{\infty i}(\cdot)$, $i = 1, \ldots, N$ are strict.

3. $\mathcal{H}_2$ and $\mathcal{H}_\infty$ norm bounds. In this section, we consider $F \in \mathcal{D}$ fixed but arbitrary. Our aim is to establish bounds to $\mathcal{H}_2$ and $\mathcal{H}_\infty$ norms in terms of an algebraic Riccati inequality.

LEMMA 3.1. *Given $(K, \gamma > 0)$, if the matrix inequality*

$$A_{\text{cl}}PA_{\text{cl}}' - P + PC_{\text{cl}}'C_{\text{cl}}P + \gamma^{-2}B_1B_1' \leq 0 \tag{13}$$

admits a symmetric positive definite solution, then $K \in \mathcal{K}_F$ and

$$H_F(\zeta)H_F(\zeta^{-1})' \leq \gamma^2(\mathbf{I} - C_{\text{cl}}PC_{\text{cl}}'), \qquad \zeta = e^{j\omega} \quad \forall\, \omega \in [-\pi, \pi]. \tag{14}$$

Proof. For the proof of this lemma, see appendix A. □

This result is a generalization of the one presented in [10]. Indeed, it provides a sufficient condition under which upper bounds to the closed-loop transfer function $\mathcal{H}_2$ and $\mathcal{H}_\infty$ norms can be determined. In addition, it enables the definition of an estimate of the difference between the actual $\|H_F\|_\infty$ and the prescribed value γ.

LEMMA 3.2. *Given $(K, \gamma > 0)$, suppose that* (13) *holds for some positive definite matrix $P \in \Re^{n\times n}$. The following hold:*

(a) *There exists* $0 \leq \theta^2 \leq \text{Tr}\ (C_{\text{cl}}PC_{\text{cl}}')$ *such that* $\|H_F\|_\infty^2 \leq \gamma^2(1 - \theta^2) \leq \gamma^2$;

(b) $\|H_F\|_2^2 \leq \gamma^2\text{Tr}\ (C_{\text{cl}}PC_{\text{cl}}')$.

Proof. For the proof of this lemma, see appendix A. □

To provide an interpretation of the above results, it is important to keep in mind that, for a given pair (K, γ), the solution of (13) is not necessarily unique and that the exact value of θ^2 cannot be determined a priori since (see appendix A) it depends on $\|H_F\|_\infty$ itself. As a result, the only information we get from part (a) of Lemma 3.2 is $\|H_F\|_\infty \leq \gamma$ or equivalently that $(K, \gamma) \in$ epi h_F whenever (13) admits a positive definite solution. However, from the part (b) of Lemma 3.2 we notice that for (K, γ) fixed, the best upper bound to the $\|H_F\|_2$ is determined by choosing P which minimizes Tr $(C_{\text{cl}}PC_{\text{cl}}')$ among all positive definite solutions of (13). By doing this, θ^2

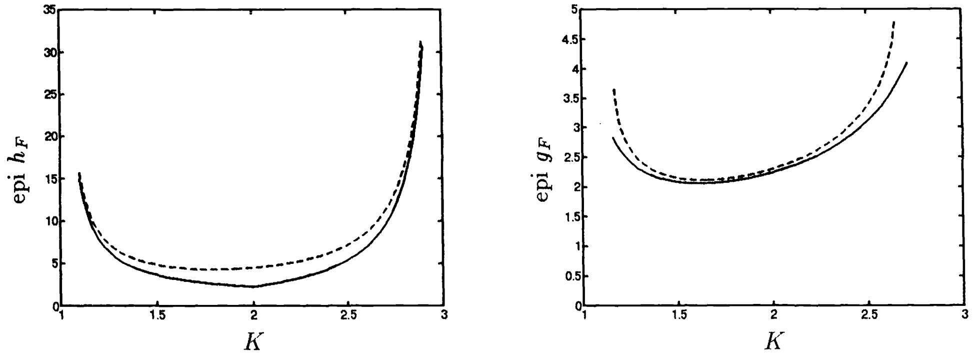


FIG. 1. epi h_F *(—), epi* g_F *(—), and approximations (– –).*

is constrained to be small, which implies that the inequality $\|H_F\|_\infty \le \gamma$ becomes tighter.

The above discussion will be important in the sequel. For the moment, let us compare numerically the results provided by Lemma 3.2 and the exact ones. Consider a linear time-invariant system such that

$$F = \begin{bmatrix} 2 & -1 \\ 0 & 0 \end{bmatrix}; \; R = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}; \; Q = \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix}. \tag{15}$$

In this case, it is obvious that $\mathcal{K}_F = \{K \in \Re \;:\; 1 < K < 3\}$. Figure 1 shows the epi h_F and its approximation given by all pairs (K, γ) such that (13) holds for some $P > 0$. Figure 1 also shows the epi g_F constrained to those $K \in \mathcal{K}_F$ such that $\|H_F\|_\infty \le \gamma = 10$ and its approximation provided by part (b) of Lemma 3.2. The solution of (P2) is calculated as $K = 2$ (exact) and $K \cong 1.8$ (approximate). For (P3), both (exact and approximate) coincide at $K \cong 1.6$. This fact (see also the numerical examples in §7) gives a measure of the "quality" of approximations introduced.

In addition, note that, in the general case, all shapes in Fig. 1 are nonconvex. Fortunately, those defined by the approximations can be generated from a nonlinear mapping acting on the elements of a convex set. This key property is proved below. Before that, let us introduce the following definition.

DEFINITION 3.3. *The pair* (A, B_2) *is said to be stabilizable with* γ *disturbance attenuation if there exist* $K \in \mathcal{K}_F$ *and* $P > 0$ *such that* (13) *holds.*

THEOREM 3.4. *Define the set* $\mathcal{C}_{\infty F}$ *as*

$$\mathcal{C}_{\infty F} \overset{\triangle}{=} \left\{ (\mathcal{W}, \mu) \;:\; \mathcal{W} \ge 0, \mu \ge 0 \quad , v'\Theta_{\infty F}(\mathcal{W}, \mu)v \le 0 \;\; \forall \, v \in \mathcal{N} \right\}. \tag{16}$$

Then $\mathcal{C}_{\infty F}$ *is convex and* $\{(W_2'W_1^{-1}, 1/\sqrt{\mu}) \;:\; (\mathcal{W}, \mu) \in \mathcal{C}_{\infty F}\} \subseteq$ epi h_F. *Furthermore, the pair* (A, B_2) *is stabilizable with* γ *disturbance attenuation if and only if, for some* $\mathcal{W}$, $(\mathcal{W}, \gamma^{-2}) \in \mathcal{C}_{\infty F}$.

Proof. Let us first prove that $\mathcal{C}_{\infty F}$ is convex. Indeed, since the set of nonnegative definite matrix is convex, it remains to show that for an arbitrary $v \in \mathcal{N}$, the function $v'\Theta_{\infty F}(\mathcal{W}, \mu)v$ is convex. To this end, note that

$$\nabla^2_\alpha \; v'\Theta_{\infty F}(\mathcal{W} + \alpha\Delta\mathcal{W}, \mu + \alpha\Delta\mu)v = 2v'\Delta\mathcal{W}R\Delta\mathcal{W}v \ge 0 \tag{17}$$

holds for all $\Delta\mathcal{W} = \Delta\mathcal{W}' \in \Re^{p\times p}$ and $\Delta\mu \in \Re$, then the convexity of $\mathcal{C}_{\infty F}$ follows. Now suppose that $(\mathcal{W}, \mu) \in \mathcal{C}_{\infty F}$. From (16), for all $v \in \mathcal{N}$,

$$0 \ge v'\Theta_{\infty F}(\mathcal{W}, \mu)v$$

$$\geq x'\Big[(A - B_2W_2'W_1^{-1})W_1(A - B_2W_2'W_1^{-1})' - W_1$$

$$+W_1(C_1 - D_1W_2'W_1^{-1})'(C_1 - D_1W_2'W_1^{-1})W_1$$

$$+ \mu B_1B_1' + B_2(W_3 - W_2'W_1^{-1}W_2)B_2'\Big]x. \tag{18}$$

Since $\mathcal{W} \geq 0$ implies $W_3 - W_2'W_1^{-1}W_2 \geq 0$, the conclusion is that (13) holds for $K = W_2'W_1^{-1}$, $\gamma = 1/\sqrt{\mu}$, and $P = W_1 > 0$. Consequently, from Lemma 3.2, $(K, \gamma) \in$ epi h_F. On the other hand, suppose the pair (A, B_2) is stabilizable with γ disturbance attenuation. From Definition 3.3, we know that there exist K and $P > 0$ such that (13) holds. Consequently, for all $x \in \Re^n$ we have

$$0 \geq x'\Big[(A - B_2K)P(A - B_2K)' - P$$

$$+ P(C_1 - D_1K)'(C_1 - D_1K)P + \gamma^{-2}B_1B_1'\Big]x$$

$$\geq v'\Theta_{\infty F}(\mathcal{W}, \gamma^{-2})v, \tag{19}$$

where the last inequality holds for

$$\mathcal{W} = \begin{bmatrix} P & PK' \\ KP & KPK' \end{bmatrix} \geq 0, \tag{20}$$

implying that $(\mathcal{W}, \gamma^{-2}) \in \mathcal{C}_{\infty F}$, and the proof is concluded. □

THEOREM 3.5. *Consider the convex set $\mathcal{C}_{\infty F}$ defined in Theorem* 3.4 *and assume that $\gamma > 0$ is given. Then $\{(W_2'W_1^{-1}, \beta) \ : \ (\mathcal{W}, \gamma^{-2}) \in \mathcal{C}_{\infty F}\} \subseteq$* epi g_F *for $\beta^2 = \gamma^2$*Tr $(R\mathcal{W})$.

Proof. Consider any $\mathcal{W}$ such that $(\mathcal{W}, \gamma^{-2}) \in \mathcal{C}_{\infty F}$. From (18) we note that (13) holds with $K = W_2'W_1^{-1}$ and $P = W_1 > 0$. We use part (b) of Lemma 3.2 to get

$$\|H_F\|_2^2 \leq \gamma^2 \mathrm{Tr}\,\Big((C_1 - D_1W_2'W_1^{-1})W_1(C_1 - D_1W_2'W_1^{-1})'\Big)$$

$$\leq \gamma^2 \mathrm{Tr}\,\Big(C_1W_1C_1' + D_1W_2'W_1^{-1}W_2D_1'\Big)$$

$$\leq \gamma^2 \mathrm{Tr}\,\Big(C_1W_1C_1' + D_1W_3D_1'\Big)$$

$$\leq \gamma^2 \mathrm{Tr}\,\Big(R\mathcal{W}\Big) \ = \ \beta^2, \tag{21}$$

which implies that $(W_2'W_1^{-1}, \beta) \in$ epi g_F. The proof is complete. □

As noticed before, both Theorem 3.4 and Theorem 3.5 provide convex approximations to epi h_F and epi g_F, respectively. Each of them is generated by a nonlinear mapping with the same convex domain $\mathcal{C}_{\infty F}$. As a result, the main difficulty to solve (P1)–(P3) directly in terms of the state feedback gain matrix K is circumvented by simply reformulating them as convex programming problems over $\mathcal{C}_{\infty F}$. Only after their solutions have been calculated, the corresponding feedback gain K is determined by $K = W_2'W_1^{-1}$ for the optimal $\mathcal{W} \in \mathcal{C}_{\infty F}$. Furthermore, the same reasoning can be easily generalized to deal with uncertain systems in convex-bounded domains.

4. The $\mathcal{H}_\infty$ guaranteed-cost control problem. This section is completely devoted to solve (P1) and (P2). For that, the key observation (see [6]) is that for any $\mathcal{W} = \mathcal{W}' \geq 0$ and $\mu \geq 0$, the function $\Theta_{\infty F}(\mathcal{W}, \mu)$ is convex with respect to $F \in \mathcal{D}$. Using the notation introduced before, this means that

$$\Theta_{\infty F}(\mathcal{W}, \mu) \leq \sum_{i=1}^{N} \xi_i \Theta_{\infty i}(\mathcal{W}, \mu) \tag{22}$$

holds for $\xi_i \geq 0$, $\sum_{i=1}^{N} \xi_i = 1$. Before proceeding, we need the following definition.

DEFINITION 4.1. *System* (1) *is said to be quadratically stabilizable with γ disturbance attenuation if there exist $K \in \mathcal{K}_F$ and $P > 0$ such that* (13) *holds for all $F \in \mathcal{D}$.*

The set of all state feedback matrix gain K satisfying the above definition is denoted by $\mathcal{K}_Q$. It is important to keep in mind that the concept of quadratic stabilizability with γ disturbance attenuation implies that for $K \in \mathcal{K}_Q$, the same matrix $P > 0$ should satisfy (13) for all $F \in \mathcal{D}$.

THEOREM 4.2. *Define the convex set $\mathcal{C}_{\infty i}$ as $\mathcal{C}_{\infty F}$ for all extreme matrices $F = F_i$, $i = 1, \ldots, N$. The following hold:*

(a) $\mathcal{C}_\infty \overset{\triangle}{=} \bigcap_{i=1}^{N} \mathcal{C}_{\infty i} \equiv \bigcap_{F \in \mathcal{D}} \mathcal{C}_{\infty F}$,

(b) $\{(W_2' W_1^{-1}, 1/\sqrt{\mu}) \ : \ (\mathcal{W}, \mu) \in \mathcal{C}_\infty\} \subseteq \bigcap_{F \in \mathcal{D}} \text{epi } h_F$.

Furthermore, the system (1) *is quadratically stabilizable with γ disturbance attenuation if and only if for some $\mathcal{W}$, $(\mathcal{W}, \gamma^{-2}) \in \mathcal{C}_\infty$.*

Proof. Only part (a) will be proved. The other statement follows immediately from part (a) and Theorem 3.4. First suppose that $(\mathcal{W}, \mu) \in \mathcal{C}_{\infty i}$, $i = 1, \ldots, N$. In this case, $v' \Theta_{\infty i}(\mathcal{W}, \mu) v \leq 0$, $i = 1, \ldots, N \ \forall \ v \in \mathcal{N}$. However, each $F \in \mathcal{D}$ can be expressed as a convex combination of the extreme matrices F_i, $i = 1, \ldots, N$. Consequently, with (22) we have

$$\begin{aligned} v' \Theta_{\infty F}(\mathcal{W}, \mu) v &\leq \sum_{i=1}^{N} \xi_i v' \Theta_{\infty i}(\mathcal{W}, \mu) v \\ &\leq 0 \quad \forall \, v \in \mathcal{N}, \quad \forall \, F \in \mathcal{D}, \end{aligned} \tag{23}$$

meaning that $(\mathcal{W}, \mu) \in \mathcal{C}_{\infty F}$, $\forall \ F \in \mathcal{D}$. The converse is straightforward because $F_i \in \mathcal{D}$, $i = 1, \ldots, N$. This concludes the proof of the theorem. □

This results furnishes the theoretical basis to the solution of both problems (P1) and (P2). Indeed, as a consequence of the convexity of the uncertain domain $\mathcal{D}$, a convex subset of epi h_F $\forall \ F \in \mathcal{D}$ can be easily generated only from the N extreme matrices that define $\mathcal{D}$. In addition, using the last part of Theorem 4.2 we conclude immediately that

$$\mathcal{K}_Q = \left\{ W_2' W_1^{-1} \ : \ (\mathcal{W}, \gamma^{-2}) \in \mathcal{C}_\infty \right\}. \tag{24}$$

Viewed the intricate dependence of (13) on K, it is somewhat surprising that $\mathcal{K}_Q$ could be generate from the elements of the convex set $\mathcal{C}_\infty$. Furthermore, given $\gamma > 0$, any $K \in \mathcal{K}_Q$ solves (P1) and the global solution of the convex programming problem

$$\max \ \left\{ \ \mu \ : \ (\mathcal{W}, \mu) \in \mathcal{C}_\infty \ \right\} \tag{25}$$

provides $K = W_2' W_1^{-1}$ and $\gamma = 1/\sqrt{\mu}$ such that the pair (K, γ) is suboptimal to (P2).

It is obvious that $\gamma = 1/\sqrt{\mu}$ is the minimum value of γ as far as the approximation introduced in part (b) of Theorem 4.2 is adopted. As noticed before, we claim that this approximation is good enough (see the numerical examples included in §7) to produce near-optimal solutions to (P2).

Being convex, the global optimal solution of (25) can be found by many powerful convex programming methods available in the literature (for details, see [2], [3], [8], and [9]). However, from the numerical point of view, the difficulty in solving (25) may be increased due to the fact that μ is not explicitly bounded. For this reason, we now determine an upper bound to μ such that the optimal solution of (25) does not change.

THEOREM 4.3. *Consider F fixed but arbitrary. Assume that matrix $V_F \in \Re^{l \times p}$ given by*

$$V_F \triangleq (B_1' B_1)^{-1} B_1' \begin{bmatrix} A - \mathbf{I} & -B_2 \end{bmatrix} \tag{26}$$

is of full rank and the perturbations $w(k)$ are "rich" in the sense that $\dim(w(k)) > n -$ rank (C_1)*. Then $(\mathcal{W}, \mu) \notin \mathcal{C}_{\infty F}$ $\forall$ $\mu > \overline{\mu}(F)$, where*

$$\overline{\mu}(F) = \min \left\{ \mu > 0 \ : \ \det(V_F' V_F - \mu R) = 0 \right\}. \tag{27}$$

Proof. Define $R_\epsilon = R + \epsilon \mathbf{I}$ for $\epsilon > 0$ sufficiently small and $\tilde{F} = F - \mathbf{I}$. After simple but tedious algebraic manipulations, we can show that the following factorization holds:

$$\begin{aligned} F\mathcal{W}F' - \mathcal{W} + \mathcal{W}R_\epsilon \mathcal{W} + \mu Q = {} & (\tilde{F} R_\epsilon^{-1} + \mathcal{W}) R_\epsilon (\tilde{F} R_\epsilon^{-1} + \mathcal{W})' \\ & + \tilde{F} \mathcal{W} \tilde{F}' - \tilde{F} R_\epsilon^{-1} \tilde{F}' + \mu Q. \end{aligned} \tag{28}$$

Since the left-hand side of (28) goes to $\Theta_{\infty F}(\mathcal{W}, \mu)$ as $\epsilon \to 0^+$ then, $\forall$ $v \in \mathcal{N}$,

$$v' \Theta_{\infty F}(\mathcal{W}, \mu) v \geq - \lim_{\epsilon \to 0^+} \ v' \left(\tilde{F} R_\epsilon^{-1} \tilde{F}' - \mu Q \right) v. \tag{29}$$

We conclude that $(\mathcal{W}, \mu) \notin \mathcal{C}_{\infty F}$ provided that μ is chosen sufficiently large in order that the right-hand side of (29) becomes positive for some $v \in \mathcal{N}$. An immediate consequence of this fact is that $(\mathcal{W}, \mu) \notin \mathcal{C}_{\infty F}$ for all $\mu > \overline{\mu}(F)$, where

$$\begin{aligned} \overline{\mu}(F) &= \lim_{\epsilon \to 0^+} \ \min \left\{ w' V_F R_\epsilon^{-1} V_F' w \ : \ w'w = 1 \right\} \\ &\geq \lim_{\epsilon \to 0^+} \ \sup \left\{ \mu \ : \ v' \left(\tilde{F} R_\epsilon^{-1} \tilde{F}' - \mu Q \right) v \geq 0 \ \forall \ v \in \mathcal{N} \right\}. \end{aligned} \tag{30}$$

Calling the full rank matrix $E = \begin{bmatrix} C_1 & D_1 \end{bmatrix}$ and $E^{\#} = E'(EE')^{-1}$ its pseudoinverse, it is a simple task to verify that

$$R_\epsilon^{-1} \longrightarrow E^{\#} E^{\#'} + \epsilon^{-1} \left(\mathbf{I} - E^{\#} E \right) \tag{31}$$

as $\epsilon \to 0^+$. Using this fact and (30) we get

$$\begin{aligned} \overline{\mu}(F) &= \min \left\{ w' V_F E^{\#} E^{\#'} V_F' w \ : \ (\mathbf{I} - E^{\#} E) V_F' w = 0 \ , \ w'w = 1 \right\} \\ &= \min \left\{ z'z \ : \ V_F' w - E'z = 0 \ , \ w'w = 1 \right\}, \end{aligned} \tag{32}$$

where the last inequality follows from the definition of the new variable z such that $E'z = V_F' w$. It is important to observe that (32) always provides $0 < \overline{\mu}(F) < +\infty$. Indeed, $w(k)$ being "rich," a $w \neq 0$ always exists such that $(\mathbf{I} - E^{\#}E)V_F' w = 0$, and on the other hand, with V_F of full rank, $V_F' w = 0$ implies $w = 0$ which is not feasible. Now, writing the optimality conditions to (32) we immediately conclude that $\mu(F)$ is given by (27), which proves the theorem. □

Defining the upper bound $\overline{\mu} = \min \{ \overline{\mu}(F_i) \ : \ i = 1, \ldots, N \}$, the convex programming problem

$$\max \left\{ \mu \ : \ (\mathcal{W}, \mu) \in \mathcal{C}_\infty, \ 0 \leq \mu \leq \overline{\mu} \right\} \tag{33}$$

is equivalent to (25) in the sense that they have the same global optimal solution. In fact, for $\mu > \overline{\mu}$, there exists an extreme matrix $F = F_i \in \mathcal{D}$ such that $\mu > \overline{\mu}(F)$. From Theorem 4.3, any pair $(\mathcal{W}, \mu) \notin \mathcal{C}_\infty$ whenever $\mu > \overline{\mu}$.

The above result is very useful for numerical computation. Actually the search interval on μ is reduced to the closed line segment $[0, \overline{\mu}]$. The upper bound $\overline{\mu}$ can be easily computed since no operations involving $\mathcal{H}_\infty$ norms calculation is needed. As a matter of fact, in many cases, $\overline{\mu}(F)$ can also be obtained directly from (30) without using equation (27), i.e.,

$$\begin{aligned} \overline{\mu}(F) &= \lim_{\epsilon \to 0^+} \min \left\{ w' V_F R_\epsilon^{-1} V_F' w \ : \ w'w = 1 \right\} \\ &= \lim_{\epsilon \to 0^+} \lambda_{\min} \left\{ V_F \left(R + \epsilon \mathbf{I} \right)^{-1} V_F' \right\}, \end{aligned} \tag{34}$$

overcoming the hypothesis concerning the richness of $w(k)$ introduced in Theorem 4.3. Note, however, that under the sufficient conditions in Theorem 4.3, the limit (34) always exists and is finite.

As a final remark concerning (P1) and (P2), we want to add that majorizations/approximations as introduced here have also been adopted in several papers [1], [14]. The main advantage of our approach is that we always obtain convex problems to be solved. As a consequence, the convergence of any applicable method is sure toward their global optimal solutions. For instance, this is not the case of the interesting procedure given in [1], for which, however, no proof of convergence is known.

5. The $\mathcal{H}_2/\mathcal{H}_\infty$ guaranteed-cost control problem.

In this section, we analyze the solution of problem (P3). The results are based on the convex combination (22) and on Theorem 3.5. Once again, it will be possible to approximate (P3) by a convex programming problem.

THEOREM 5.1. *Consider the convex set $\mathcal{C}_\infty$ as defined in Theorem 4.2 and assume that $\gamma > 0$ is given. Then*

$$\left\{ (W_2' W_1^{-1}, \beta) \ : \ (\mathcal{W}, \gamma^{-2}) \in \mathcal{C}_\infty \right\} \subseteq \bigcap_{F \in \mathcal{D}} \operatorname{epi} g_F, \tag{35}$$

where $\beta^2 = \gamma^2 \mathrm{Tr}\ (R\mathcal{W})$.

Proof. From Theorem 4.2 we note that $(\mathcal{W}, \gamma^{-2}) \in \mathcal{C}_\infty$ implies $(\mathcal{W}, \gamma^{-2}) \in \mathcal{C}_{\infty F}$ $\forall\ F \in \mathcal{D}$. Using Theorem 3.5 we have $\|H_F\|_2^2 \leq \gamma^2 \mathrm{Tr}\ (R\mathcal{W}) = \beta^2\ \forall\ F \in \mathcal{D}$. Consequently (35) holds. □

With Theorem 5.1, we are able to rewrite problem (P3) as a convex problem, that is,

$$\min \left\{ \gamma^2 \mathrm{Tr}\ (R\mathcal{W}) \ : \ (\mathcal{W}, \gamma^{-2}) \in \mathcal{C}_\infty \right\}. \tag{36}$$

For γ fixed, its objective function is linear. Furthermore, its global optimal solution provides $K = W_2'W_1^{-1}$, which is feasible to (P3) being consequently a suboptimal solution. Indeed, with $(\mathcal{W}, \gamma^{-2}) \in \mathcal{C}_\infty$, the associated state feedback gain given above is such that $(K, \gamma) \in \text{epi } h_F \; \forall \; F \in \mathcal{D}$ and $(K, \beta) \in \text{epi } g_F \; \forall \; F \in \mathcal{D}$. Clearly, the objective function of (36) is the available upper bound to $\|H_F\|_2^2$ for all $F \in \mathcal{D}$, which is reduced to its minimum level. An observation that has perhaps occurred to the reader is that the above conclusions hold only because the same nonlinear mapping $(W_2'W_1^{-1})$ generates elements on both epi g_F and epi h_F.

Finally, we have to notice that problems (33) and (36) are of the same kind. Actually, both are convex with linear objectives. The main difference between them is that the former presents one more variable than (36).

We are now in position to generalize our results along the lines of [14]. Suppose that to the basic model (1), we add a new output variable $y(k) = C_2x(k) + D_2u(k)$, where $C_2'D_2 = 0$ and $D_2'D_2 > 0$ and whose closed-loop transfer function from $w(k)$ to $y(k)$ is denoted by $T_F(\zeta)$. Our goal is to solve the $\mathcal{H}_2/\mathcal{H}_\infty$ guaranteed-cost control problem considering $\|T_F\|_2 \leq \beta$ subject to $\|H_F\|_\infty \leq \gamma \; \forall \; F \in \mathcal{D}$. To this end, defining the matrix

$$R_2 = \begin{bmatrix} C_2'C_2 & \mathbf{0} \\ \mathbf{0} & D_2'D_2 \end{bmatrix} \in \Re^{p\times p},$$

the functions $g_F(K) = \|T_F\|_2$ and $h_F(K) = \|H_F\|_\infty$, and using the same reasoning adopted before, we get

$$\min \left\{ \gamma^2 \text{Tr } (R_2\mathcal{W}) \; : \; (\mathcal{W}, \gamma^{-2}) \in \mathcal{C}_\infty \right\}. \tag{37}$$

This shows that the above generalization is easily incorporated in problem (P3) without loss of any of its geometric properties.

Another important generalization of the above problem was first proposed in [10] in the case of precisely known systems. In our context, it consists of the solution of (P3) with a decentralized control structure. Under the approximation and majorizations introduced before, it can be solved by simply noticing that if $\mathcal{W}$ is constrained as (see (12))

$$\mathcal{W} = \mathcal{W}_D = \begin{bmatrix} W_{1D} & W_{2D} \\ W_{2D}' & W_3 \end{bmatrix} \geq 0, \tag{38}$$

where the subscript D imposes a prespecified decentralized structure on each matrix, then the mapping which defines the control gain turns out to be

$$K = W_{2D}'W_{1D}^{-1} = K_D, \tag{39}$$

meaning that $K = K_D$ presents the same decentralized structure imposed on $\mathcal{W}$ in (38). Consequently, (P3) can be rewritten as

$$\min \left\{ \gamma^2 \text{Tr } (R_2\mathcal{W}_D) \; : \; (\mathcal{W}_D, \gamma^{-2}) \in \mathcal{C}_\infty \right\}. \tag{40}$$

Problem (40) is once again convex and easier to solve than (37). In fact, $\mathcal{W} = \mathcal{W}_D$ imposes a priori the zero value on many elements of $\mathcal{W}$ which clearly can be simply eliminate in the optimization process, reducing consequently the number of free variables to be determined. As before, the solution of (40) (if any) provides $\mathcal{W}_D$ which is used with (39) to get the decentralized control gain.

Before concluding this section, a final remark. In many practical applications, the "inverse" $\mathcal{H}_2/\mathcal{H}_\infty$ guaranteed-cost control problem may be more attractive. It consists of the determination of the smaller $\mathcal{H}_\infty$ norm level $\gamma > 0$ such that $\|H_F\|_\infty \leq \gamma$ subject to the $\mathcal{H}_2$ norm constraint $\|T_F\|_2 \leq \beta$, for some $\beta > 0$ fixed. Even in the case $F \in \mathcal{D}$ fixed, this problem cannot be solved by the procedures introduced in [1], [14]. On the contrary, in our context, it can be approximated by a convex programming problem. To show this, note that from Theorem 5.1 we have

$$\|T_F\|_2^2 \leq \mu^{-1}\text{Tr }(R_2\mathcal{W}) \quad \forall\ (\mathcal{W}, \mu) \in \mathcal{C}_\infty, \tag{41}$$

which implies that for $\beta > 0$ fixed, Tr $(R_2\mathcal{W}) - \beta^2\mu \leq 0$ imposes $\|T_F\|_2 \leq \beta$. Consequently, the problem stated before can be written as

$$\max \left\{\mu \ : \ \text{Tr }(R_2\mathcal{W}) - \beta^2\mu \leq 0 \ , \ (\mathcal{W}, \mu) \in \mathcal{C}_\infty \right\}, \tag{42}$$

which exhibits the same number of variables as and only one additional linear constraint than (P2). Once (42) is solved, the corresponding state feedback gain is given by $K = W_2'W_1^{-1}$, and as expected, (42) reduces to (25) whenever β is chosen sufficiently large.

6. Comparisons. We are able to compare our results with those of [24], where the necessary and sufficient counterpart of Lemma 3.2, part (a) has been proposed. For that we have to restrict our analysis only to the case of precisely known systems $(N = 1)$.

We immediately observe that (13) is a good approximation to the necessary and sufficient condition of [24] if its solution is such that

$$(\mathbf{I} + C_{\text{cl}}PC_{\text{cl}}')^{-1} \simeq \mathbf{I}. \tag{43}$$

For $\mathcal{W} \geq 0$ being the optimal solution of (25) or (36), condition (43) must be verified for $P = W_1$ and $C_{\text{cl}} = C_1 - D_1W_2'W_1^{-1}$. Using the fact that $W_3 \geq W_2'W_1^{-1}W_2$, (43) holds whenever

$$\lambda_{\max}\left(\left[\begin{array}{cc} C_1 & -D_1 \end{array}\right]\mathcal{W}\left[\begin{array}{c} C_1' \\ -D_1' \end{array}\right]\right) \ll 1, \tag{44}$$

which shows that the quality of our results can be verified by a simple postoptimization test. Furthermore, it is interesting to note (see the discussion after Lemma 3.2) that the trace of the matrix indicated in (44) equals Tr $(R\mathcal{W})$, yielding the conclusion that for (P3) the condition (44) generally holds because an upper bound of its left-hand side is minimized. To show that this fact actually occurs, we have calculated, for the simple example included in §3, the maximum eigenvalue indicated in (44). For the optimal solution of (25) it is equal to $\simeq 0.47$, while for the one of (36) it is equal to $\simeq 0.04$.

As a final and important remark we note that problems (25) and (36) always provide feasible solutions to (P2) and (P3), respectively, even when (44) is not verified. The suboptimality of their solutions is the price to be paid for the approximation of both (P2) and (P3) by convex programming problems. Unfortunately, if we replace (13) by the necessary and sufficient condition of [24], the same reasoning as used before does not lead to convex problems being thus an open problem for future research.

We now turn our attention to problem (36). The goal is to analyze its optimal solution as $\gamma \to +\infty$ in the particular case of precisely known systems. We observe that since $N = 1$, then $\mathcal{C}_\infty \equiv \mathcal{C}_{\infty F}$, and with $\mu = 1/\gamma^2$ and $\mathcal{W} \geq 0$ fixed we get

$$\lim_{\mu \to 0^+} \mu^{-1}\Theta_{\infty F}(\mu\mathcal{W}, \mu) = F\mathcal{W}F' - \mathcal{W} + Q$$

$$\triangleq \Theta_{2F}(\mathcal{W}). \tag{45}$$

On the other hand, calling $J(\mathcal{W}) = \mu^{-1}\mathrm{Tr}\ (R\mathcal{W})$ the objective function of (36), it is obvious that $J(\mu\mathcal{W}) = \mathrm{Tr}\ (R\mathcal{W})$. This fact together with (45) allows us to conclude that as $\gamma \to +\infty$, problem (36) degenerates to the convex problem

$$\min \left\{ \mathrm{Tr}\ (R\mathcal{W}) \ : \ \mathcal{W} \in \mathcal{C}_2 \right\}, \tag{46}$$

where $\mathcal{C}_2 = \{\mathcal{W} = \mathcal{W}' \geq 0 \ : \ v'\Theta_2(\mathcal{W})v \leq 0 \ \forall \ v \in \mathcal{N}\}$. It is possible to prove that the optimal solution of (46) coincides with the optimal solution of the linear quadratic regulator problem (see [18]). In other words, as $\gamma \to +\infty$, the optimal solution of the approximate problem (36) coincides with the optimal solution of (P3), which in this case is nothing more than the classical linear quadratic problem.

7. Numerical examples. In this section we consider a numerical example first proposed in [16] and also analyzed in a preliminary version of [24]. Using the notation introduced before, the data are as follows:

$$F = F_1 = \begin{bmatrix} 0.9974 & 0.0539 & -0.0013 \\ -0.1078 & 1.1591 & -0.0539 \\ 0 & 0 & 0 \end{bmatrix},$$

$$Q = \begin{bmatrix} 1 & 1 & 0 \\ 1 & 1 & 0 \\ 0 & 0 & 0 \end{bmatrix}, \quad R = \begin{bmatrix} 1 & -1 & 0 \\ -1 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}. \tag{47}$$

By reducing successively the γ parameter and testing for the existence of a positive definite solution to the necessary and sufficient conditions provided in [24], we get a near-optimum solution to the $\mathcal{H}_\infty$ optimal control problem as being $K^* \cong \left[\ 37.1242 \quad 22.4845\ \right]$, $\|H_F\|_\infty \cong \gamma^* \cong 36.44$ **dB**. The numerical difficulty involved increases as $\gamma \to \gamma^*$ because the condition to be tested becomes ill conditioned.

Considering $N = 1$ and $\bar{\mu} = 4.0 \times 10^{-4}$, problem (33) has been solved, yielding

$$\gamma = 1/\sqrt{\bar{\mu}} \cong 38.65 \ \mathbf{dB},$$

$$\mathcal{W} \cong \begin{bmatrix} 0.1375 & -0.1641 & -0.4410 \\ -0.1641 & 0.3954 & 2.5492 \\ -0.4410 & 2.5492 & 21.8803 \end{bmatrix} \times 10^{-2}. \tag{48}$$

The quality of this suboptimal solution (against the optimal one) can be measured a priori by simply verifying that the left-hand side of (44) gives $\simeq 0.2230$. Indeed (48) provides

$$K \cong \left[\ 8.8829 \quad 10.1327\ \right],$$

$$\|H_F\|_\infty \cong 37.63 \ \mathbf{dB}, \tag{49}$$

which corresponds to a loss of optimality of about 3.2%. It is important to remark that in solving problem (33) we did not observe the ill-conditioning of its solution as pointed out before.

We turn now our attention to the case of uncertain systems. Our goal is to obtain an approximate solution to (P2) for $N = 2$ and $\mathcal{D}$ given by (4), where F_1 is as before and

$$F_2 = \begin{bmatrix} 0.9974 & 0.0539 & -0.0013 \\ -0.1078 & 1.1591 & -0.1078 \\ 0 & 0 & 0 \end{bmatrix}, \tag{50}$$

which has been calculated assuming that the $(2,1)$ element of the input matrix B_2 may increase 100%. Once again, problem (33) has been solved with the same algorithm and the same upper bound $\overline{\mu}$ as before, yielding

$$\gamma = 1/\sqrt{\overline{\mu}} \cong 38.82 \text{ dB},$$

$$\mathcal{W} \cong \begin{bmatrix} 0.1231 & 0.1478 & 0.2768 \\ 0.1478 & 0.4608 & 3.0178 \\ 0.2768 & 3.0178 & 26.0642 \end{bmatrix} \times 10^{-2}, \tag{51}$$

$$K \cong \begin{bmatrix} 9.1290 & 9.4792 \end{bmatrix}.$$

Since every $F \in \mathcal{D}$ can be expressed as $F = F_1 + \xi(F_2 - F_1)$ for $\xi \in [0,1]$, Fig. 2 shows the "shadow" singular value diagram corresponding to the closed-loop uncertain system under consideration, that is, $\sigma_{\max}[H_F(e^{j\omega})]$ versus ω, parametrized on $\xi \in [0,1]$. Figure 2 also shows the closed-loop system $\mathcal{H}_\infty$ norm versus ξ with both the guaranteed $\mathcal{H}_\infty$ cost control gain (51) (—) and the optimal $\mathcal{H}_\infty$ state feedback gain K^* (– –) (which is optimal for $F = F_1$ or equivalently for $\xi = 0$). The following conclusions can be drawn.

(i) Although both feedback gains are quite different, comparing (51) with the optimal solution for $F = F_1$, we note that the guaranteed cost is very close to the optimal one. This occurs mainly because for each $\xi \in [0,1]$, the associate singular value diagram is kept near enough to the optimal one.

(ii) Figure 2 makes evident the importance of the suboptimal solution to the $\mathcal{H}_\infty$ guaranteed-cost control problem proposed here. The state feedback gain K is stabilizing and $\|H_F\|_\infty \leq \gamma \; \forall \; \xi \in [0,1]$. The same conclusion does not hold if instead of K we use the state feedback gain K^* which is $\mathcal{H}_\infty$-optimal for the "nominal" model F_1 (or equivalently $\xi = 0$). Indeed, with K^*, the closed-loop system becomes unstable for $\xi > 0.8$ and, worse, the $\mathcal{H}_\infty$ norm of the closed-loop transfer function becomes greater than the one with K for $\xi > 0.05$. This shows that for very small parameter variations (about 5%) the solution of (33) is already better than K^*.

8. Conclusions. In this paper, we have investigated three important problems arising in $\mathcal{H}_\infty$ and mixed $\mathcal{H}_2/\mathcal{H}_\infty$ control design. The results follow from two basic facts. The first one concerns the proposition of a simple sufficient condition for testing whenever the $\mathcal{H}_\infty$ norm of a certain closed-loop transfer function is smaller than a prespecified level γ. The second one consists of the definition of a suitable majorization to the $\mathcal{H}_2$ norm of the same closed-loop transfer function. Based only on that, the three problems considered are approximated by convex problems all with linear objective functions. Obviously an immediate consequence is that all machinery of convex programming can be used to solve them in order to find directly the

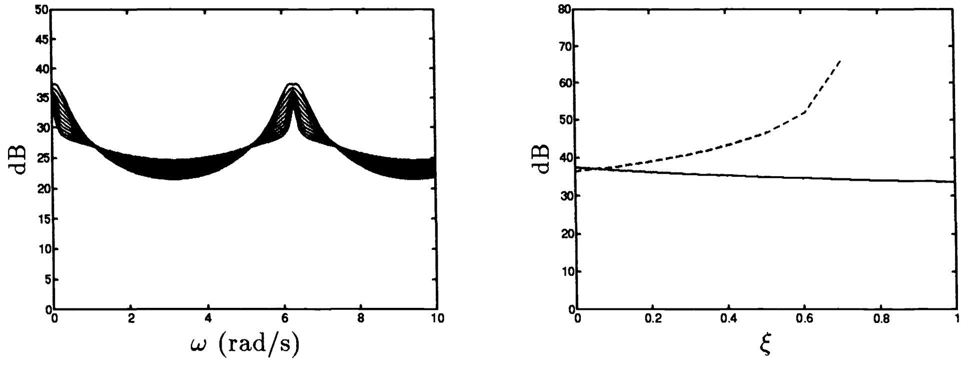


FIG. 2. *Singular value diagram and $\mathcal{H}_\infty$-norm level.*

corresponding optimal state feedback gain matrix. Furthermore, it is important to emphasize that the convex problems are defined in a special augmented parameter space whose elements generate through a nonlinear mapping the set of all stabilizing state feedback gains of a given linear discrete-time system.

In the real world all systems are subject to parameter uncertainties. Then the results presented here are important because, for the first time, they generalize to actual systems many basic and well-known control design tools available in the literature to date.

In this sense, based on the numerical solution of some examples we claim that our suboptimal and easy-to-determine solution to the three problems proposed before may constitute a useful and valid tool design when convex-bounded uncertainties have to be considered. The suboptimality is largely compensated by taking into account the uncertainties, and in many cases, the results presented here can be even better than the ones obtained by using an optimal policy calculated for the "nominal" model.

Appendix A. Consider $f(\cdot) \; : \; S \longrightarrow \Re$ a real-valued function defined for all $x \in S \subset \Re^n$. Following [20] we define the epigraph of $f(\cdot)$ as

$$\text{epi}\, f \triangleq \Big\{(x,\gamma) \; : \; x \in S \; , \; f(x) \leq \gamma\Big\} \subseteq \Re^{n+1}. \tag{52}$$

An important property relating a function and its epigraph is that $f(\cdot)$ is convex if and only if epi f is convex. In our context, the function $h_F(K) \; : \; \mathcal{K}_F \longrightarrow \Re$ given by

$$h_F(K) = \|(C_1 - D_1 K)\left(\zeta \mathbf{I} - (A - B_2 K)\right)^{-1} B_1\|_\infty \tag{53}$$

is well defined for all $K \in \mathcal{K}_F$ since as required in (53) any element of $\mathcal{K}_F$ stabilizes the closed-loop system. Consequently

$$\text{epi}\, h_F \triangleq \Big\{(K,\gamma) \; : \; K \in \mathcal{K}_F \; , \; \|H_F\|_\infty \leq \gamma\Big\}. \tag{54}$$

Unfortunately epi h_F is not a convex set because generally the set $\mathcal{K}_F$ is not convex. The set epi g_F is similarly defined.

Proof of Lemma 3.1. Given (K,γ), assume that $P = P' > 0$ satisfies the matrix inequality (13). In this case it follows that

$$A_{\text{cl}} P A_{\text{cl}}' - P \;\leq\; -\gamma^{-2} B_1 B_1' \;\leq\; 0. \tag{55}$$

However, we also assume that the pair (A, B_1) is controllable and range $(B_2) \subseteq$ range (B_1). This implies that $(A_{\rm cl}, B_1)$ is also controllable, which together with (55) enables us to say that $A_{\rm cl}$ is asymptotically stable. Consequently $K \in \mathcal{K}_F$. To prove the inequality (14) we first note that for $\zeta = e^{j\omega}$, $\omega \in [-\pi, \pi]$, we have

$$(\zeta\mathbf{I} - A_{\rm cl})P(\zeta^{-1}\mathbf{I} - A'_{\rm cl}) + A_{\rm cl}P(\zeta^{-1}\mathbf{I} - A'_{\rm cl}) + (\zeta\mathbf{I} - A_{\rm cl})PA'_{\rm cl} = P - A_{\rm cl}PA'_{\rm cl}. \tag{56}$$

So inequality (13) can be rewritten as

$$\begin{aligned} &-(\zeta\mathbf{I} - A_{\rm cl})P(\zeta^{-1}\mathbf{I} - A'_{\rm cl}) - A_{\rm cl}P(\zeta^{-1}\mathbf{I} - A'_{\rm cl}) - (\zeta\mathbf{I} - A_{\rm cl})PA'_{\rm cl} \\ &+ PC'_{\rm cl}C_{\rm cl}P + \gamma^{-2}B_1B_1' \leq 0, \end{aligned} \tag{57}$$

which after some algebraic manipulations gives

$$\begin{aligned} &C_{\rm cl}PC'_{\rm cl} - \zeta C_{\rm cl}(\zeta\mathbf{I} - A_{\rm cl})^{-1}PC'_{\rm cl} - \zeta^{-1}C_{\rm cl}P(\zeta^{-1}\mathbf{I} - A'_{\rm cl})^{-1}C'_{\rm cl} \\ &+ C_{\rm cl}(\zeta\mathbf{I} - A_{\rm cl})^{-1}PC'_{\rm cl}C_{\rm cl}P(\zeta^{-1}\mathbf{I} - A'_{\rm cl})^{-1}C'_{\rm cl} \\ &+ \gamma^{-2}C_{\rm cl}(\zeta\mathbf{I} - A_{\rm cl})^{-1}B_1B_1'(\zeta^{-1}\mathbf{I} - A'_{\rm cl})^{-1}C'_{\rm cl} \leq 0. \end{aligned} \tag{58}$$

Keeping in mind the definition of $H_F(\zeta)$ (see (5)) and defining the auxiliary transfer function $L_F(\zeta) = \zeta C_{\rm cl}(\zeta\mathbf{I} - A_{\rm cl})^{-1}PC'_{\rm cl}$, from (58) we have

$$C_{\rm cl}PC'_{\rm cl} - L_F(\zeta) - L_F(\zeta^{-1})' + L_F(\zeta)L_F(\zeta^{-1})' + \gamma^{-2}H_F(\zeta)H_F(\zeta^{-1})' \leq 0, \tag{59}$$

which after completing squares becomes

$$\begin{aligned} H_F(\zeta)H_F(\zeta^{-1})' &\leq \gamma^2\mathbf{I} - \gamma^2C_{\rm cl}PC'_{\rm cl} - \gamma^2[\mathbf{I} - L_F(\zeta)][\mathbf{I} - L_F(\zeta^{-1})]' \\ &\leq \gamma^2(\mathbf{I} - C_{\rm cl}PC'_{\rm cl}). \end{aligned} \tag{60}$$

Since (60) holds for all $\zeta = e^{j\omega}$, $\omega \in [-\pi, \pi]$ the lemma is proved. □

Proof of Lemma 3.2. From the definition of the $\mathcal{H}_\infty$ norm, there exist $\omega \in [-\pi, \pi]$ and a vector z ($z^\sim$ denotes its complex conjugate) such that

$$\begin{aligned} \|H_F\|_\infty^2 &= \frac{z^\sim H_F(e^{j\omega})H_F(e^{-j\omega})'z}{z^\sim z} \\ &\leq \gamma^2 - \gamma^2\frac{z^\sim C_{\rm cl}PC'_{\rm cl}z}{z^\sim z} \\ &\leq \gamma^2(1 - \theta^2), \end{aligned} \tag{61}$$

where the second inequality follows from Lemma 3.1 and

$$\begin{aligned} \theta^2 &= \frac{z^\sim C_{\rm cl}PC'_{\rm cl}z}{z^\sim z} \\ &\leq \lambda_{\max}(C_{\rm cl}PC'_{\rm cl}) \ \leq \ \mathrm{Tr}\ (C_{\rm cl}PC'_{\rm cl}). \end{aligned} \tag{62}$$

It is clear that the exact value of θ^2 cannot be calculated only in terms of the inequality (13) because it depends on the value of $\|H_F\|_\infty$. The above inequality provides an upper bound for it, thus proving part (a).

For part (b) we note once again that if $P = P' > 0$ satisfies (13), then it also satisfies (55). As a result, L_c being the controllability Gramian defined in (8) we get

$$\begin{aligned} P &\geq \gamma^{-2} \sum_{k=0}^{\infty} A_{\mathrm{cl}}^k B_1 B_1' {A_{\mathrm{cl}}'}^k \\ &\geq \gamma^{-2} L_c. \end{aligned} \tag{63}$$

This inequality together with (7) yields

$$\begin{aligned} \|H_F\|_2^2 &= \mathrm{Tr}\ (C_{\mathrm{cl}} L_c C_{\mathrm{cl}}') \\ &\leq \gamma^2 \mathrm{Tr}\ (C_{\mathrm{cl}} P C_{\mathrm{cl}}'), \end{aligned} \tag{64}$$

which concludes the proof of Lemma 3.2. □

REFERENCES

[1] D. S. Bernstein and W. M. Haddad, *LQG control with an $\mathcal{H}_\infty$ performance bound: A Riccati equation approach*, IEEE Trans. Automat. Control, 34 (1989), pp. 293–305.

[2] J. Bernussou, P. L. D. Peres, and J. C. Geromel, *A linear programming oriented procedure for quadratic stabilization of uncertain systems*, Systems Control Lett., 13 (1989), pp. 65–72.

[3] S. Boyd and Q. Yang, *Structured and simultaneous Lyapunov functions for system stability problems*, Internat. J. Control, 49 (1989), pp. 2215–2240.

[4] J. C. Doyle, K. Glover, P. P. Khargonekar, and B. A. Francis, *State-space solutions to standard $\mathcal{H}_2$ and $\mathcal{H}_\infty$ control problems*, IEEE Trans. Automat. Control, 34 (1989), pp. 831–847.

[5] K. Furuta and S. Phoojaruenchanachai, *An algebraic approach to discrete-time $\mathcal{H}_\infty$ control problems*, in Proc. 1990 American Control Conference, San Diego, May 1990, pp. 3067–3072.

[6] J. C. Geromel, P. L. D. Peres, and J. Bernussou, *On a convex parameter space method for linear control design of uncertain systems*, SIAM J. Control Optim., 29 (1991), pp. 381–402.

[7] J. C. Geromel, P. L. D. Peres, and S. R. Souza, *Quadratic stabilizability of linear uncertain systems with prescribed $\mathcal{H}_\infty$ norm bounds*, in Proc. 1st IFAC Symp. on Design Meth. Control Systems, Zurich, Switzerland, September 1991, pp. 302–307.

[8] ———, *$\mathcal{H}_\infty$ control of discrete-time uncertain systems*, IEEE Trans. Automat. Control, 39 (1994), pp. 1072–1075.

[9] K. Gu, Y. H. Chen, M. A. Zohdy, and N. K. Loh, *Quadratic stabilizability of uncertain systems: a two level optimization setup*, Automatica, 27 (1991), pp. 161–165.

[10] W. M. Haddad, D. S. Bernstein, and C. N. Nett, *Decentralized $\mathcal{H}_2/\mathcal{H}_\infty$ controller design: The discrete-time case*, in Proc. 28th Conference on Decision and Control, Tampa, December 1989, pp. 932–933.

[11] W. M. Haddad, D. S. Bernstein, and D. Mustafa, *Mixed-norm $\mathcal{H}_2/\mathcal{H}_\infty$ regulation and estimation: The discrete-time case*, Systems Control Lett., 16 (1991), pp. 235–247.

[12] P. A. Iglesias and K. Glover, *State space approach to discrete-time $\mathcal{H}_\infty$ control*, Internat. J. Control, 54 (1991), pp. 1031–1073.

[13] P. P. Khargonekar, I. R. Petersen, and K. Zhou, *Robust stabilization of uncertain linear systems: Quadratic stabilizability and $\mathcal{H}_\infty$ control theory*, IEEE Trans. Automat. Control, 35 (1990), pp. 356–361.

[14] P. P. Khargonekar and M. A. Rotea, *Mixed $\mathcal{H}_2/\mathcal{H}_\infty$ control: A convex optimization approach*, IEEE Trans. Automat. Control, 36 (1991), pp. 824–837.

[15] I. KAMINER, P. P. KHARGONEKAR, AND M. A. ROTEA, *Mixed $\mathcal{H}_2/\mathcal{H}_\infty$ control for discrete-time systems via convex optimization*, in Proc. 1992 American Control Conference, Chicago, June 1992, pp. 1363–1367.

[16] E. D. KIRK, *Optimal Control Theory, An Introduction*, Prentice-Hall, Englewood Cliffs, NJ, 1970.

[17] D. J. N. LIMEBEER, M. GREEN, AND D. WALKER, *Discrete time $\mathcal{H}_\infty$ control*, in Proc. 28th Conference on Decision and Control, Tampa, December 1989, pp. 392–396.

[18] P. L. D. PERES AND J. C. GEROMEL, *An alternate numerical solution to the linear quadratic problem*, IEEE Trans. Automat. Control, 39 (1994), pp. 198–202.

[19] P. L. D. PERES, J. C. GEROMEL, AND S. R. SOUZA, *Convex analysis of discrete-time uncertain $\mathcal{H}_\infty$ control problems*, in Proc. 30th Conference on Decision and Control, Brighton, UK, December 1991, pp. 521–526.

[20] R. T. ROCKAFELLAR, *Convex Analysis*, Princeton University Press, Princeton, NJ, 1970.

[21] C. SCHERER, *$\mathcal{H}_\infty$-control by state-feedback and fast algorithms for the computation of optimal $\mathcal{H}_\infty$-norms*, IEEE Trans. on Automat. Control, 35 (1990), pp. 1090–1099.

[22] C. E. DE SOUZA, M. FU, AND L. XIE, *$\mathcal{H}_\infty$ analysis and synthesis of discrete-time systems with time-varying uncertainty*, IEEE Trans. on Automat. Control, 38 (1993), pp. 459–462.

[23] R. J. VEILLETTE, J. V. MEDANIC, AND W. R. PERKINS, *Robust stabilization and disturbance rejection for systems with structured uncertainty*, in Proc. 28th Conference on Decision and Control, Tampa, December 1989, pp. 936–941.

[24] I. YAESH AND U. SHAKED, *A transfer function approach to the problems of discrete-time systems: $\mathcal{H}_\infty$ optimal linear control and filtering*, IEEE Trans. on Automat. Control, 36 (1991), pp. 1264–1271.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 33, No. 6, pp. 1834–1846, November 1995

© 1995 Society for Industrial and Applied Mathematics
011

A FINITE-DIMENSIONAL RISK-SENSITIVE CONTROL PROBLEM*

ALAIN BENSOUSSAN† AND ROBERT J. ELLIOTT‡

Abstract. A partially observed stochastic control problem with exponential running cost is considered. The dynamics are linear and the running cost is quadratic, although the control may enter nonlinearly. Explicit solutions are found to a modified Zakai equation and a backward adjoint equation. This enables the problem to be expressed in terms of observable finite-dimensional dynamics and a separation principle to be applied.

Key words. risk-sensitive partially observed stochastic control

AMS subject classifications. 93E20, 93E03, 93E11

1. Introduction.

There are few partially observed stochastic control problems for which the optimal control can be given in terms of finite-dimensional sufficient statistics. The linear quadratic Gaussian case is one. In [12] Wonham extended this result to a nonquadratic cost function. In recent years there has been interest in the case of linear dynamics for the state and observation and a cost which is the exponential of a standard quadratic cost. This model can be used to study risk aversion, or preference, in terms of a real parameter θ, and it was first discussed by Jacobson [6]. Whittle solved the linear discrete time partially observed problem in [11]. In continuous time the existence of finite-dimensional sufficient statistics was studied in particular cases by Jacobson [6], Speyer [10], and Kumar and van Schuppen [8] and was finally solved by Bensoussan and van Schuppen [2]. A detailed treatment can be found in Bensoussan [1].

In this paper, motivated by Wonham's contributions, we consider dynamics and running cost which are, respectively, linear and exponential quadratic in the state variables but which may be nonlinear functionals of the control. The terminal cost is also a general measurable function. An explicit solution of a modified Zakai equation is found, and the problem can then be described in terms of an information state defined by finite-dimensional fully observed dynamics. The value function, in turn, is a function of these finite-dimensional parameters and so is not a finite-dimensional quantity.

In §1 a general partially observed stochastic control problem with an exponential running cost is discussed. A modified Zakai equation is introduced whose solution is a measure related not only to the state of the process but also to the exponential running cost. A robust form of this equation is defined, and an adjoint measure valued process introduced. Section 3 discusses a verification result for a fully observed stochastic control problem. The case when the dynamics are linear and the exponential running cost is quadratic is considered in §4, although the control parameter may enter both dynamics and cost nonlinearly. It is shown that in this situation there is

* Received by the editors September 20, 1993; accepted for publication (in revised form) June 17, 1994. This research was performed during a visit to Institut National de Recherche en Informatique et en Automatique in May 1993 and is supported by Natural Sciences and Engineering Research Council of Canada grant A7964.

† Institut National de Rechereche en Informatique et en Automatique, Rocquencourt, 78153 Le Chesnay Cedex, France.

‡ Department of Mathematical Sciences, University of Alberta, Edmonton, Alberta T6G 2G1, Canada.

an explicit solution of the modified Zakai equation in terms of a modified mean r and variance Π of the state process and also an explicit solution for the backward, adjoint process. Results of §4 extend those in [1] and [2] and permit the partially observed stochastic control problem, with exponential cost, to be written as a completely observed stochastic control problem with finite-dimensional state variables r and Π. If the verification theorem of §3 can be applied, we can conclude the separation principle holds.

2. Dynamics. Suppose $(\Omega, \mathcal{A}, P)$ is a probability space with a complete filtration $\{\mathcal{F}_t\}$, $t \geq 0$, on which are given two independent $\mathcal{F}_t$-Wiener processes $w(t)$ and $z(t)$. We assume that

- $w(\cdot)$ takes values in R^n and has covariance matrix $Q(\cdot)$;
- $z(\cdot)$ takes values in R^m and has covariance matrix $R(\cdot)$;
- $R(\cdot)$ is uniformly positive definite.

ξ is an R^n-valued random variable which is $\mathcal{F}_0$ measurable and independent of $w(\cdot)$ and $z(\cdot)$. The distribution of ξ is the measure $\Pi_0(\cdot)$.

U is a nonempty subset of R^k. Consider the Borel functions

$$g : R^n \times U \times [0, \infty) \to R^n,$$
$$\sigma : R^n \times [0, \infty) \to L(R^n, R^n),$$

where

$$|g(x_1, v, t) - g(x_2, v, t)| \leq k|x_1 - x_2|$$

and

$$\|\sigma(x_1, t) - \sigma(x_2, t)\| \leq k|x_1 - x_2|.$$

$Z_t = \sigma\{z_s : s \leq t\}$ is the complete filtration generated by z and an admissible control is a Z_t-adapted process which takes values in U. For any admissible control v_t there is a strong solution $x^v_\cdot = x_\cdot$ of the stochastic differential equation

$$\begin{aligned} dx_t &= g(x_t, v_t)dt + \sigma(x_t)dw_t, \\ x_0 &= \xi. \end{aligned} \tag{2.1}$$

Equation (2.1) describes the dynamics of the state process. Note that here, and in what follows, to simplify notation we omit t when writing g, σ, etc.

Consider the Borel function $h : R^n \times [0, \infty) \to R^m$, where we suppose

$$|h(x, t)| \leq k(1 + |x|).$$

For any admissible control v, with $x_\cdot$ the corresponding solution of (2.1), we suppose

$$E\Big[\exp\Big(\frac{1}{2}\int_0^t h^*(x_s, s)R_s^{-1}h(x_s, s)ds\Big)\Big] < \infty.$$

Define

$$\Lambda_{0,t} = \exp\Big(\int_0^t h^*(x_s, s)R_s^{-1}dz_s - \frac{1}{2}\int_0^t h^*(x_s, s)R_s^{-1}h(x_s, s)ds\Big).$$

Then from Novikov's result, (see Thm. 13.27 of [3]), Λ is an $\mathcal{F}_t$ martingale and

$$E[\Lambda_{0,t}] = 1.$$

A new probability measure $\widehat{P} = \widehat{P}(v)$ can be defined if we put

$$\frac{d\widehat{P}}{dP}\Big|_{\mathcal{F}_t} = \Lambda_{0,t}.$$

Define the process $b_t = b_t^v$ by the formula $b_t = z_t - \int_0^t h(x_s, s)ds$. Then $b.$ is a Wiener process under $\widehat{P}$ with covariance matrix $R(\cdot)$. Therefore, under $\widehat{P}$

$$z_t = \int_0^t h(x_s, s)ds + b_t. \tag{2.2}$$

Note it is under measure $\widehat{P}$ that z, satisfying (2.2), describes the noisy observations of the state process. Note also that $\Lambda_{0,t}$ is given by

$$\begin{aligned} d\Lambda_{0,t} &= \Lambda_{0,t}h^*(x_t, t)R_t^{-1}dz_t, \\ \Lambda_{0,0} &= 1. \end{aligned} \tag{2.3}$$

Consider Borel functions

$$\begin{aligned} L &: R^n \times U \times [0, \infty) \to R, \\ \Phi &: R^n \to R. \end{aligned}$$

For any admissible control v and real number θ we consider the expected exponential risk-sensitive cost

$$\begin{aligned} J(v) &= \theta \widehat{E}\Big[\exp \theta \Big(\int_0^T L(x_s, v_s)ds + \Phi(x_T)\Big)\Big] \\ &= \theta E\Big[\Lambda_{0,T} \exp \theta \Big(\int_0^T L(x_s, v_s)ds + \Phi(x_T)\Big)\Big]. \end{aligned}$$

Write

$$D_{0,t} = D_t = \exp\Big(\theta \int_0^t L(x_s, v_s)ds\Big)$$

so

$$dD_t = \theta L(x_t, v_t)D_t dt, \tag{2.4}$$

with $D_0 = 1$.

Following [7] we now define an information state σ.

Notation 2.1. For any function $\phi : R^n \to R$ for which the expectation is defined, write

$$\sigma(\phi)_t = E[\Lambda_{0,t}D_t\phi(x_t)|Z_t].$$

In case the measure defined by $\sigma(\cdot)_t$ has a density $q(x, t)$ we have

$$\sigma(\phi)_t = \int_{R^n} \phi(x)q(x, t)dx.$$

Write

$$A = \sum_i g_i(x_t, v_t)\frac{\partial}{\partial x_i} + \sum_{i,j} a_{ij}(x_t)\frac{\partial^2}{\partial x_i \partial x_j},$$

where $a(x_t) = \frac{1}{2}\sigma(x_t)Q_t\sigma^*(x_t)$. We now obtain a modified Zakai equation for $\sigma(\phi)_t$.

PROPOSITION 2.2. *Suppose $\phi : R^n \to R$ is a C^2 function with compact support. Then*

$$\sigma(\phi)_t = \sigma(\phi_0) + \theta \int_0^t \sigma(L\phi)_r dr + \int_0^t \sigma(\phi h^*(x_r, r) R_r^{-1}) dz_r + \int_0^t \sigma(A\phi)_r dr,$$

where $\sigma(\phi)_0 = E[\phi(x_0)] = \int \phi(\xi)\Pi_0(d\xi)$.

Proof. The Itô rule implies

$$\phi(x_t) = \phi(x_0) + \int_0^t (A\phi)(x_r)dr + \int_0^t D\phi \cdot \sigma(x_r)dw_r, \tag{2.5}$$

where $D = \left(\frac{\partial}{\partial x_1}, \ldots, \frac{\partial}{\partial x_n}\right)$. From (2.3), (2.4), and (2.5) we have

(2.6)

$$\Lambda_{0,t}D_t\phi(x_t) = \phi(x_0) + \int_0^t \Lambda_{0,r}D_r\phi(x_r)h^*(x_r, r)R_r^{-1}dz_r + \theta \int_0^t \Lambda_{0,r}D_rL_r\phi(x_r)dr$$
$$+ \int_0^t \Lambda_{0,r}D_r(A\phi)(x_r)dr + \int_0^t \Lambda_{0,r}D_r(D\phi) \cdot \sigma(x_r)dw_r.$$

We now condition each side of (2.6) on Z_t and use the facts that $w(\cdot)$ and $z(\cdot)$ are independent and that $z(\cdot)$ has independent increments under P (it is Wiener). (See Lemma 3.2 of Chapter 7 of Hajek and Wong [5].) We thus obtain

$$\sigma(\phi)_t = \sigma(\phi)_0 + \int_0^t \sigma(L\phi)_r dr + \int_0^t \sigma(\phi h^*(x_r, r)R_r^{-1})dz_r + \int_0^t \sigma(A\phi)_r dr. \tag{2.7}$$

COROLLARY 2.3. *If $\sigma(\cdot)_t$ has a density $q(x,t)$, integrating each term of (2.7) by parts implies that q satisfies*

(2.8)

$$q(x,t) = q_0(x) + \int_0^t q(x,r)h^*(x,r)R_r^{-1}dz_r + \theta \int_0^t L(x,r)q(x,r)dr + \int_0^t (A^*q)(x,r)dr.$$

Here A^* is the adjoint of A, and $q_0(x)$ is the density such that

$$\Pi_0(dx) = q_0(x)dx.$$

Remark 2.4. Analogs of (2.8) appear in [2] and [7]. For any admissible control v we have defined

$$\begin{aligned} J(v) &= \theta E[\Lambda_{0,T}D_T \exp(\theta\Phi(x_T))] \\ &= \theta E\big[E[\Lambda_{0,T}D_T \exp(\theta\Phi(x_T)) \mid Z_T]\big] \\ &= \theta E\Big[\int_{R^n} \exp(\theta\Phi(x)) \cdot q(x,T)dx\Big]. \end{aligned}$$

With $\Lambda_{t,T}$ and $D_{t,T}$ defined in the obvious way and with $\mathcal{X}_t = \sigma\{x_s : s \le t\}$ the filtration generated by $x(\cdot)$,

$$\begin{aligned} E[\Lambda_{0,T}D_T \exp(\theta\Phi(x_T)) \mid Z_T] &= E\big[E[\Lambda_{0,t}\Lambda_{t,T}D_{0,t}D_{t,T} \exp(\theta\Phi(x_T)) \mid Z_T, \mathcal{X}_t]Z_T\big] \\ &= E\big[\Lambda_{0,t}D_{0,t}E[\Lambda_{t,T}D_{t,T} \exp(\theta\Phi(x_T)) \mid Z_T, x_t]Z_T\big], \end{aligned}$$

using the Markov property. Write

$$p(x,t) = E[\Lambda_{t,T}D_{t,T} \exp(\theta\Phi(x_T)) \mid Z_T, x_t = x].$$

Then

$$\begin{aligned}E[\Lambda_{0,T}D_T\exp(\theta\Phi(x_T)) \mid Z_T] &= E[\Lambda_{0,t}D_tp(x_t,t) \mid Z_T]\\ &= \int_{R^n} q(x,t)p(x,t)dx.\end{aligned}$$

Clearly this quantity is independent of t. The expression is similar to results of Pardoux [9] on filtering and smoothing. Pardoux obtains the forward Zakai equation for q and an adjoint backward stochastic partial differential equation for the analog of p. However, in the control case which we are considering, we cannot immediately find the backward stochastic partial differential equation for p because, although the Wiener process $z(\cdot)$ could be reversed, the coefficients and running cost $L(x_r, v_r)$ involve the control $v(\cdot)$ which is adapted to the forward filtration. Consequently, we consider the robust form of the equation for q.

That is, we introduce the factor

$$\Psi_t(x) := \exp(-h^*(x,t)R_t^{-1}z_t).$$

We suppose the derivative

$$\frac{d}{dt}(h^*(x,t)R_t^{-1}) = (h^*(x,t)R_t^{-1})'$$

exists so that

$$d\Psi = \Psi_t(-h^*(x,t)R_t^{-1}dz_t - (h^*(x,t)R_t^{-1})'z_tdt + \frac{1}{2}(h^*(x,t)R_t^{-1}h(x,t))dt). \tag{2.9}$$

Define $\hat{q}(x,t) := q(x,t)\Psi_t(x)$. Then $d\hat{q} = dq\cdot\Psi + q\cdot d\Psi + \langle dq, d\Psi\rangle$, and from (2.8) this is

$$= \theta q\Psi.Ldt + \Psi.(A^*q)dt - q\Psi.(h^*R^{-1})'z_tdt - \frac{1}{2}q\Psi.(h^*R^{-1}h)dt. \tag{2.10}$$

Note here that the stochastic integral, dz-term has cancelled. The $\Psi.(A^*q)$ coefficient on the right side can be written as $\Psi.(A^*(\Psi^{-1}\hat{q}))$ so that $\hat{q}(x,t)$ is defined by a forward parabolic partial differential equation in which the observation process z appears as a parameter. That is,

$$\begin{aligned}\frac{\partial\hat{q}}{\partial t} &= \theta\hat{q}Ldt + \Psi.(A^*(\Psi^{-1}\hat{q}))dt - \hat{q}.(h^*R^{-1})'z_tdt - \frac{1}{2}\hat{q}.(h^*R^{-1}h)dt\\ &= B(\hat{q}),\ \text{say}, \qquad \hat{q}_0(x) = q_0(x),\end{aligned}$$

where B is a second-order operator. We can then consider a backward adjoint equation for a function $\hat{p}(x,t)$:

$$-\frac{\partial\hat{p}}{\partial t} = B^*(\hat{p}), \qquad \hat{p}(x,T) = \exp(\Phi(x) + h^*(x,T)R_T^{-1}z_T).$$

Defining $p(x,t) = \hat{p}(x,t)\Psi_t^{-1}$ we have

$$\int_{R^n} p(x,t)q(x,t)dx = \int_{R^n} \hat{p}(x,t)\hat{q}(x,t)dx,$$

a quantity independent of t.

We can, therefore, write

$$\begin{aligned} J(v) &= E\Big[\int_{R^n} p(x,t)q(x,t)dx\Big] = E\Big[\int_{R^n} \hat{p}(x,t)\hat{q}(x,t)dx\Big] \\ &= E\Big[\int_{R^n} \exp(\theta\Phi(x))\cdot q(x,T)dx\Big] = E\Big[\int_{R^n} p(x,0)q_0(x)dx\Big]. \end{aligned}$$

For any times $0 \leq s < t \leq T$ write $\mathcal{U}_{s,t}$ for the admissible controls on the interval $[s,t]$, that is, the functions $v : [s,t] \times \Omega \to U$ which are adapted to $\{Z_r\}$. The minimum cost from time 0 is of course

$$V(q_0,0) = \inf_{v\in\mathcal{U}_{0,T}} J(v).$$

The density $q(x,t)$, or alternatively $\hat{q}(x,t)$, can be taken as the observable state of the system at any time $t \in [0,T]$. In general, therefore, this state is infinite dimensional. Consider an intermediate time t and suppose the system is in state $q = q(x,t)$ at time t. For $s > t$ and any $v \in \mathcal{U}_{t,T}$, using (2.8), write $q_q^v(x,s) = q_q^v$ for the solution of

$$\text{(2.11)}\quad q_q^v := q + \int_t^s q(x,r)h^*(x,r)R_r^{-1}dz_r + \theta\int_t^s L(x,v_r)q(x,r)dr + \int_t^s (A^*q)(x,r)dr.$$

Then the cost for this process, using control v and starting in state q at time t, is

$$\begin{aligned} J(v,q,t) &= E\Big[\int_{R^n} \exp(\theta\Phi(x))q_q^v(x,T)dx\Big] \\ &= E\Big[\int_{R_n} p^v(x,t)q(x,t)dx\Big]. \end{aligned}$$

The minimum cost starting from state q at time t is, therefore,

$$V(q,t) = \inf_{v\in\mathcal{U}_{t,T}} J(v,q,t).$$

Consider $h > 0$ such that $0 \leq t < t+h \leq T$. We have the following dynamic programming result.

THEOREM 2.5.

$$\text{(2.12)}\qquad V(q,t) = \inf_{u\in\mathcal{U}_{t,t+h}} E\big[V(q_q^u(\cdot,t+h),t+h) \mid q(\cdot,t) = q\big].$$

Proof.

$$\begin{aligned} V(q,t) &= \inf_{v\in\mathcal{U}_{t,T}} J(v,q,t) = \inf_{v\in\mathcal{U}_{t,T}} E\Big[\int_{R^n} p^v(x,t)q(x,t)dx \mid q(\cdot,t) = q\Big] \\ &= \inf_{u\in\mathcal{U}_{t,t+h}} \inf_{v\in\mathcal{U}_{t+h,T}} E\Big[\int_{R^n} p^v(x,t+h)q_q^u(x,t+h)dx \mid q(\cdot,t) = q\Big] \\ &= \inf_{u\in\mathcal{U}_{t,t+h}} \inf_{v\in\mathcal{U}_{t+h,T}} E\Big[E\Big[\int_{R^n} p^v(x,t+h)q_q^u(x,t+h)dx \mid Z_{t+h}, q(\cdot,t)\Big]q(\cdot,t) = q\Big] \\ &= \inf_{u\in\mathcal{U}_{t,t+h}} E\Big[\inf_{v\in\mathcal{U}_{t+h,T}} E\Big[\int_{R^n} p^v(x,t+h)q_q^u(x,t+h)dx \mid q_q^u(\cdot,t+h)\Big]q(\cdot,t) = q\Big] \\ &= \inf_{u\in\mathcal{U}_{t,t+h}} E[V(q_q^u(x,t+h),t+h) \mid q(\cdot,t) = q]. \end{aligned}$$

The interchange of minimization and conditional expectation is justified because of the lattice property of the set of controls. See Elliott [3, Chap. 16].

Remark 2.6. There is a similar result in terms of the robust state $\hat{q}(\cdot, t)$:

$$V(\hat{q}, t) = \inf_{u \in \mathcal{U}_{t,t+h}} E[V(\hat{q}^u_{\hat{q}}(\cdot, t+h), t+h) \mid \hat{q}(\cdot, t) = \hat{q}].$$

Here $\hat{q}$ evolves according to (2.10). We have noted that the state q, a density, is in general infinite-dimensional. Consequently, it is difficult to differentiate (2.12) and obtain a minimum principle. In §4, however, we consider a situation where $q(\cdot, t)$ (and $\hat{q}$, $\hat{p}$) are defined in terms of finite-dimensional parameters.

3. A verification theorem.

The notation is that of §2. In this section we consider a fully observed risk-sensitive control problem with dynamics

$$dx_t = g(x_t, v_t)dt + \sigma(x_t)dw_t, \qquad x_0 = \xi \in R^n, \qquad 0 \le t \le T, \tag{3.1}$$

and cost

$$J(v) = \theta E\Big[\exp \theta\Big(\int_0^T L(x_s, v_s)ds + \Phi(x_T)\Big)\Big]. \tag{3.2}$$

The admissible controls are those functions on $U \times [0, \infty)$ with values in U which are adapted to the filtration generated by x and for which the expectation (3.2) is finite. Sufficient conditions for the latter result to hold are given in [2].

Recall $a = (a_{ij}(x)) = \frac{1}{2}\sigma Q \sigma^*$. Suppose there is a solution $\mathcal{X}(x, t)$ of the nonlinear parabolic equation

$$\frac{\partial \mathcal{X}}{\partial t} + \min_v [D\mathcal{X}.g(x, v) + L(x, v)] + TrD^2\mathcal{X}.a + \theta D\mathcal{X}^*_. a.D\mathcal{X} = 0, \tag{3.3}$$

with $\mathcal{X}(x, T) = \Phi(x)$. Write

$$h_t := \int_0^t L(x_s, v_s)ds + \mathcal{X}(x_t, t).$$

Then $J(v) = \theta E[\exp \theta h_T]$. Suppose $u^* : R^n \to U$ is the function such that for each $x \in R^n$

$$D\mathcal{X}.g(x, u^*(x)) + L(x, u^*(x)) = \min_{v \in U}[D\mathcal{X}.g(x, v) + L(x, v)]. \tag{3.4}$$

We can then establish the following verification theorem. A similar proof is given in the paper by Fleming and McEneaney [4].

THEOREM 3.1. *Suppose there is a function $\mathcal{X} \in C^{2,1}$ which is a solution of* (3.3). *With $u^*(\cdot)$ defined by* (3.4) *the feedback control $u^*(x_t)$ is optimal for the fully observed, risk-sensitive control problem* (3.1), (3.2).

Proof. Using Itô's rule

$$\mathcal{X}(x_t, t) = \mathcal{X}(x_0, 0) + \int_0^t \Big(\frac{\partial \mathcal{X}}{\partial s} + D\mathcal{X}.g(x_s, v_s) + TrD^2\mathcal{X}.a)ds + \int_0^t D\mathcal{X}.\sigma dw_s.$$

Therefore,

$$\begin{aligned}
J(v) &= \theta E\Big[\exp\theta\Big(\int_0^T L(x_s,v_s)ds + \mathcal{X}(x,T)\Big)\Big]\\
&= \theta E\Big[\exp\theta\Big(\mathcal{X}(x_0,0) + \int_0^T \Big(\frac{\partial\mathcal{X}}{\partial s} + D\mathcal{X}.g(x_s,v_s) + L(x_s,v_s) + TrD^2\mathcal{X}.a\Big)ds\\
&\qquad + \int_0^T D\mathcal{X}.\sigma dw_s\Big)\Big]\\
&= \theta\exp\theta\mathcal{X}(x_0,0).E\Big[\exp\theta\Big(\int_0^T \Big(\frac{\partial\mathcal{X}}{\partial s} + D\mathcal{X}.g(x_s,v_s) + L(x_s,v_s)\\
&\qquad\qquad + TrD^2\mathcal{X}.a + \theta D\mathcal{X}^*.a.D\mathcal{X}\Big)ds.\Big)\\
&\qquad\qquad \times\exp\Big(\theta\int_0^T D\mathcal{X}.\sigma dw_s - \theta^2\int_0^T D\mathcal{X}^*.a.D\mathcal{X}ds\Big)\Big].
\end{aligned} \tag{3.5}$$

From (3.3), therefore,

$$\begin{aligned}
J(u^*) &= \theta\exp\theta\mathcal{X}(x_0,0).E\Big[\exp\Big(\theta\int_0^T D\mathcal{X}.\sigma dw - \theta^2\int_0^T D\mathcal{X}^*.a.D\mathcal{X}ds\Big)\Big]\\
&= \theta\exp\theta\mathcal{X}(x_0,0)
\end{aligned}$$

because the term in the expectation is a martingale. For any other control v

$$\begin{aligned}
&\frac{\partial\mathcal{X}}{\partial s} + D\mathcal{X}.g(x_s,v_s) + L(x_s,v_s) + TrD^2\mathcal{X}.a + \theta D\mathcal{X}^*.a.D\mathcal{X}\\
&\quad\geq \frac{\partial\mathcal{X}}{\partial s} + D\mathcal{X}.g(x_s,u^*(x_s)) + L(x_s,u^*(x_s)) + TrD^2\mathcal{X}.a + \theta D\mathcal{X}^*.a.D\mathcal{X} = 0.
\end{aligned}$$

Consequently, from (3.5)

$$\begin{aligned}
J(v) &\geq \theta\exp(\theta\mathcal{X}(x_0,0))E\Big[\exp\Big(\theta\int_0^T D\mathcal{X}.\sigma dw_s - \theta^2\int_0^T D\mathcal{X}^*.a.D\mathcal{X}ds\Big)\Big]\\
&= \theta\exp(\theta\mathcal{X}(x_0,0)) = J(u^*).
\end{aligned}$$

Therefore, u^* is an optimal control.

4. Linear dynamics. The results of §2 are now specialized to the situation where

$$\begin{aligned}
g(x,v) &= F(v).x + G(v),\\
\sigma &= I, \quad \text{the } n\times n \text{ identity matrix},\\
a &= \frac{1}{2}Q(t), \quad \text{a time-varying } n\times n \text{ matrix},\\
h(x,t) &= H_t x + h_t,\\
L(x,v) &= \frac{1}{2}M(v)x\cdot x + m(v)\cdot x + N(v),\\
\Phi(x) &= \Phi(x).
\end{aligned}$$

Here $F(\cdot)$ and $M(\cdot)$ are maps from U into the space of $n\times n$ matrices $L(R^n,R^n)$, $G(\cdot)$ and $m(\cdot)$ are maps from U into R^n, $H_t\in L(R^n,R^m)$, $h_t\in R^m$, and N is a real function defined on U. We suppose ξ is described by a normal density

$$p_0(x) = \exp\frac{\left(-\frac{1}{2}P_0^{-1}(x-x_0)\cdot(x-x_0)\right)}{(2\Pi)^{n/2}|P_0|^{1/2}}.$$

The mean of ξ is, therefore, x_0. We shall often write H for H_t, Q for Q_t, etc.

LEMMA 4.1. *In this case the equation* (2.8) *for* $q(x,t)$ *is*

$$\begin{aligned}(4.1)\quad dq = \Big[&- Dq.(F(v)x + G(v)) - qTr.F(v) \\ &+ \theta q\Big(\frac{1}{2}M(v)x\cdot x + m(v)\cdot x + N(v)\Big) + \frac{1}{2}TrQD^2q\Big]dt \\ &+ q(Hx+h)^* R_t^{-1} dz_t,\end{aligned}$$

with $q(x,0) = p_0(x)$.

THEOREM 4.2. *The solution of* (4.1) *is*

$$(4.2)\qquad q(x,t) = \frac{\nu_t \cdot \exp\left(-\frac{1}{2}\Pi_t^{-1}(x-r_t)\cdot(x-r_t)\right)}{(2\Pi)^{n/2}|\Pi_t|^{1/2}}.$$

Here $r_t = r_t^v$ *is given by*

$$\begin{aligned}(4.3)\quad dr_t = \big(F(v)\cdot r_t + G(v) + \theta\Pi_t(M(v)\cdot r_t + m(v))\big)dt \\ + \Pi_t H^* R_r^{-1}(dz_t - (Hr_t + h)dt),\end{aligned}$$

with $r(0) = x_0$; $\Pi_t = \Pi_t^v$ *is given by the Riccati equation*

$$(4.4)\qquad \dot{\Pi}_t = F(v)\Pi_t + \Pi_t F^*(v) + Q + \Pi_t(\theta M(v) - H^* R_t^{-1} H)\Pi_t,$$

with $\Pi_0 = P_0$; *and*

$$\begin{aligned}(4.5)\quad \nu_t = \exp\Big[&\int_0^t (Hr_s + L)^* R_s^{-1} dz_s - \frac{1}{2}\int_0^t (Hr_s+h)^* R_s^{-1}(Hr_s+h)ds \\ &+ \theta\int_0^t \Big(\frac{1}{2}M(v)r_s\cdot r_s + m(v)\cdot r_s + N(v) + \frac{1}{2}Tr.\Pi_s M(v)\Big)ds\Big].\end{aligned}$$

Note that we are making the following assumption: we suppose (4.4) has a bounded, symmetric solution $\Pi^v_\cdot$.

Proof. Differentiating $q(x,t)$ defined by (4.2), (4.3), (4.4), and (4.5) and substituting in (4.1) verify the result.

Remark 4.3. Recall that under measure $\widehat{P}$ $b^v_\cdot = b$ is a Wiener process where

$$b_t = z_t - \int_0^t (Hr_s + h)ds;$$

consequently equation (4.3) for r can be written

$$(4.6)\qquad dr_t = \big(F(v)r_t + G(v) + \theta\Pi_t(M(v)r_t + m(v))\big)dt + \Pi_t H^* R_t^{-1} db_t.$$

PROPOSITION 4.4. *Suppose the derivative* $\frac{d}{dt}((Hx+h)R_t^{-1}) = ((Hx+h)R_t^{-1})'$ *exists. If we write* $\Psi_t(x) = \exp(-(Hx+h)^* R_t^{-1} z_t)$, *the robust process* $\hat{q}(x,t) =$

$q(x,t)\Psi_t(x)$ is the solution of the parabolic partial differential equation

$$\begin{aligned} d\hat{q} = \Big[&- (\hat{q}H^* R_t^{-1} z_t + D\hat{q})(F(v)x + G(v)) \\ &- \hat{q} Tr F(v) + \theta \hat{q}\Big(\frac{1}{2} M(v)x \cdot x + m(v) \cdot x + N(v)\Big) \\ &- \hat{q}((Hx+h)^* R_t^{-1})' z_t - \frac{\hat{q}}{2}(Hx+h)^* R_t^{-1}(Hx+h) \\ &+ \frac{1}{2} Tr Q_t \cdot D^2 \hat{q} + \frac{1}{2} \hat{q} z_t^* R_t^{-1} H Q_t H^* R_t^{-1} z_t + z_t^* H^* R_t^{-1} Q_t \cdot D\hat{q}\Big] dt, \end{aligned} \tag{4.7}$$

$$\hat{q}(x,0) = p_0(x).$$

Proof. From (2.10) and (4.1) we have

$$\begin{aligned} d\hat{q} = \Big[&- \Psi \cdot Dq(F(v)x + G(v)) - \Psi q \cdot Tr F(v) \\ &+ \theta \Psi q \cdot \Big(\frac{1}{2} M(v)x \cdot x + m(v) \cdot x + N(v)\Big) + \frac{1}{2}\Psi . Tr Q D^2 q \\ &- \Psi q.((Hx+h)R_t^{-1})' z_t - \frac{1}{2}\Psi q.(Hx+h)^* R_t^{-1}(Hx+h)\Big] dt. \end{aligned} \tag{4.8}$$

We wish to write the right-hand side in terms of $\hat{q} = \Psi q$. Now

$$\begin{aligned} D\Psi_t &= -\Psi_t \cdot H^* R_t^{-1} z_t, \\ D^2 \Psi_t &= \Psi_t H^* R_t^{-1} z_t \otimes H^* R_t^{-1} z_t. \end{aligned}$$

Also, $D(\Psi q) = D\Psi \cdot q + \Psi \cdot Dq$ and

$$D^2(\Psi q) = D^2 \Psi \cdot q + 2(D\Psi) \otimes (Dq) + \Psi \cdot D^2 q,$$

so

$$-\Psi \cdot Dq = -(\Psi q) H^* R_t^{-1} z_t dt - D(\Psi q)$$

and

$$\Psi \cdot D^2 q = D^2(\Psi q) - (\Psi q) H^* R_t^{-1} z_t \otimes H^* R_t^{-1} z_t + 2H^* R_t^{-1} z_t \otimes (D(\Psi q) - (D\Psi) \cdot q).$$

Substituting in (4.8) we obtain (4.7).

Remark 4.5. The adjoint backward parabolic equation for $\hat{p}(x,t)$ is, therefore,

$$\begin{aligned} -\frac{\partial \hat{p}}{\partial t} = \Big[&(D\hat{p} - \hat{p} H^* R_t^{-1} z_t)(F(v)x + G(v)) \\ &+ \theta \hat{p}\Big(\frac{1}{2} M(v)x \cdot x + m(v) \cdot x + N(v)\Big) - \hat{p}((Hx+h)^* R_t^{-1})' z_t \\ &- \frac{\hat{p}}{2}(Hx+h)^* R_t^{-1}(Hx+h) + \frac{1}{2} Tr Q_t \cdot D^2 \hat{p} + \frac{1}{2}\hat{p} z_t^* R_t^{-1} H Q_t H^* R_t^{-1} z_t \\ &+ z_t^* H^* R_t^{-1} Q_t \cdot D\hat{p}\Big] dt, \end{aligned} \tag{4.9}$$

$$\hat{p}(x,T) = \exp(\Phi(x) + (Hx+h)^* R_T^{-1} z_T).$$

It is of interest that we can give a finite-dimensional solution of (4.9).

THEOREM 4.6. *The solution of* (4.9) *is*

$$\hat{p}(x,t) = \mu_t \cdot \int_{R^n} \frac{\exp\big(-\frac{1}{2}S_t^{-1}(x-\gamma_t(y))\cdot(x-\gamma_t(y))\big)}{(2\Pi)^{n/2}|S_t|^{1/2}} \cdot \exp(\Phi(y) + z_T^* R_T^{-1} H_T y)dy. \tag{4.10}$$

Here $S_t = S_t^v$ *is given by*

$$\dot{S}_t = F(v)S_t + S_t F^*(v) - Q + S_t(H^*R^{-1}H - \theta M(v))S_t, \tag{4.11}$$

with $S_T = 0$; $\gamma_t = \gamma_t(y) = \gamma_t^v(y)$ *is given by*

$$\begin{aligned}\dot{\gamma}_t = {} & S_t(H^*R^{-1}H - \theta M(v))\gamma_t + F(v)\gamma_t + S_t(H^*R^{-1})'z_t \\ & + S_t F^*(v)H^*R^{-1}z_t + S_t H^*R^{-1}h - \theta S_t m(v) + G(v) - QH^*R^{-1}z_t,\end{aligned} \tag{4.12}$$

with $\gamma_T = y \in R^n$, *and*

$$\begin{aligned}\mu_t = \exp\Big(-\frac{1}{2}\int_t^T & \Big[2Tr.F(v) + Tr.(S_s H^*R^{-1}H) - \theta Tr.(S_s M(v)) \\ & + \gamma_s^* H^*R^{-1}H\gamma_s - \gamma_s^* M(v)\gamma_s + 2\gamma_s^* F^*(v)H^*R^{-1}z_s - \gamma_s^*(H^*R^{-1})'z_s \\ & - 2\gamma_s^* m(v) - 2\theta N(v) + 2\gamma_s^* H^*R^{-1}h - z_s^* R^{-1}HQH^*R^{-1}z_s \\ & + 2(h^*R^{-1})'z_s + h^*R^{-1}h + 2z_s^*R^{-1}HG(v)\Big]ds - h_T^* R_T^{-1} z_T\Big).\end{aligned} \tag{4.13}$$

Proof. Differentiating $\hat{p}(x,t)$ defined by (4.10), (4.11), (4.12), and (4.13) and substituting in (4.9) verify the result.

5. A separation principle. With the linear dynamics and cost of §4, for any admissible, Z_t-adapted control v

$$\begin{aligned}J(v) &= \theta\hat{E}\Big[\exp\Big(\theta\int_0^T L(x_s,v_s)ds\Big)\cdot\exp\theta\Phi(x_T)\Big] \\ &= \theta E\Big[\int_{R^n}\exp(\theta\Phi(x))\cdot q(x,T)dx\Big].\end{aligned}$$

Using the explicit form of equation (4.2) for q, this is

$$\begin{aligned}= \theta E\Big[\int_{R^n} & \exp(\theta\Phi(x)).\exp\Big[\int_0^T (Hr_s+h)^*R_s^{-1}dz_s - \frac{1}{2}\int_0^T (Hr_s+h)^*R_s^{-1}(Hr_s+h)ds\Big] \\ & \times\exp\Big[\theta\int_0^T\Big(\frac{1}{2}M(v)r_s\cdot r_s + m(v)\cdot r_s + N(v) + \frac{1}{2}Tr\Pi_s\cdot M(v)\Big)ds\Big] \\ & \times\exp\frac{\big(-\frac{1}{2}\Pi_T^{-1}(x-r_T)\cdot(x-s_T)\big)}{(2\Pi)^{n/2}|\Pi_T|^{1/2}}\cdot dx\Big],\end{aligned}$$

and, in terms of the measure $\hat{P}$, this is

$$\begin{aligned}= \theta\hat{E}\Big[\int_{R^n} & \exp(\theta\Phi(x))\cdot\frac{\exp\big(-\frac{1}{2}\Pi_T^{-1}(x-r_T)\cdot(x-r_T)\big)}{(2\Pi)^{n/2}|\Pi_T|^{1/2}}\cdot dx \\ & \times\exp\Big[\theta\int_0^T\Big(\frac{1}{2}M(v)r_s\cdot r_s + m(v)\cdot r_s + N(v) + \frac{1}{2}Tr\Pi_s\cdot M(v)\Big)ds\Big]\Big].\end{aligned}$$

Write $x = r_T^v + (\Pi_T^v)^{1/2}$, $\xi = r_T + \Pi_T^{1/2}\xi$, and

$$\tilde{\phi}(x, \Pi_T) = (2\Pi)^{-n/2} \int_{R^n} \exp(\theta\Phi(x + \Pi_T^{1/2}\xi)) \exp\Big(- \frac{1}{2}|\xi|^2\Big) d\xi.$$

Then

$$J(v) = \theta\widehat{E}\Big[\exp\theta\Big(\hat{\phi}(r_T, \Pi_T) + \int_0^T \Big(\frac{1}{2}M(v)r_s \cdot r_s + m(v)\cdot r_s + N(v) + \frac{1}{2}Tr\Pi_s \cdot M(v)\Big)ds\Big)\Big], \tag{5.1}$$

where $\hat{\phi}(x, \Pi) = \theta^{-1}\log\tilde{\phi}(x, \Pi)$.

The partially observed, stochastic control problem of §4 is now expressed as a fully observed stochastic control problem with finite-dimensional state (r, Π) defined by (4.6) and (4.4). Consequently, if there is a function $\mathcal{X}$ which is the solution of the analog of equation (3.3) for these dynamics and cost (5.1), we can conclude that the optimal control u^* is the feedback control determined by the function which minimizes the Hamiltonian (3.4). Therefore, in this situation the separation principle holds.

Alternatively, we note that the observed state q is determined by the parameters ν, r, and Π:

$$q(x,t) = q(x,t,\nu,r,\Pi) = \nu \cdot \frac{\exp\big(- \frac{1}{2}\Pi^{-1}(x-r)\cdot(x-r)\big)}{(2\Pi)^{n/2}|\Pi|^{1/2}}.$$

Consequently, the cost process $V(q,t)$ defined in §4 can be considered as a function of ν, r, Π: $V(\nu, r, \Pi, t)$.

If there is an optimal control u^*, consider the $\hat{p}(x,t) = \hat{p}(x,t,\mu_t^{u^*}, S_t^{u^*}, \gamma_t^{u^*})$ defined by (4.10). Write

$$p^*(x,t) = \hat{p}(x,t,\mu_t^{u^*}, S_t^{u^*}, \gamma_t^{u^*})\exp((Hx+h)^*R_t^{-1}z_t).$$

Then

$$V(q,t) = V(\nu, r, \Pi, t) = E\Big[\int_{R^n} p^*(x,t)q(x,t,\nu,r,\Pi)dx\Big].$$

The dynamic programming result, Theorem 2.5, becomes

$$V(\nu, r, \Pi, t) = \inf_{u\in\mathcal{U}_{t,t+h}} E[V(\nu_{t+h}^u, r_{t+h}^u, \Pi_{t+h}^u, t+h) \mid \nu_t = \nu, r_t = r, \Pi_t = \Pi]. \tag{5.2}$$

Suppose V is twice diffentiable in ν and r and differentiable in Π and t so that

$$\begin{aligned} V(\nu_{t+h}^u, r_{t+h}^u, \Pi_{t+h}^u, t+h) &= V(\nu, r, \Pi, t) + \int_t^{t+h} \frac{\partial V}{\partial t}\cdot ds + \int_t^{t+h}\frac{\partial V}{\partial \nu}\cdot d\nu_s \\ &+ \int_t^{t+h}\frac{\partial V}{\partial r}\cdot dr_s + \int_t^{t+h}\frac{\partial V}{\partial \Pi}\cdot d\Pi_s + \frac{1}{2}\int_t^{t+h}\frac{\partial^2 V}{\partial \nu^2}\Pi H^* R^{-1}H\Pi ds \\ &+ \frac{1}{2}\int_t^{t+h}\frac{\partial^2 V}{\partial r^2}\nu(Hr+h)^*R^{-1}(Hr+h)ds. \end{aligned}$$

Substituting in (5.2), dividing by $h > 0$, and letting h go to 0 we obtain the separated minimum principle which states that, for almost all t, the optimal control

$u^*(\nu, r, \Pi)$ is the value which minimizes the Hamiltonian

$$\frac{\partial V}{\partial r} \cdot \big(F(v)r + G(v) + \theta\Pi(M(v)r + m(v))\big) + \frac{\partial V}{\partial \Pi} \cdot (F(v)\Pi + \Pi F^*(v) + \theta\Pi M(v)\Pi) + \frac{\partial V}{\partial v} \cdot v\theta\Big(\frac{1}{2}M(v)r \cdot r + m(v) \cdot r + N(v) + \frac{1}{2}Tr.\Pi M(v)\Big).$$

6. Conclusion. In a partially observed stochastic control problem with an exponential running cost of modified Zakai equation was introduced whose solution is a measure which is related not only to the state of the process but also to the cost. The situation when the dynamics are linear and the exponential running cost is quadratic is discussed, although the control parameter may enter both dynamics and cost nonlinearly. In this case the solution of the modified Zakai equation can be found explicitly. An explicit solution was also found for the backward adjoint process. This permits the partially observed stochastic control problem to be written as a fully observed problem in terms of an information state with finite-dimensional dynamics so that the separation principle holds.

Acknowledgment. The hospitality and support of the Institut National de Recherche en Informatique et en Automatique are gratefully acknowledged.

REFERENCES

[1] A. Bensoussan, *Stochastic Control of Partially Observable Systems*, Cambridge University Press, Cambridge, UK, 1992.

[2] A. Bensoussan and J. H. van Schuppen, *Optimal control of partially observable stochastic systems with an exponential-of-integral performance index*, SIAM J. Control Optim., 23 (1985), pp. 599–613.

[3] R. J. Elliott, *Stochastic Calculus and Applications*, in Applications of Mathematics 18, Springer-Verlag, Berlin, Heidelberg, New York, 1982.

[4] W. H. Fleming and W. M. McEneaney, *Risk Sensitive Optimal Control and Differential Games*, Springer Lecture Notes in Control and Information Science 184, Springer-Verlag, New York, 1992, pp. 185–197.

[5] B. Hajek and E. Wong, *Stochastic Processes in Engineering Systems*, Springer-Verlag, New York, 1985.

[6] D. H. Jacobson, *Optimal stochastic linear systems with exponential criteria and their relation to deterministic differential games*, IEEE Trans. Automat. Control, AC-18 (1973), pp. 124–131.

[7] M. R. James, J. S. Baras, and R. J. Elliott, *Output feedback risk-sensitive control and differential games for continuous time nonlinear systems*, Presented at 32nd IEEE Conf. on Decision and Control, San Antonio, TX, December 1993, pp. 3357–3360.

[8] P. R. Kumar and J. H. van Schuppen, *On the optimal control of stochastic systems with exponential-of-integral performance index*, J. Math. Anal. Appl., (1981), pp. 312–332.

[9] E. Pardoux, *Equations du filtrage nonlinéaire, de la prediction et du lissage*, Stochastics, 6 (1982), pp. 193–232.

[10] J. L. Speyer, *An adaptive terminal guidance scheme based on an exponential cost criterion with application to homing missile guidance*, IEEE Trans. Automat. Control, AC-21 (1976), pp. 371–375.

[11] P. Whittle, *Risk sensitive linear quadratic Gaussian control*, Adv. in Appl. Prob, 13 (1981), pp. 764–777.

[12] W. M. Wonham, *On the separation theorem of stochastic control*, SIAM J. Control, 6 (1968), pp. 312–326.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 33, No. 6, pp. 1847–1856, November 1995

© 1995 Society for Industrial and Applied Mathematics

012

ON THE TOPOLOGY OF THE KARUSH–KUHN–TUCKER SET UNDER MANGASARIAN–FROMOVITZ CONSTRAINT QUALIFICATION*

HARALD GÜNZEL†

Abstract. This paper deals with smooth optimization problems $\mathcal{P}$ in $I\!R^n$ depending on parameter $y \in I\!R^p$. The problem $\mathcal{P}(y)$ is defined by means of a finite number of equality and inequality constraints. We study the set $\mathbf{\Sigma}_{KKT}$ of pairs (x,y) such that x is a Karush–Kuhn–Tucker point of the problem $\mathcal{P}(y)$. Let $\mathbf{\Sigma}$ denote the subset of $\mathbf{\Sigma}_{KKT}$ at which the Mangasarian–Fromovitz constraint qualification is fulfilled. For problem data in general position we prove that $\mathbf{\Sigma}$ is a topological manifold of dimension p.

Key words. parametric optimization, Karush–Kuhn–Tucker set, Mangasarian–Fromovitz constraint qualification, topological manifold

AMS subject classification. 90C31

1. Introduction. Consider the optimization problem $\mathcal{P}(y)$, where $y \in I\!R^p$ is a parameter:

$$\mathcal{P}(y) \quad : \quad \max\{\, f(x,y) \mid x \in M(y) \,\}, \qquad \text{where}$$

$$M(y) \;=\; \{\, x \in I\!R^n \mid h_i(x,y) = 0,\; i \in \mathcal{H}\;;\; g_j(x,y) \le 0,\; j \in \mathcal{G} \,\}.$$

The sets $\mathcal{H} = \{1,\ldots,\mathrm{h}\}$ and $\mathcal{G} = \{1,\ldots,\mathrm{g}\}$ are finite index sets. If not explicitly stated otherwise, all appearing mappings and manifolds are assumed to be smooth, i.e., of class C^∞. Let $\mathbf{\Sigma}_{KKT} \subset I\!R^n \times I\!R^p$ denote the set of Karush–Kuhn–Tucker (KKT) points of $\mathcal{P}$; a pair $(x,y) \in I\!R^n \times I\!R^p$ is called a KKT-point if the following conditions hold with some $\lambda \in I\!R^{\mathrm{h}}$ and $\mu \in I\!R^{\mathrm{g}}$:

(KKT1) $x \in M(y)$,
(KKT2) $\mu_j \ge 0\;,\quad j \in \mathcal{G}_0(x,y)$,
(KKT3) $\mu_j = 0\;,\quad j \notin \mathcal{G}_0(x,y)$,
(KKT4) $D_x f \;=\; \sum_i \lambda_i\, D_x h_i \;+\; \sum_j \mu_j\, D_x g_j\;\Big|_{(x,y)}$.

Here, $\mathcal{G}_0(x,y) := \{\, j \in \mathcal{G} \mid g_j(x,y) = 0 \,\}$ denotes the set of active inequality constraints. A pair (λ,μ) satisfying conditions (KKT1)–(KKT4) is called a Lagrange multiplier associated with (x,y). Note that the set of Lagrange multipliers associated with (x,y) is a convex polyhedron.

The Mangasarian–Fromovitz constraint qualification (MFCQ) is said to hold at (x,y) if (KKT1)–(KKT3), together with the relation $\sum_i \lambda_i\; D_x h_i(x,y) + \sum_j \mu_j\; D_x g_j(x,y) \;=\; 0$, imply that (λ,μ) vanishes. In that case, the set of Lagrange multipliers for (x,y) is compact (cf. [3]) and will be referred to as a Lagrange polytope. Let $\mathbf{\Sigma} \subset \mathbf{\Sigma}_{KKT}$ denote the subset of those KKT-points at which the MFCQ is satisfied. The set $\mathbf{\Sigma}$ is called the KKT-set.

MAIN THEOREM. *There exists a C^2-open and C^∞-dense subset $\mathcal{F}$ of $C^\infty(I\!R^{n+p}, I\!R^{1+\mathrm{h}+\mathrm{g}})$ with the following property: If $(f,H,G) \in \mathcal{F}$, then the corresponding KKT-set $\mathbf{\Sigma}$ is a topological manifold of dimension p.*

* Received by the editors October 27, 1993; accepted for publication (in revised form) June 21, 1994. This research was supported by Deutsche Forschungsgemeinschaft, Graduiertenkolleg “Analyse und Konstruktion in der Mathematik.”

† Department of Mathematics (C), Aachen University of Technology, D-52056 Aachen, Germany (guenzel@RWTH-Aachen.de).

Here, $H = (h_1, \dots, h_{\mathrm{h}})$. The topology used above is defined as follows. First, let $s \in I\!N$. For $\bar{F} \in C^\infty(I\!R^n, I\!R^m)$ and $\varepsilon \in C^0(I\!R^n, I\!R_+)$, we define an ε-neighborhood of $\bar{F}$:

$$U_\varepsilon(\bar{F}) := \left\{ F \in C^\infty(I\!R^n, I\!R^m) \;\middle|\; \| \partial_\alpha F(x) - \partial_\alpha \bar{F}(x) \| < \varepsilon(x) \ \ \forall x \in I\!R^n \ , \ |\alpha| \le s \right\}.$$

Here, α stands for the multiindex of a partial derivative, $I\!R_+ := (0, \infty)$. The sets $U_\varepsilon(\bar{F})$ form a basis of the C^s-topology. The C^∞-topology is defined to be the union of all C^s-topologies. Let $\mathcal{X}$ be the KKT-set blown up by means of the associated Lagrange polytope, i.e.,

$$\mathcal{X} = \{(x, y, \lambda, \mu) \in I\!R^n \times I\!R^p \times I\!R^{\mathrm{h}} \times I\!R^{\mathrm{g}} \mid \text{(KKT1)–(KKT4) and MFCQ hold at } (x, y)\}.$$

We call $\mathcal{X}$ the extended KKT-set. We actually will prove the following theorem, which implies our main theorem.

THEOREM 1.1. *There exists a C^2-open and C^∞-dense subset $\mathcal{F}$ of $C^\infty(I\!R^{n+p}, I\!R^{1+\mathrm{h}+\mathrm{g}})$ with the following property: If $(f, H, G) \in \mathcal{F}$, then the extended KKT-set $\mathcal{X}$ is a topological manifold of dimension p. Moreover, there exists a homeomorphism $\varphi : \mathcal{X} \to \mathbf{\Sigma}$, arbitrarily C^0-close to the natural projection $\mathcal{X} \to \mathbf{\Sigma}$.*

The main point in Theorem 1.1 is the existence of the homeomorphism φ. The fact that the extended KKT-set $\mathcal{X}$ is a topological manifold was already known (cf. [12]). For a nice outline of the latter idea see [9]. The one-parametric case (including the KKT-set $\mathbf{\Sigma}$) was treated by Kojima and Hirabayashi (cf. [10]).

The reason for the homeomorphy between $\mathcal{X}$ and $\mathbf{\Sigma}$ can be illustrated as follows. There is a natural equivalence relation on $\mathcal{X}$, namely, to belong to the same Lagrange polytope. The set $\mathbf{\Sigma}$ then becomes the quotient space. The family of Lagrange polytopes is upper semicontinuous; i.e., each Lagrange polytope has an arbitrarily small open neighborhood which itself is union of Lagrange polytopes (cf. [11]). There are general techniques dealing with topological spaces (like $\mathbf{\Sigma}$) obtained by simultaneous identification of an upper semicontinuous family of closed discs in a manifold (like $\mathcal{X}$) (cf. [2]). However, it seems hard to verify the assumptions made there, in particular the so-called disjoint disk property of the target space. Nevertheless, in our case we can exploit the special structure.

There are only a few partial results about the (local) homeomorphy of $\mathcal{X}$ and $\mathbf{\Sigma}$. In fact, the two-parameter case is treated by Hirabayashi, Shida, and Shindoh (cf. [7]). Under additional assumptions, the multiparametric case is studied by Schecter [12], where our main theorem was conjectured. In particular, Schecter assumes a certain rank condition as well as a condition on the "nondegeneracy" of the Lagrange polytopes. Both conditions, however, are not generic.

Roughly speaking, our homeomorphy proof is based on induction. Suppose we have established a homeomorphism sending the subset of $\mathcal{X}$ formed by all Lagrange polytopes with dimension less than d to its projection in $\mathbf{\Sigma}$. The induction step consists of perturbing the latter homeomorphism in order to make an extension to the critical set, which is the union of Lagrange polytopes of dimension equal to d. To this end, one considers certain fibres in $\mathcal{X}$ which properly contain exactly one Lagrange polytope of dimension d. These fibres are pairwise disjoint and form a partition for a neighborhood of the critical set. For a specific fibre we make a special construction, and then we extend this construction over the bundle of all fibres. For the latter extension we apply Thom's isotopy lemma.

The rest of this paper is organized as follows. In §2 we cover the background material needed for the application of Thom's isotopy lemma in a nonproper setting.

Section 3 is devoted to the special construction on a specific fibre. Finally, in §4 all information is put together in order to prove Theorem 1.1.

2. The characteristic set. Our situation modelled in terms of jet extensions. First applications of this technique to mathematical programming are due to Jongen, Jonker, and Twilt (cf. [8]). We define characteristic sets $\mathbf{S} \subset \mathrm{Jet}_{\mathbf{\Sigma}}$ and $\mathbf{K} \subset \mathrm{Jet}_{\mathcal{X}} \equiv \mathrm{Jet}_{\mathbf{\Sigma}} \times I\!R^{\mathrm{h}} \times I\!R^{\mathrm{g}}$ in jet spaces and a characteristic mapping $j^1 : I\!R^{n+p} \longrightarrow \mathrm{Jet}_{\mathbf{\Sigma}}$ such that we have the fundamental relations

$$z \in \mathbf{\Sigma} \iff j^1(z) \in \mathbf{S} \qquad \text{and} \qquad (z, \lambda, \mu) \in \mathcal{X} \iff (j^1(z), \lambda, \mu) \in \mathbf{K}.$$

The characteristic mapping j^1 will in fact be a (reduced) 1-jet extension of (f, H, G). In this section we show that $\mathbf{S}$ admits a Whitney regular disc stratification. Here, we use tools from real algebraic geometry (cf. [1]). Finally, we introduce a nonproper version of the well-known first isotopy lemma by R. Thom.

Put $\mathrm{Jet}_{\mathbf{\Sigma}} := I\!R^{\mathrm{h}} \times I\!R^{\mathrm{g}} \times \mathcal{M}_{n,\mathrm{h}} \times \mathcal{M}_{n,\mathrm{g}} \times \mathcal{M}_{n,1}$, where $\mathcal{M}_{n,k}$ stands for the space of real (n, k)-matrices. For problem data (f, H, G), define the following reduced 1-jet extension:

$$\begin{aligned} j^1_{(f,H,G)} &: I\!R^{n+p} \longrightarrow \mathrm{Jet}_{\mathbf{\Sigma}}, \\ j^1_{(f,H,G)} &:= (H, G, D_x^{\top} H, D_x^{\top} G, D_x^{\top} f). \end{aligned}$$

Here, $D_x f$ stands for $(\partial_{x_1} f, \ldots, \partial_{x_n} f)$ and $D_x^{\top}$ for the transposed matrix of D_x. When there is no ambiguity, we omit the index (f, H, G). We define the following conditions for a point $(a, b, V, W, v, \lambda, \mu) \in \mathrm{Jet}_{\mathcal{X}}$:

(K1) $a = 0,\ b \leq 0,\ \mu \geq 0,$ *and* $b^{\top}\mu = 0$;
(K2) $v = V\lambda + W\mu$;
(MF) (K1) and (K2) imply $(\lambda', \mu') = (0, 0)$ for any $(a, b, V, W, 0, \lambda', \mu') \in \mathrm{Jet}_{\mathcal{X}}$.

Using the above abbreviations, we define the characteristic set $\mathbf{K}$:

$$\mathbf{K} = \{\ (a, b, V, W, v, \lambda, \mu) \in \mathrm{Jet}_{\mathcal{X}} \ \big|\ (K1), (K2), \text{ and } (MF)\ \}.$$

Denote the restriction of the natural projection by $\Pi : \mathbf{K} \longrightarrow \mathrm{Jet}_{\mathbf{\Sigma}}$; we then have the relation $\mathbf{S} = \Pi(\mathbf{K})$.

2.1. Whitney regular disc stratifications. We prove the existence of a Whitney regular disc stratification of $\mathbf{S}$ which extends as such to its closure $\overline{\mathbf{S}}$. By virtue of [6], both $\mathbf{S}$ and its closure are semialgebraic sets. For details on semialgebraic sets see, e.g., [1].

DEFINITION 2.1. *Let A be a subset of $I\!R^m$ with a locally finite partition $\mathcal{A}$. If the elements of $\mathcal{A}$ are smooth manifolds, then $\mathcal{A}$ is called a stratification of A and the pair $(A, \mathcal{A})$ a stratified set. In the latter case, the elements of $\mathcal{A}$ are also called strata. The stratification $\mathcal{A}$ is called a disc stratification if its strata are open discs, i.e., diffeomorphic to open real unit balls.*

DEFINITION 2.2. *Let $(A, \mathcal{A})$ be a stratified subset of $I\!R^m$. Let $X, Y \in \mathcal{A}$ be distinct strata and $\bar{x} \in X$. Then Y is called Whitney regular over X at point $\bar{x}$ if for any pair of sequences $x^k \longrightarrow \bar{x}$ and $y^k \longrightarrow \bar{x}$ with the following properties*

1. $x^k \in X$, $y^k \in Y$;
2. $T_{y^k}Y \longrightarrow T$ *in* $G(m, \mathrm{dim}Y)$;
3. $\mathcal{L}(y^k - x^k) \longrightarrow L$ *in* $G(m, 1)$

the inclusion $L \subset T$ holds. Y is called Whitney regular over X if this holds at any point $\bar{x}$ of X. The stratification $\mathcal{A}$ is called Whitney regular if any stratum is over any other stratum.

Here $G(m,k)$ denotes the Grassmann manifold of k-dimensional linear subspaces of $\mathbb{R}^m$, whereas $\mathcal{L}$ stands for the linear hull and T_y for the tangent space considered as a subspace of the embedding space. Note that Whitney regularity is well defined on manifolds as well, since it is invariant with respect to diffeomorphisms.

DEFINITION 2.3. *Let $\mathcal{A}$ and $\mathcal{B}$ be families of subsets of a given set M. Then $\mathcal{A}$ is called a refinement of $\mathcal{B}$ if each $X \in \mathcal{A}$ is contained in some $Y \in \mathcal{B}$.*

LEMMA 2.4 (cf. [1]). *Let A be a semialgebraic subset of $\mathbb{R}^m$ and assume a finite family $\mathcal{B}$ of subsets of A, all semialgebraic. Then there exists a finite refinement $\mathcal{A}$ of $\mathcal{B}$ into semialgebraic manifolds constituting a Whitney regular stratification of A.*

LEMMA 2.5 (cf. [1]). *Let A be a semialgebraic subset of $\mathbb{R}^m$. Then there exists a finite disc stratification of A with semialgebraic strata.*

In the present paper we need the following corollary of Lemmas 2.4 and 2.5.

COROLLARY 2.6. *Let A be a semialgebraic subset of $\mathbb{R}^m$ and $\mathcal{B}$ a finite family of semialgebraic subsets of A. Then A admits a (finite) Whitney regular disc stratification refining $\mathcal{B}$.*

Proof. For arbitrary $d \in \mathbb{N}$ we prove the following claim.

CLAIM (d). *There exists a finite Whitney regular stratification $\mathcal{A}$ of A refining $\mathcal{B}$ with semialgebraic strata such that any stratum is a disc or has dimension less than d.*

Claim (0) yields the desired result. Claim (m+1) follows from Lemma 2.4. It remains to show that claim (d+1) implies claim (d).

Put $\mathcal{A}_d := \{ X \in \mathcal{A} \mid \dim(X) = d \}$ and $\mathcal{A}_{>d} := \{ X \in \mathcal{A} \mid \dim(X) > d \}$, where $\mathcal{A}$ has been chosen according to claim (d+1). By virtue of Lemma 2.5, for any $X \in \mathcal{A}_d$ there exists a finite semialgebraic disc stratification $\mathcal{C}_X$ of X. Let $\mathcal{C}_d$ denote the set of all d-dimensional strata in the union of all $\mathcal{C}_X$, $X \in \mathcal{A}_d$. Then, the semialgebraic set $C := A \setminus \bigcup \mathcal{A}_{>d} \setminus \bigcup \mathcal{C}_d$ has dimension less than d; i.e., each stratum of any stratification of C is of dimension less than D. Moreover, $\mathcal{A}_{>(d-1)} := \mathcal{A}_{>d} \cup \mathcal{C}_d$ is a finite Whitney regular disc stratification refining $\mathcal{B}$ (since $\mathcal{C}_d$ refines $\mathcal{A}_d$). Let $\mathcal{D}$ be a refinement of the family $\mathcal{A}_{>(d-1)} \cup \mathcal{B}$ as stated in Lemma 2.4. We replace the subset of $\mathcal{D}$ refining $\mathcal{A}_{>(d-1)}$ by $\mathcal{A}_{>(d-1)}$ itself. In such a way we again obtain a Whitney regular stratification. Now, the stratification $\mathcal{D}$ satisfies the assertion of claim (d). □

COROLLARY 2.7. *There exist Whitney regular disc stratifications $\mathcal{A}$ of $\mathbf{S}$ and $\bar{\mathcal{A}}$ of $\overline{\mathbf{S}}$, respectively, such that $\mathcal{A} \subset \bar{\mathcal{A}}$.*

2.2. Thom's isotopy lemma. We need a slight extension of Thom's first isotopy lemma; see Proposition 2.9 below.

Notation. Let $(A, \mathcal{A})$ be a stratified set in a manifold N, and let N' be an open subset of N. Define $(A, \mathcal{A})|_{N'} := (A \cap N', \mathcal{A}|_{N'})$, where $\mathcal{A}_{N'} := \{X \cap N' | X \in \mathcal{A}\}$.

Note that the restriction $(A, \mathcal{A})|_{N'}$ again is a stratified set. Whitney regularity is carried over as well.

DEFINITION 2.8. *Let $f : N \to M$ be a mapping between manifolds and $(A, \mathcal{A})$ be a stratified subset of N. Then $(A, \mathcal{A})$ is said to be topologically trivial over M (with respect to f) if there exist a stratified set $(B, \mathcal{B})$ and a homeomorphism $h : M \times B \to A$ such that*

1. *$f \circ h = \Pi_M$, Π_M denoting the natural projection;*
2. *for each $X \in \mathcal{B}$ there exists some $Y \in \mathcal{A}$ such that $h : M \times X \to Y$.*

For later use, we emphasize that in the following proposition the set B in Definition 2.8 can be chosen as $A \cap f^{-1}(y)$, and $\mathcal{B}$ can be chosen as $\{ X \cap f^{-1}(y) \mid X \in \mathcal{A} \}$ for some $y \in M$.

PROPOSITION 2.9. *Let $f : N \to I\!R^m$ be a mapping between manifolds and $(A, \mathcal{A})$ be a locally closed Whitney regularly stratified subset of N. Let $\tilde{\mathcal{A}} \subset \mathcal{A}$ be such that $\tilde{A} := \bigcup \tilde{\mathcal{A}}$ is locally closed and such that the following conditions hold for any $X \in \tilde{\mathcal{A}}$:*

1. *$f|_X$ is a submersion,*
2. *$f|_{\overline{X} \cap \tilde{A}}$ is proper.*

Then there exists an open set $N' \subset N$ containing $\tilde{A}$ such that $(A, \mathcal{A})|_{N'}$ is trivial over $I\!R^m$ (with respect to f).

Sketch of Proof. We use the proof (cf. [4]) of the proper version of the isotopy lemma (i.e., Proposition 2.9 for the case $A = \tilde{A}$). In order to keep the exposition short, we focus on merely pointing out the modifications to be made. For missing details we refer to [4]. Starting from the fibre $\{0\}$ in $I\!R^m$, we can fill out the space $I\!R^m$ by successive integration of the constant coordinate vector fields ∂_{x_i}. The main idea of the proof is to perform the same construction on A, starting from the fibre $A \cap f^{-1}(0)$ and integrating controlled lifts of the coordinate vector fields. Any vector field ξ on $I\!R^m$ admits a controlled lift ξ^A onto A, at least in a neighborhood of $\tilde{A}$; i.e., we have $Df\ \xi^A = \xi \circ f$, and certain control relations hold with respect to a tube system. We essentially use the fact that ξ^A is globally integrable if ξ is. Beyond [4], it remains to show that ξ^A (and not just $\xi^A|_{\tilde{A}}$) is globally integrable. In order to see this, recognize that the vector field ξ^A lifts $(\xi^A|_X, 0)$ from $X \times I\!R$ with respect to the mapping (π_X, ρ_X), where $X \in \tilde{\mathcal{A}}$. This is due to the control relations with respect to the tube system; π_X is standing for the tubular projection mapping and ρ_X for the distance function of the tube at X. Since (π_X, ρ_X) is proper, an analogous argument as above yields that ξ^A is globally integrable on a tubular neighborhood of X. Finally, N' can be chosen as a union of tubular neighborhoods of the strata in $\tilde{\mathcal{A}}$. □

3. Controlled explosions in special fibres.

For an arbitrary stratification $\mathcal{A}$ of $\mathbf{S}$ and $X \in \mathcal{A}$, $\bar{x} \in X$, we construct a special projection mapping π onto X such that the fibre $\Pi^{-1}\pi^{-1}(\bar{x}) \subset \mathbf{K}$ can be handled as well as a differentiable manifold. To this end we use a topological transformation of $\mathbf{K}$ which removes the natural creases. This is done in §3.2. Then we can obtain a controlled explosion of the Lagrange polytope $\Pi^{-1}(\bar{x})$ within this fibre by virtue of the construction made in §3.1.

3.1. Controlled explosions.

DEFINITION 3.1. *Let P be a compact subset of a metric space M and $\bar{y} \in P$. A continuous mapping $\psi : (0,1) \times (M \setminus \{\bar{y}\}) \to M \setminus P$ is called a controlled explosion of P (in M) if the following conditions hold for any $\varepsilon \in (0,1)$:*

1. *$\psi_\varepsilon : M \setminus \{\bar{y}\} \to M \setminus P$ is a homeomorphism,*
2. *$\psi_\varepsilon : P \ \setminus \{\bar{y}\} \to U_\varepsilon(P) \setminus P$,*
3. *$\psi_\varepsilon|_{M \setminus U_{3\varepsilon}(P)} = \mathrm{id}_{M \setminus U_{3\varepsilon}(P)}$,*
4. *$\psi_\varepsilon^{(-1)} : M \setminus P \to M \setminus \{\bar{y}\}$ extends continuously by $P \to \{\bar{y}\}$.*

Here define $\psi_\varepsilon(y) := \psi(\varepsilon, y)$ and $\psi_\varepsilon^{(-1)} := {\psi_\varepsilon}^{-1}$. $U_\varepsilon(P)$ stands for the union of the ε-neighborhoods $U_\varepsilon(y)$, $y \in P$. Point $\bar{y}$ is called the centre of explosion.

PROPOSITION 3.2. *Let $M \subset I\!R^n$ be a smooth manifold and $P \subset M$ be a compact convex subset of $I\!R^n$. Then P admits a controlled explosion in M.*

The following technical lemma is essential for the proof of Proposition 3.2.

LEMMA 3.3. *There exists a continuous mapping $\eta : (0,1) \times [0,1] \times I\!R_+ \to I\!R_+$ such that the following conditions hold for any $\varepsilon \in (0,1)$ and $\delta \in [0,1]$:*

1. $\eta_{\varepsilon,\delta} : \mathbb{R}_+ \to \mathbb{R}_+$ *is a homeomorphism if* $\delta > 0$;
2. $\eta_{\varepsilon,0} : \mathbb{R}_+ \to (1,\infty)$ *is a homeomorphism*;
3. $\eta_{\varepsilon,\delta} : (0,1) \to (0, 1+\varepsilon)$;
4. $\eta_{\varepsilon,1} = \mathrm{id}_{\mathbb{R}_+}$;
5. $\eta_{\varepsilon,\delta}\big|_{(1+2\varepsilon,\infty)} = \mathrm{id}_{(1+2\varepsilon,\infty)}$;
6. *by* $(0,1) \times \{0\} \times (0,1] \to \{0\}$, $\eta^{(-1)}$ *extends to a continuous mapping* $(0,1) \times [0,1] \times \mathbb{R}_+ \to \overline{\mathbb{R}_+}$.

Here define $\eta_{\varepsilon,\delta}(x) := \eta(\varepsilon,\delta,x)$ *and* $\eta^{(-1)}(\varepsilon,\delta,x) := {\eta_{\varepsilon,\delta}}^{-1}(x)$.

Proof. Put, for instance,

$$\eta_{\varepsilon,\delta}(x) := \begin{cases} x/\delta, & x \in (0,\delta], \\ 1 \;+\; \varepsilon(x-\delta), & x \in (\delta,1], \\ x \;+\; (1+2\varepsilon-x)(1-\delta)/2, & x \in (1,1+2\varepsilon], \\ x, & x > 1+2\varepsilon. \end{cases} \qquad \square$$

Proof of Proposition 3.2. Reduction Step. Let d be the dimension of P. Assume the nontrivial case $d > 0$. In a neighborhood of M there is defined a projection mapping $\pi : \mathbb{R}^n \to M$; i.e., we have $\pi|_M = \mathrm{id}_M$. Consider the affine hull $\mathrm{aff}(P) \subset \mathbb{R}^n$, which is a submanifold of $\mathbb{R}^n$. Note that $\pi|_P = \mathrm{id}_P$. Hence $\pi|_{\mathrm{aff}(P)}$ is a diffeomorphism in an open neighborhood of P. Consequently, $\tilde{M} := \pi(\mathrm{aff}(P))$ is a d-dimensional submanifold of M containing P. Finally, near P the manifold M is diffeomorphic $\tilde{M} \times \mathbb{R}^{\dim M - d}$. Hence we have without loss of generality $M = \mathbb{R}^n$ and $P \subset \mathbb{R}^d \subset \mathbb{R}^n$. Using a standard coordinate transformation in $\mathbb{R}^d$, we also can assume that $P = D^d \times \{0\}$, where D^d stands for the d-dimensional closed unit ball in $\mathbb{R}^d$.

Application of Lemma 3.3. We may assume that $M = \mathbb{R}^n \equiv \mathbb{R}^d \times \mathbb{R}^{n-d}$ and $P = D^d \times \{0\}$. Now we define $\psi_\varepsilon : \mathbb{R}^n \setminus \{0\} \to \mathbb{R}^n \setminus D^d$ as follows:

$$\psi_\varepsilon(x,y) := \begin{cases} \Big(\frac{x}{\|x\|}\eta_{\varepsilon,\min\{1,\|y\|/\varepsilon\}}(\| x \|),\ y \Big), & x \neq 0, \\ \Big(0 \ , \ y \Big), & x = 0. \end{cases}$$

Here η has been chosen according to Lemma 3.3. A moment of reflection shows that ψ is a controlled explosion of P in M with centre in 0. $\square$

3.2. The special fibre. Consider the following topological embedding:

$$\begin{array}{lccl} \phi & : & \mathcal{M}_{n,k} \times \mathcal{M}_{n,g} \times \mathbb{R}^{\mathrm{h}} \times \mathbb{R}^{\mathrm{g}} & \hookrightarrow & \mathrm{Jet}_{\mathcal{X}}, \\ \phi & : & (V, W, \lambda, \mu) & \mapsto & (0, \mu_-, V, W, V\lambda + W\mu_+, \lambda, \mu_+). \end{array}$$

Here let $\mu_+ := \max\{0,\mu\}$ and $\mu_+ + \mu_- = \mu$. Put $\mathcal{K} := \phi^{-1}(\mathbf{K})$; note that $\mathcal{K}$ is open in its embedding space (hence a manifold) and that $\phi : \mathcal{K} \to \mathbf{K}$ is a homeomorphism. Hence the set $\mathbf{K}$ constitutes a topological manifold which has codimension $n + \mathrm{h} + \mathrm{g}$ in $\mathrm{Jet}_{\mathcal{X}}$.

Let be $I \subset \{1,\dots,\mathrm{g}\}$. For any $b \in \mathbb{R}^{\mathrm{g}}$ let $b^I \in \mathbb{R}^{\mathrm{g}}$ be defined by means of the following relations: $b^I_i = b_i$ if $i \in I$, and $b^I_i = 0$ otherwise. Note that $(b^I)_+ = (b_+)^I$. Define $\mathbb{R}^I := \{b^I \mid b \in \mathbb{R}^{\mathrm{g}}\}$. Put $CI := \{1,\dots,g\} \setminus I$. Let $\mathbf{S}^I$ denote the subset of $\mathbf{S}$ with active index set I, i.e.,

$$\mathbf{S}^I = \{\ (a,b,V,W,v) \in \mathbf{S} \mid b_i = 0 \iff i \in I\ \}.$$

PROPOSITION 3.4. *Let X be a smooth submanifold of* $\mathrm{Jet}_{\boldsymbol{\Sigma}}$ *such that* $X \subset \mathbf{S}^I$ *for some* $I \subset \{1,\dots,\mathrm{g}\}$, *and let* $\bar{x} \in X$. *Then there exists a projection mapping*

$\pi : \mathrm{Jet}_{\mathbf{\Sigma}} \to X$ *(defined in a neighborhood of* $\bar{x}$*) such that* $\phi^{-1}\Pi^{-1}\pi^{-1}(\bar{x})$ *is a smooth manifold.*

Proof. Consider the following linear subspaces of the Euclidean space $\mathrm{Jet}_{\mathbf{\Sigma}} \equiv \mathbb{R}^{\mathrm{h}} \times \mathbb{R}^{\mathrm{g}} \times \mathbb{R}^{\mathrm{h}n} \times \mathbb{R}^{\mathrm{g}n} \times \mathbb{R}^{n}$. $L_1 := \{0\} \times \mathbb{R}^n$ and $L_2 := {L_1}^{\perp}$, the orthogonal complement. Let $\Pi_{L_2} : \mathrm{Jet}_{\mathbf{\Sigma}} \to L_2$ denote the orthogonal projection and $T_{\bar{x}}X$ the tangent space considered a subspace of $\mathrm{Jet}_{\mathbf{\Sigma}}$. Let us consider the subspaces $T_1 := T_{\bar{x}}X \cap L_1$ and $T_2 := \Pi_{L_2} T_{\bar{x}}X$. Put $\Pi_i := \Pi_{T_i}$ and note that $(\Pi_1, \Pi_2)|_X$ is a diffeomorphism at $\bar{x}$. We define the following mapping:

$$\begin{array}{rccc} \alpha & : & \mathrm{Jet}_{\mathbf{\Sigma}} & \to & \mathrm{Jet}_{\mathbf{\Sigma}}, \\ \alpha & : & (a, b, V, W, v) & \mapsto & (a, b^{CI}, V, W, v + Wb^I). \end{array}$$

Since $X \subset \mathbf{S}^I$, we have $\alpha|_X = \mathrm{id}_X$. Hence there is defined a projection mapping $\pi : \mathrm{Jet}_{\mathbf{\Sigma}} \to X$ (defined on a neighborhood of $\bar{x}$) by the relation $(\Pi_1, \Pi_2) \circ \pi = (\Pi_1, \Pi_2) \circ \alpha$. It suffices to show that $\psi^{-1}(\bar{y})$ is a manifold, where $\psi : \mathcal{K} \to T_1 \times T_2$ is defined by $\psi := (\Pi_1, \Pi_2) \circ \alpha \circ \Pi \circ \phi$ and $\bar{y}$ by $\bar{y} := (\Pi_1, \Pi_2)(\bar{x})$. Now

$$(V, W, \lambda, \mu) \overset{\Pi\circ\phi}{\mapsto} (0, \mu_-^I + \mu_-^{CI}, V, W, V\lambda + W\mu_+).$$

In a neighborhood of $\phi^{-1}\Pi^{-1}(\bar{x})$ we have $\mu_+ = \mu_+^I$ and $\mu_-^{CI} = \mu^{CI}$. Hence

$$(V, W, \lambda, \mu) \overset{\alpha\circ\Pi\circ\phi}{\mapsto} (0, \mu^{CI}, V, W, V\lambda + W\mu^I).$$

The latter mapping is smooth and thus ψ is as well. Note that at point $\bar{x}$ the Jacobian $D_{(V,W,\mu^{CI})}\psi$ maps surjectively onto T_2 and $D_{(\lambda,\mu^I)}\psi$ onto T_1. An application of the theorem on implicit functions yields the desired result. □

Remark. The problem in the latter construction was to find a differentiable mapping defined on $\mathrm{Jet}_{\mathbf{\Sigma}}$ such the fibre induced to $\mathcal{K}$ became a differentiable manifold, although this induction was due to a mapping which not had been differentiable at all.

4. Proof of the theorem. The concepts of active index set and Lagrange polytope are extended to $\mathbf{S}$ in a natural way. For $x \in \mathbf{S}$ we put $\mathcal{G}_0(x) = I$ iff $x \in \mathbf{S}^I$. Using the abbreviation $x = (a, b, V, W, v)$ for elements of $\mathrm{Jet}_{\mathbf{\Sigma}}$, we define the Lagrange polytope $P(x)$ for $x \in \mathbf{S}$:

$$P(x) := (V|W)^{-1}(v) \cap (\mathbb{R}^{\mathrm{h}} \times \overline{\mathbb{R}_+}^{\mathcal{G}_0(x)}).$$

Here $\overline{\mathbb{R}_+} = [0, \infty)$. Note that $\Pi^{-1}(x) = \{x\} \times P(x)$. It is useful to consider the faces of the Lagrange polytope $P(x)$ by means of its generating index sets:

$$\mathcal{J}(x) := \{J \subset \mathcal{G}_0(x) \mid P(x) \cap (\mathbb{R}^{\mathrm{h}} \times {\mathbb{R}_+}^J) \neq \emptyset\}.$$

We have an estimation result for the fineness of disc stratifications of $\mathbf{S}$.

PROPOSITION 4.1 (cf. [5]). *Let* $\mathcal{A}$ *be a Whitney regular disc stratification for* $\mathbf{S}$ *and* $X \in \mathcal{A}$. *Then both set-valued mappings* $\mathcal{G}_0 : X \to 2^{\mathcal{G}}$ *and* $\mathcal{J} : X \to 2^{2^{\mathcal{G}}}$ *are constant.*

From now on, let us use the following notation. If X is a set, then $\widehat{X} := \mathbb{R}^{n+p} \times X$, and if $f : X \to Y$ is a mapping between sets, then $\widehat{f} : \widehat{X} \to \widehat{Y}$ denotes the mapping $\widehat{f} := \mathrm{id}_{\mathbb{R}^{n+p}} \times f$.

Note that the mappings $(\mathrm{id}_{\mathbb{R}^{n+p}}, j^1) : \mathbb{R}^{n+p} \to \widehat{\mathrm{Jet}_{\mathbf{\Sigma}}}$ and $(\mathrm{id}_{\mathbb{R}^{n+p}}, j^1) \times \mathrm{id}_{\mathbb{R}^{\mathrm{h}+\mathrm{g}}}$ are embeddings of $\mathbf{\Sigma}$ into $\widehat{\mathbf{S}}$ and of $\mathcal{X}$ into $\widehat{\mathbf{K}}$, respectively. In this setting, the mapping

$\widehat{\Pi} : \widehat{\mathbf{K}} \to \widehat{\mathbf{S}}$ coincides on $\mathcal{X} \subset \widehat{\mathbf{K}}$ with the natural projection onto $\mathbf{\Sigma} \subset \widehat{\mathbf{S}}$. We have a mapping $\xi : \widehat{\mathrm{Jet}_{\mathbf{\Sigma}}} \to \mathrm{Jet}_{\mathbf{\Sigma}}$ defining $\mathbf{\Sigma}$ as a subset of $\widehat{\mathbf{S}}$, i.e., $\mathbf{\Sigma} = \widehat{\mathbf{S}} \cap \xi^{-1}(0)$. This mapping is given by $\xi := j^1 \circ \Pi_{\mathbb{R}^{n+p}} - \Pi_{\mathrm{Jet}_{\mathbf{\Sigma}}}$, where $\Pi_{\mathbb{R}^{n+p}}$ and $\Pi_{\mathrm{Jet}_{\mathbf{\Sigma}}}$ denote the natural projections. Note that $\xi = \xi_{(f,H,G)}$ depends on the problem data (f, H, G) which have been used.

For $J \subset I \subset \mathcal{G}$, we define the following subset of $\mathbf{K}$:

$$K_J^I := \big\{(a, b, V, W, v, \lambda, \mu) \in \mathbf{K} \ \big| \ b_i = 0 \iff i \in I \text{ and } \mu_j > 0 \iff j \in J\big\}.$$

In view of Corollary 2.7, there exist Whitney regular disc stratifications $\mathcal{A} \subset \bar{\mathcal{A}}$ of $\mathbf{S}$ and $\overline{\mathbf{S}}$, respectively. With $\widehat{\mathcal{A}} := \{\widehat{X} \mid X \in \mathcal{A}\}$ we have a disc stratification of $\widehat{\mathbf{S}}$. Applying the jet transversality theorem as described in [6], we see the existence of a C^2-open and C^∞-dense subset $\mathcal{F} \subset C^\infty(\mathbb{R}^{n+p}, \mathbb{R}^{1+\mathrm{h}+\mathrm{g}})$ such that $\xi|_{\widehat{X}}$ is submersive on $\xi^{-1}(0)$ for any $X \in \mathcal{A}$. Here we assume $\xi = \xi_{(f,H,G)}$ with $(f, H, G) \in \mathcal{F}$.

For $d \in \mathbb{N}$, let $\mathbf{S}_d$ be the union of all strata $X \in \mathcal{A}$ with dimension $\dim X \geq d$. Put $\mathbf{K}_d := \Pi^{-1}\mathbf{S}_d$. Let $\widehat{\mathbf{S}}_d := \widehat{\mathbf{S}_d}$. Now let $\bar{d}$ denote the dimension of $\mathbf{S}$; i.e., we have $\mathbf{S}_{\bar{d}} \neq \emptyset$ and $\mathbf{S}_{\bar{d}+1} = \emptyset$. For $d = 0, 1, \ldots, \bar{d}$ we are going to prove the following assertion.

Assertion (d). There exists a continuous mapping $\varphi_d : \widehat{\mathbf{K}} \to \widehat{\mathbf{S}}$, arbitrarily C^0-close to $\widehat{\Pi}$, such that the following conditions are satisfied:

1. φ_d maps $\mathcal{X}$ onto $\mathbf{\Sigma}$;
2. φ_d maps $\widehat{\mathbf{K}}_d$ homeomorphically onto $\widehat{\mathbf{S}}_d$;
3. outside $\widehat{\mathbf{K}}$, φ_d and $\widehat{\Pi}$ coincide.

Besides showing that $\mathcal{X}$ is a topological manifold, it suffices to verify Assertion (0) in order to prove the theorem.

Let $X \in \mathcal{A}$ be of dimension $\bar{d}$. Whitney regularity of $\mathcal{A}$ guarantees that X is open in $\mathbf{S}$; hence $\Pi^{-1}(X)$ is open in $\mathbf{K}$. A moment of reflection shows that the dimensions of the topological manifolds X and $\Pi^{-1}(X)$ coincide. According to Proposition 4.1, the Lagrange polytope has a constant dimension on X; hence this dimension is zero. Consequently, again in view of Proposition 4.1, we have $\Pi^{-1}(X) \subset \mathbf{K}^{\mathcal{G}_0}_{\mathcal{G}_0}$. That implies $\widehat{\Pi} : \widehat{\Pi}^{-1}(\widehat{X}) \to \widehat{X}$ to be a homeomorphism. Altogether, Assertion $(\bar{d})$ holds with $\varphi_{\bar{d}} = \widehat{\Pi}$.

Now assume Assertion (d) for some $d \in \{1, \ldots, \bar{d}\}$ and $X \in \mathcal{A}$ with $\dim X = d-1$. Since $\mathcal{A}$ is locally finite, it suffices to verify the following assertion in order to prove Assertion $(d-1)$.

Assertion (X). There exists a continuous mapping $\varphi : \widehat{\mathbf{K}} \to \widehat{\mathbf{S}}$ such that the following conditions are satisfied:

1. outside an arbitrarily small neighborhood of $\widehat{\Pi}^{-1}(\widehat{X})$, φ coincides with φ_d;
2. φ maps $\mathcal{X}$ onto $\mathbf{\Sigma}$;
3. φ maps $\widehat{\mathbf{K}}_X$ homeomorphically onto $\widehat{\mathbf{S}}_X$, where $\widehat{\mathbf{S}}_X := \widehat{\mathbf{S}}_d \cup \widehat{X}$ and $\widehat{\mathbf{K}}_X := \widehat{\Pi}^{-1}(\widehat{\mathbf{S}}_X)$;
4. outside $\widehat{\mathbf{K}}_X$, φ coincides with $\widehat{\Pi}$.

If the Lagrange polytopes have dimension zero on X, then we have finished by taking $\varphi = \varphi_d$. In the remainder of this paper we exclude this trivial case.

Step 1 (*fitting together special fibre and problem data*). The first step combines two projection mappings onto $\widehat{X}$. The first one, $\pi_1 : \widehat{\mathrm{Jet}_{\mathbf{\Sigma}}} \to \widehat{X}$, is defined in a neighborhood of $\widehat{X} \cap \xi^{-1}(0)$ and compatible with ξ; i.e., we have $\xi \circ \pi_1 = \xi$. This will prove that the approximation φ to be constructed maps $\mathcal{X}$ onto $\mathbf{\Sigma}$. The existence of π_1 follows from that of a tube at X which is compatible with ξ (cf. [4]).

The codimension of $\{0\}$ in $\mathrm{Jet}_{\mathbf{\Sigma}}$ is greater than zero; hence that of $\xi^{-1}(0)$ in $\widehat{X}$ is as well. Thus, we can assume a certain point $(\bar{z}, \bar{x}) \in \widehat{X}$ ($\bar{z} \in \mathbb{R}^{n+p}$, $\bar{x} \in X$) outside the open set on which π_1 is defined. X is stratum in a Whitney regular disk stratification. According to Proposition 4.1, it holds $X \subset \mathbf{S}^{\mathcal{G}_0}$ and the application of Proposition 3.4 yields a projection mapping $\pi_2 : \mathrm{Jet}_{\mathbf{\Sigma}} \to X$ (near $\bar{x}$) such that $\phi^{-1}\Pi^{-1}\pi_2^{-1}(\bar{x})$ is a smooth manifold. The second projection mapping onto $\widehat{X}$ is just $\widehat{\pi}_2$. It will provide the special fibre. In order to combine π_1 and $\widehat{\pi}_2$ we use a third projection mapping, π_3, defined in a neighborhood of the whole set $\widehat{X}$. By means of an appropriate partition of the unity we obtain a convex combination $\check{\pi} : \mathrm{Jet}_{\mathbf{\Sigma}} \to \widehat{X}$, which coincides with π_1 near $\xi^{-1}(0)$ and with $\widehat{\pi}_2$ near $(\bar{z}, \bar{x})$. Obviously, $\check{\pi}$ constitutes a projection mapping.

Step 2 (*application of the isotopy lemma*). Identify $\mathcal{G}_0$ and $\mathcal{J}$ with their (constant) images on X (in view of Proposition 4.1). The dimension of the affine space $\mathcal{A}\!f\!f_J(x) := (V|W)^{-1}(v) \cap (\mathbb{R}^{\mathrm{h}} \times \mathbb{R}^J)$ does not depend on the particular choice of $x \equiv (a, b, V, W, v)$ in X, provided that $J \in \mathcal{J}$. Hence $A_J := \{(x, \lambda, \mu) \mid x \in X, (\lambda, \mu) \in \mathcal{A}\!f\!f_J(x)\}$ is a manifold which contains $X_J := \Pi^{-1}(X) \cap \mathbf{K}_J^{\mathcal{G}_0}$ as an open subset. Finally, we have $A_{J'} \subset A_J$ for $J', J \in \mathcal{J}$ with $J' \subset J$. Altogether, $\{X_J \mid J \in \mathcal{J}\}$ gives a Whitney regular stratification of $\Pi^{-1}(X)$ refining $\{\mathbf{K}_J^I \mid J \subset I \subset \mathcal{G}\}$, which is a Whitney regular stratification of $\mathbf{K}$. Let $\mathcal{U}$ be an open set in $\widehat{\mathrm{Jet}_{\mathcal{X}}}$ containing $\widehat{\Pi}^{-1}(\widehat{X})$ as a closed subset. Then the following is a Whitney regular stratification of $\widehat{\mathrm{Jet}_{\mathcal{X}}} \cap \mathcal{U}$:

$$\begin{aligned} \mathcal{B} &:= \widetilde{\mathcal{B}} \cup \{\widehat{\mathbf{K}}_J^I \setminus \widehat{\Pi}^{-1}(\widehat{X}) \mid J \subset I \subset \mathcal{G}\}\Big|_{\mathcal{U}}, \text{ where} \\ \widetilde{\mathcal{B}} &:= \{\widehat{X}_J \mid J \in \mathcal{J}\}. \end{aligned}$$

$\widehat{\mathbf{K}} \cap \mathcal{U}$ and $\widehat{\Pi}^{-1}(\widehat{X}) = \bigcup \widetilde{\mathcal{B}}$ are closed subsets of $\mathcal{U}$. Restricting the mapping $\check{\pi} \circ \widehat{\Pi} : \mathcal{U} \to \widehat{X}$ to $\widehat{\Pi}^{-1}(\widehat{X})$ one obtains a proper mapping and, restricting it to the strata from $\widetilde{\mathcal{B}}$, a submersion. Since $\mathcal{A}$ is a disk stratification, it holds $\widehat{X} \equiv \mathbb{R}^m$ for some m. We define the special fibre $\bar{F} := \widehat{\Pi}^{-1}\check{\pi}^{-1}(\bar{z}, \bar{x})$. This fibre contains the polytope $\bar{P} := \widehat{\Pi}^{-1}(\bar{z}, \bar{x})$. After we shrink the open set $\mathcal{U}$ if necessary, an application of Proposition 2.9 yields a homeomorphism $h : \widehat{X} \times \bar{F} \to \widehat{\mathbf{K}} \cap \mathcal{U}$ with the following properties:

1. $\check{\pi} \circ \widehat{\Pi} \circ h = \Pi_{\widehat{X}}$, the natural projection;
2. $h : \widehat{X} \times (\bar{F} \cap Y) \to Y$ for any $Y \in \mathcal{B}$.

Since h preserves strata, it maps $\widehat{X} \times \bar{P}$ onto $\widehat{\Pi}^{-1}(\widehat{X})$. In order to see that $\mathcal{X}$ is a topological manifold, recall that $\mathcal{X} = \widehat{\mathbf{K}} \cap \widehat{\Pi}^{-1}\xi^{-1}(0)$ and that $\xi \circ \widehat{\Pi} \circ h$ coincides with $\xi \circ \Pi_{\widehat{X}}$ (where defined). (The latter relation is in view of 1. above and Step 1.) Hence $h^{-1}(\mathcal{X}) = \big(\widehat{X} \cap \xi^{-1}(0)\big) \times \bar{F}$. For problem data from $\mathcal{F}$, $\widehat{X} \cap \xi^{-1}(0)$ is a manifold. We will see that $\widehat{\phi}(\bar{F})$ is a topological manifold and thus $\mathcal{X}$ is as well.

Step 3 (*tame explosion*). Recall the construction of $\check{\pi}$ near the point $(\bar{z}, \bar{x})$ made in Step 1. We have $\check{\pi}^{-1}(\bar{z}, \bar{x}) = \{\bar{z}\} \times {\pi_2}^{-1}(\bar{x})$. Thus, $\widehat{\phi}^{-1}(\bar{F})$ is a smooth manifold which contains the polytope $\widehat{\phi}^{-1}(\bar{P})$. Using the homeomorphism $\widehat{\phi}$, Proposition 3.2 yields a controlled explosion ψ of $\bar{P}$ in $\bar{F}$. Let $\bar{y}$ denote the centre of explosion and T the trace of $\bar{y}$ under application of h, defined by $T := h(\widehat{X}, \{\bar{y}\})$. Also, put $\partial\widehat{X} := \overline{\widehat{X}} \setminus \widehat{X}$.

Let $\varepsilon : \widehat{X} \to (0,1)$ be a continuous mapping such that $\varepsilon(x_i)$ tends to zero for any sequence $\{x_i\} \subset \widehat{X}$ converging to some point in $\partial\widehat{X}$. There exists a continuous mapping $\delta : \widehat{X} \to (0,1)$ such that h maps $\{x\} \times U_{4\delta(x)}(\bar{P})$ into $U_{\varepsilon(x)}\widehat{\Pi}^{-1}(x)$. The fact that ε can be chosen arbitrarily small proves the approximation property claimed in Assertion (X). Let denote $U := \bigcup_{x \in \widehat{X}} h\big[\, \{x\} \times U_\varepsilon(\bar{P}) \,\big]$ and $U' := \bigcup_{x \in \widehat{X}} h\big[\, \{x\} \times$

$U_{3\delta(x)}(\bar{P})\,]$. The sets U and U' are open subsets of $\widehat{\mathbf{K}}$, and it holds $\overline{U}'\setminus U \subset \widehat{\Pi}^{-1}(\partial\widehat{X})$. Define an explosion $\Psi : U \setminus T \to U \setminus \widehat{\Pi}^{-1}(\widehat{X})$ by $\Psi := h \circ (\mathrm{id}_{\widehat{X}} \times \psi_\delta) \circ h^{-1}$. Then

(T1) $\Psi|_{\widehat{\mathbf{K}}_X}$ is a homeomorphism;

(T2) outside U', Ψ and $\mathrm{id}_{\widehat{\mathbf{K}}}$ coincide;

(T3) $\widehat{\Pi} \circ \Psi$ extends continuously onto $T \cup \widehat{\Pi}^{-1}(\partial\widehat{X})$ by $\widehat{\Pi}$;

(T4) $\xi \circ \widehat{\Pi} \circ \Psi = \xi \circ \widehat{\Pi}$, where defined.

In fact, Ψ operates within the fibres $h(x, \bar{F}) = \widehat{\Pi}^{-1}\check{\pi}^{-1}(x)$, and we also have $\|\, \widehat{\Pi} \circ \Psi - \widehat{\Pi}\, \| < \varepsilon \circ \check{\pi} \circ \widehat{\Pi}$. If we use $\mathrm{id}_{\widehat{\mathbf{K}}}$, Ψ extends continuously to $\Psi : \widehat{\mathbf{K}} \setminus (T \cup \widehat{\Pi}^{-1}(\partial\widehat{X})) \to \widehat{\mathbf{K}}\setminus\widehat{\Pi}^{-1}(\widehat{\widehat{X}})$, being globally defined and satisfying (T1), (T3), and (T4).

Step 4 (*perturbed extension of the homeomorphism*). Define

$$\varphi := \begin{cases} \varphi_d & \text{on } T \cup \widehat{\Pi}^{-1}(\partial\widehat{X}), \\ \varphi_d \circ \Psi & \text{elsewhere.} \end{cases}$$

By the very construction it suffices to check that φ is continuous at $\tilde{y} \in \widehat{\Pi}^{-1}(\partial\widehat{X})$. To this end assume a sequence $y^k \in \widehat{\mathbf{K}}$ converging to $\tilde{y}$. Assume the nontrivial case that $y^k \notin T \cup \widehat{\Pi}^{-1}(\partial\widehat{X})$. By (T3), $\widehat{\Pi} \circ \Psi(y^k)$ tends to $\widehat{\Pi}(\tilde{y})$. Since the family of Lagrange polytopes is upper semicontinuous, $\Psi(y^k)$ is arbitrarily close to the polytope $\widehat{\Pi}^{-1}(\widehat{\Pi}(\tilde{y}))$, provided that k is sufficiently large. Hence, $\varphi_d \circ \Psi(y^k)$ converges to $\widehat{\Pi}(\tilde{y}) = \varphi_d(\tilde{y})$. This completes the proof of the theorem. □

Acknowledgments. The author is kindly indebted to B. Bank, J. Guddat, and H. Th. Jongen for valuable suggestions.

REFERENCES

[1] J. Bochnak, M. Coste, and M.-F. Roy, *Géometrie Algébrique Réelle*, Springer-Verlag, Berlin, 1987.

[2] R. J. Daverman, *Decompositions of Manifolds*, Academic Press, London, 1986.

[3] J. Gauvin, *A necessary and sufficient regularity condition to have bounded multipliers in nonconvex programming*, Math. Programming, 12 (1977), pp. 136–138.

[4] C. G. Gibson, K. Wirthmüller, A. A. du Plessis, and E. J. N. Looijenga, *Topological Stability of Smooth Mappings*, Lecture Notes in Mathematics 552, Springer-Verlag, Berlin, 1976.

[5] H. Günzel, *The Crease Structure of the Karush-Kuhn-Tucker Set in Parametric Optimization*, Reports of the Faculty of Technical Mathematics and Informatics 92-62, Delft University of Technology, 1992.

[6] H. Günzel, R. Hirabayashi, H. Th. Jongen, and S. Shindoh, *A note on the stratification of the Karush–Kuhn–Tucker set*, in Parametric Optimization and Related Topics III, J. Guddat, H.Th. Jongen, B. Kummer and F. Nožička, eds., Peter Lang Verlag, Frankfurt am Main, 1993, pp. 215–225.

[7] R. Hirabayashi, M. Shida, and S. Shindoh, *Manifold structure of the Karush–Kuhn–Tucker stationary solution set with two parameters*, SIAM J. Optim., 3 (1993), pp. 564–581.

[8] H. Th. Jongen, P. Jonker, and F. Twilt, *Nonlinear Optimization in $\mathbb{R}^n$*, II. *Transversality, Flows, Parametric Aspects*, Peter Lang Verlag, Frankfurt am Main, 1986.

[9] ———, *Parametric optimization: The Kuhn-Tucker set*, in Parametric Optimization and Related Topics, J. Guddat, H. Th. Jongen, B. Kummer and F. Nožička, eds., Akademie-Verlag, Berlin, 1987, pp. 196–208.

[10] M. Kojima and R. Hirabayashi, *Continuous deformation of nonlinear programs*, Mathematical Programming Study, Vol. 21 (1984), pp. 150-198.

[11] S. M. Robinson, *Generalized equations and their solutions, Part* II: *Applications to nonlinear programming*, Mathematical Programming Study, 19 (1982), pp. 200–221.

[12] S. Schecter, *Structure of the first order solution set for a class of nonlinear programs with parameters*, Math. Programming, 34 (1986), pp. 84–110.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 33, No. 6, pp. 1857–1880, November 1995
© 1995 Society for Industrial and Applied Mathematics
013

PONTRYAGIN MAXIMUM PRINCIPLE FOR SEMILINEAR AND QUASILINEAR PARABOLIC EQUATIONS WITH POINTWISE STATE CONSTRAINTS*

BEI HU† AND JIONGMIN YONG‡

Abstract. This paper studies the first-order necessary conditions for the optimal controls of semilinear and quasilinear parabolic partial differential equations with pointwise state constraints. A Pontryagin-type maximum principle is obtained.

Key words. parabolic equations, pointwise state constraints, optimal control, Pontryagin maximum principle, Ekeland variational principle

AMS subject classifications. 49K20, 35K10, 35K20

1. Introduction.

In this paper, we are concerned with the following parabolic equation:

$$\left\{\begin{array}{ll} y_t - \sum_{i,j=1}^{n} (a_{ij}(x,t)y_{x_i})_{x_j} = f(x,t,y,u(t,x)), & \text{in } \Omega_T, \\ y\,\big|_{\partial\Omega} = 0, & \\ y\,\big|_{t=0} = y_0(x), \qquad x \in \Omega, & \end{array}\right. \tag{1.1}$$

where a_{ij} and f are some given functions, $\Omega_T = \Omega \times (0,T)$, with $\Omega \subset \mathbb{R}^n$ being a bounded domain and $T > 0$ being a given time duration. The function $u(x,t)$ is called the control, which takes value in some separable metric space U. The solution $y(x,t)$ to (1.1) (for given $y_0(x)$ and $u(x,t)$) is called the state of the system, and $y_0(x)$ is referred to as the initial state. We set $\mathcal{U} = \{u : \Omega_T \to U \mid u \text{ is measurable }\}$. Under proper conditions (see §2), we have that for any $y_0 \in C_0^\alpha(\bar\Omega)$ and $u \in \mathcal{U}$, (1.1) admits a unique solution $y(x,t)$ which is in $C(\bar\Omega_T)$ (actually, it is even better; see §2). Then, we may talk about the state constraint of form,

$$G(y) \in Q, \tag{1.2}$$

for some continuously Fréchet differentiable map $G : C_0(\bar\Omega_T) \to Z$, where $C_0(\bar\Omega_T) = \{\eta \in C(\bar\Omega_T) \mid \eta\big|_{\partial\Omega\times[0,T]} = 0\}$, Z is some Banach space, and $Q \subset Z$. Let us take a look at two important examples of the above type of constraint. First, let $Z = C_0(\bar\Omega_T)$,

$$Q = \{\eta \in Z \mid \eta(x,t) \le 0, \;\; (x,t) \in \bar\Omega_T \}, \tag{1.3}$$

and $G(y)(x,t) = g(x,t,y(x,t))$ for some function $g : \bar\Omega_T \times \mathbb{R} \to \mathbb{R}$. Then, (1.2) reads

$$g(x,t,y(x,t)) \le 0, \qquad (x,t) \in \bar\Omega_T. \tag{1.4}$$

*Received by the editors June 9, 1993; accepted for publication (in revised form) June 24, 1994.

†Department of Mathematics, University of Notre Dame, Notre Dame, Indiana 46556. The research of this author was partially supported by National Science Foundation grants DMS 90-24986 and DMS 92-24935.

‡Department of Mathematics, Fudan University, Shanghai 200433, China. The research of this author was partially supported by the National Science Foundation of China and the Fok Ying Tung Education Foundation. Part of this research was performed while this author was visiting the Institute for Mathematics and its Applications, University of Minnesota.

Our second example is as follows: We again let $Z = C_0(\bar{\Omega}_T)$, and define

$$Q = \{\eta \in Z \mid \eta(x_i, t_i) = b_i, \quad 1 \leq i \leq m\} \tag{1.5}$$

for some given (different) points $(x_i, t_i) \in \Omega_T \bigcup(\Omega \times \{T\})$ and numbers b_i, and $G = I$, the identity operator on Z. Then, (1.2) reads

$$y(x_i, t_i) = b_i, \qquad 1 \leq i \leq m. \tag{1.6}$$

For elliptic equations, similar constraints like (1.4) and (1.6) were considered in [4], [5], and [19]. The above two examples all require pointwise behavior of the state $y(x, t)$. There are many other examples covered by (1.2) (see §5). It is seen that our state constraint (1.2) is very general.

Now, we introduce the following functional:

$$J(u) = \int_{\Omega_T} f^0(x, t, y(x, t), u(x, t))dxdt, \tag{1.7}$$

for some function f^0, where $y(x, t)$ is the solution of (1.1) corresponding to u. This is called the cost functional. Next, we set

$$\mathcal{U}_{\text{ad}} \equiv \{u \in \mathcal{U} \mid \text{the corresponding } y \text{ satisfies (1.2)}\}. \tag{1.8}$$

Any element $u \in \mathcal{U}_{\text{ad}}$ is called an admissible control. In what follows, we assume that $\mathcal{U}_{\text{ad}} \neq \emptyset$. Then, we may state our optimal control problem as follows.

PROBLEM C. Find $\bar{u} \in \mathcal{U}_{\text{ad}}$, such that

$$J(\bar{u}) = \inf_{\mathcal{U}_{\text{ad}}} J(u). \tag{1.9}$$

Whenever such a $\bar{u} \in \mathcal{U}_{\text{ad}}$ exists, we call it an optimal control; the corresponding state $\bar{y}$ is called an optimal state and $(\bar{y}, \bar{u})$ is called an optimal pair.

Our goal is to obtain a set of first-order necessary conditions for the optimal pairs. This set of conditions is called the Pontryagin maximum principle.

In recent papers [3]–[5], [20], the Pontryagin maximum principle was derived for semilinear and quasilinear elliptic partial differential equations with pointwise state constraints (see [2] also). For parabolic equations, an abstract evolution equation setting was used a little earlier to obtain similar results [9], [10], [14], [15], [18]. We note that by using the abstract framework for parabolic equations, people treat the time variable t and the spatial variable x unequally, in the sense that the variable x is "averaged" and actually does not appear explicitly in the whole process. Consequently, some pointwise information on the state $y(x, t)$, like Hölder continuity, and values at some particular points $(x_0, t_0) \in \Omega_T$ are lost. In particular, the problem with the state constraint (1.6) cannot be covered by abstract framework. In this paper, we use the idea of [5] (see [13] and [20] also) to discuss the optimal control problem for parabolic equations without using the abstract evolution equations. By this approach, we retain some pointwise behavior of the state $y(x, t)$. Consequently, we can treat general (pointwise) state constraint (1.2), which contains (1.6) as a special case.

Due to the fact that U is just a separable metric space, only the spike perturbation of the control is allowed when we derive the necessary conditions. On the other hand, the pointwise state constraint is presented. These two together cause the main difficulty in our procedure. The key which overcomes this main difficulty is to find the

"Taylor expansion" of the state with respect to the spike variation of the control, in a strong enough topology that is sufficient for us to treat the pointwise constraint. We achieve this by improving a technical lemma found in [5] and using proper estimates for parabolic equations. Once this is obtained, we then use the usual procedure of applying Ekeland's variational principle to derive the desired conclusion.

We refer readers to [17] and [19] for some classical relevant results.

The rest of the paper is organized as follows. In §2, we give some preliminary results and state the main result. Section 3 is devoted to proving some technical lemmas. The proof of the Pontryagin maximum principle is carried out in §4. Some applications are given in §5. The result corresponding to quasilinear parabolic equations is briefly discussed in §6.

2. Preliminary results and the main result. Let us first give some assumptions and preliminary results. We let $\Omega \subset \mathbb{R}^n$ be a bounded domain with $\partial\Omega \in C^2$, $\Omega_T = \Omega \times (0,T)$, $\partial_p\Omega_T = (\bar{\Omega} \times \{0\}) \bigcup (\partial\Omega \times [0,T])$ be the parabolic boundary of Ω_T, and U be a separable metric space. We use $|\cdot|$ or $\|\cdot\|$ (sometimes with a subscript) as the norm in various spaces, which can be identified from the context. For any measurable set $S \subset \mathbb{R}^n$, we use $|S|$ to denote the Lebesgue measure of the set S. In what follows, we will denote by $C_0(\bar{\Omega}_T) \subset C(\bar{\Omega}_T)$ the set of all continuous functions on $\bar{\Omega}_T$ which vanish on $\partial\Omega \times [0,T]$, by $C^{\beta,\beta/2}(\bar{\Omega}_T)$ the set of all continuous functions on $\bar{\Omega}_T$ which are Hölder continuous in (x,t) with the exponent β in x and $\beta/2$ in t ($\beta \in (0,1)$), and by $C_0(\bar{\Omega})$ the set of all continuous functions on $\bar{\Omega}$ which vanish on $\partial\Omega$.

The following assumptions will be assumed throughout the paper.

(A1) The function $a_{ij} : \Omega_T \to \mathbb{R}$ is measurable, $a_{ij} = a_{ji}$, and there exist constants $\Lambda > \lambda > 0$, such that

$$\lambda|\xi|^2 \leq \sum_{i,j=1}^{n} a_{ij}(x,t)\xi_i\xi_j \leq \Lambda|\xi|^2, \qquad \text{for a.e. } (x,t) \in \bar{\Omega}_T, \quad \xi \in \mathbb{R}^n. \tag{2.1}$$

(A2) The function $f : \bar{\Omega} \times [0,T] \times \mathbb{R} \times U \to \mathbb{R}$ has the following properties: $f(\cdot,\cdot\ ,y,u)$ is measurable on $\Omega \times [0,T]$, $f(x,t,\cdot,u)$ is in $C^1(\mathbb{R})$ with $f(x,t,\cdot,\cdot)$ and $f_y(x,t,\cdot,\cdot)$ being continuous on $\mathbb{R} \times U$. There exists a constant $C > 0$, such that

$$f(x,t,y,u)y \leq C(|y|^2 + 1), \qquad \forall (x,t,y,u) \in \bar{\Omega} \times [0,T] \times \mathbb{R} \times U. \tag{2.2}$$

Moreover, for any $R > 0$, there exists an $M_R > 0$, such that

$$|f(x,t,y,u)| + |f_y(x,t,y,u)| \leq M_R, \qquad \forall (x,t,u) \in \bar{\Omega} \times [0,T] \times U, |y| \leq R. \tag{2.3}$$

The same conditions, except (2.2), hold for the function $f^0 : \bar{\Omega} \times [0,T] \times \mathbb{R} \times U \to \mathbb{R}$.

(A3) Z is a Banach space with the dual Z^* being strictly convex. $Q \subset Z$ is convex and closed, and is of finite codimension in Z (see below or [15, Def. 2.2]). The map $G : C_0(\bar{\Omega}_T) \to Z$ is continuously Fréchet differentiable.

Let us make some remarks on (A3). First, a set $Q \subset Z$ is said to be finite codimensional in Z if for some $z_0 \in Q$, the space Z_0 spanned by $Q - z_0 \equiv \{z - z_0 \mid z \in Q\}$ is a finite codimensional subspace of Z and the convex hull $\overline{\text{co}}(Q - z_0)$ of $Q - z_0$ has a nonempty relative interior in Z_0. It is not hard to see that the set Q defined by (1.3) has a nonempty interior in Z and hence is of codimension 0 in Z; the set Q defined by (1.5) is of codimension m in Z.

Next, we consider the case $Z = C_0(\bar{\Omega}_T)$. Then, by the Hahn–Banach Theorem, for any $\mu \in Z^* \equiv C_0(\bar{\Omega}_T)^*$, there exists a $\widetilde{\mu} \in C(\bar{\Omega}_T)^* = \mathcal{M}(\bar{\Omega}_T)$ (the set of all Radon measures on $\bar{\Omega}_T$), such that $\mu = \widetilde{\mu}\big|_{C_0(\bar{\Omega}_T)}$ and

$$(2.4) \qquad \langle\, \widetilde{\mu}, \eta\, \rangle = \int_{\bar{\Omega}_T} \eta d\widetilde{\mu}, \qquad \forall \eta \in C(\bar{\Omega}_T).$$

Then, for any $\eta \in C_0(\bar{\Omega}_T)$ (note $\eta\big|_{\partial\Omega} = 0$)

$$(2.5) \qquad \langle\, \mu, \eta\, \rangle = \int_{\bar{\Omega}_T} \eta d\mu = \int_{\Omega_T \bigcup (\Omega \times \{0,T\})} \eta d\mu.$$

In what follows, we let $\mathcal{M}_0(\bar{\Omega}_T)$ be the set of all Radon measures on $\bar{\Omega}_T$ with the support contained in $\Omega_T \bigcup (\Omega \times \{0,T\})$. Clearly, $\mathcal{M}_0(\bar{\Omega}_T) = C_0(\bar{\Omega}_T)^*$, with the identification being (2.5). It is known that if we use the usual norm in $C_0(\bar{\Omega}_T)$, the dual $C_0(\bar{\Omega}_T)^*$ of it is not strictly convex. However, since the space $C_0(\bar{\Omega}_T)$ is a separable Banach space, by [7, p. 167], there exists a norm, denoted by $|\cdot|_0$, which is equivalent to the norm $\|\cdot\|_{C_0(\bar{\Omega}_T)}$, such that the dual of $(C_0(\bar{\Omega}_T), |\cdot|_0)$ is strictly convex. It is clear that any element $\mu \in (C_0(\bar{\Omega}_T), |\cdot|_0)^*$ can still be identified with an element of $\mathcal{M}_0(\bar{\Omega}_T)$, such that (2.5) holds. This will be useful when we discuss the case with $Z = C(\bar{\Omega}_T)$ (see §5).

Next, we define

$$(2.6) \qquad d_Q(\eta) = \inf_{\eta \in Q} |z - \eta|, \qquad \forall \eta \in Z.$$

Then, $d_Q : Z \to \mathbb{R}$ is convex and Lipschitz continuous (with the Lipschitz constant being 1). From [6], we know that the Clarke's generalized gradient, denoted by ∂d_Q, which coincides with the subdifferential in the sense of the convex analysis in this case [6, Prop. 2.2.7], is convex and weak*-compact. Therefore, given $\xi \in \partial d_Q(\eta)$, we have that

$$(2.7) \qquad \langle\, \xi, z - \eta\, \rangle + d_Q(\eta) \le d_Q(z), \qquad \forall z \in Z.$$

This implies that $|\langle\, \xi, z - \eta\, \rangle| \le |z - \eta|$, for all $z \in Z$, since $d_Q(\cdot)$ is Lipschitz continuous with Lipschitz constant 1. Thus, $\|\xi\|_{Z^*} \le 1$. The identity $\|\xi\|_{Z^*} = 1$ is true whenever $\eta \notin Q$; see [15, Lem. 3.4]. Since Z^* is strictly convex, $\partial d_Q(\eta)$ is a singleton for every $\eta \notin Q$ [15, Cor. 3.5]. Furthermore, $d_Q : Z \to \mathbb{R}$ is Gâteaux differentiable at every point $\eta \notin Q$ and $\{\nabla d_Q(\eta)\} = \partial d_Q(\eta)$ [6, Prop. 2.2.4], where $\nabla d_Q(\eta)$ is the Gâteaux derivative of $d_Q(\eta)$ at η. Hence

$$(2.8) \qquad \|\nabla d_Q(\eta)\|_{Z^*} = 1, \qquad \forall \eta \notin Q.$$

The following result is basic.

PROPOSITION 2.1. *Let* (A1)–(A2) *hold. Then, for any* $u \in \mathcal{U}$ *and* $y_0 \in C^\alpha(\bar{\Omega}) \bigcap C_0(\bar{\Omega})$ $(0 < \alpha < 1)$*, there exists a* $\beta \in (0,1)$*, such that* (1.1) *has a unique solution* $y \equiv y(\cdot,\cdot\,;u) \in C^{\beta,\beta/2}(\bar{\Omega}_T) \bigcap L^2(0,T;H_0^1(\Omega))$*. Furthermore, there exists a constant* $C > 0$*, independent of* $u \in \mathcal{U}$*, such that*

$$(2.9) \qquad \|y(\cdot,\cdot\,;u)\|_{C^{\beta,\beta/2}(\bar{\Omega}_T)} \le C, \qquad \forall u \in \mathcal{U}.$$

Sketch of the proof. Uniqueness follows immediately from the energy estimates. For the existence, it suffices to establish the a priori estimates for the solution. The assumption (2.2) immediately gives us the $L^\infty(\Omega_T)$ estimates. Then the standard energy inequality gives $L^2(0,T;H^1(\Omega))$ estimates, and the existence of the solution follows. The estimates (2.9) is standard and can be found in [12, Chap. III, §10]. □

In what follows, any pair $(y,u) \in \big(C^{\beta,\beta/2}(\bar{\Omega}_T)\bigcap C_0(\bar{\Omega}_T)\big) \times \mathcal{U}$ satisfying (1.1) is called a feasible pair and we refer to the corresponding y and u as feasible state and control, respectively. Clearly, under (A1)–(A2), $\mathcal{U}$ coincides with the set of all feasible controls and for each feasible control $u \in \mathcal{U}$ there corresponds a unique feasible state. Also, we see that the cost functional $J(u)$ is well defined for each $u \in \mathcal{U}$ and the state constraint (1.2) clearly makes sense.

Now, we assume that the set $\mathcal{U}_{\text{ad}}$ defined in (1.8) is nonempty and there exists an optimal pair $(\bar{y},\bar{u})$ to Problem C. Our main result then can be stated as follows.

THEOREM 2.2 (Maximum Principle). *Let (A1)– (A3) hold and let the following compatibility condition, for the set Q, the map G, and the initial state y_0, hold:*

$$\begin{aligned}&\operatorname{supp} G'(\eta)^*\partial d_Q(G(\eta)) \subset \Omega_T \bigcup(\Omega \times \{T\}),\\ &\qquad\qquad \forall \eta \in C_0(\bar{\Omega}_T) \textit{ with } G(\eta) \in Q, \quad \eta\big|_{t=0} = y_0(x).\end{aligned} \tag{2.10}$$

Let $(\bar{y},\bar{u})$ be an optimal pair of Problem C. Then, there exists a constant $\psi^0 \le 0$, a function $\psi \in L^q(0,T;W_0^{1,q}(\Omega))$ $(1 < q < \frac{n+2}{n+1})$, and a $\varphi \in \partial d_Q(G(\bar{y})) \subset Z^*$, such that

$$|\psi^0| + \|\varphi\|_{Z^*} > 0, \tag{2.11}$$

$$\begin{cases} \psi_t + \sum\limits_{i,j=1}^n (a_{ij}(x,t)\psi_{x_j})_{x_i} = -f_y(x,t,\bar{y}(x,t),\bar{u}(x,t))\psi \\ \qquad\qquad - \psi^0 f_y^0(x,t,\bar{y}(x,t),\bar{u}(x,t)) + (G'(\bar{y})^*\varphi)\big|_{\Omega_T}, \qquad \textit{in } \Omega_T, \\ \psi\ \big|_{\partial\Omega} = 0, \\ \psi\ \big|_{t=T} = (G'(\bar{y})^*\varphi)\big|_{\Omega\times\{T\}}, \end{cases} \tag{2.12}$$

$$\langle\, z - G(\bar{y}), \varphi\,\rangle \le 0, \qquad \forall z \in Q. \tag{2.13}$$

$$\begin{aligned}&H(x,t,\bar{y}(x,t),\bar{u}(x,t),\psi^0,\psi(x,t)) = \max_{v\in U} H(x,t,\bar{y}(x,t),v,\psi^0,\psi(x,t)),\\ &\qquad\qquad a.e.(x,t) \in \Omega \times [0,T],\end{aligned} \tag{2.14}$$

where

$$\begin{aligned}&H(x,t,y,u,\psi^0,\psi) =\psi^0 f^0(x,t,y,u) + \psi f(x,t,y,u),\\ &\qquad\qquad \forall (x,t,y,u,\psi^0,\psi) \in \Omega \times [0,T] \times \mathbb{R} \times U \times \mathbb{R} \times \mathbb{R}.\end{aligned} \tag{2.15}$$

In the above, (2.12) is called the adjoint equation, (2.13) is called the transversality condition, and (2.14) is called the maximum condition. It will be seen that in the proof, we only need (2.10) to hold for $\eta = \bar{y}$. Also, in §5, we will give some examples for which such a compatible condition holds.

3. Some technical lemmas. In order to derive the first-order necessary conditions for optimal pairs, we need some sort of "directional derivatives" of the state y and the cost functional $J(u)$ in the control variable u. However, since the control domain U is just a metric space and there is no convexity in general, the perturbation of the control variable is restricted to be of "spike" type. Thus, to find the "directional derivative" is not obvious. In this section, we will present some technical lemmas which will give us exactly the "directional derivatives" we need in the proof of the maximum principle in §4. The results of this section are comparable with those in [5] for elliptic equations (see [13]–[15] and [20] also).

LEMMA 3.1. *Let $h^0 \in L^1(\Omega)$ and $h \in L^p(\Omega)$, $1 < p < \infty$. For any $\rho \in (0,1)$, we define*

$$\mathcal{E}_\rho = \{E \subset \Omega \mid E \text{ measurable and } |E| = \rho|\Omega|\}. \tag{3.1}$$

Let $\mathcal{Y}$ be a Banach space with the embedding $\mathcal{Y} \hookrightarrow L^{p'}(\Omega)$ being compact ($p' = \frac{p}{p-1}$). Then,

$$\inf_{E \in \mathcal{E}_\rho} \left\{ \left| \int_\Omega \left(1 - \frac{1}{\rho}\chi_{E_\rho}(x)\right) h^0(x)dx \right| + \left\| \left(1 - \frac{1}{\rho}\chi_{E_\rho}\right) h \right\|_{\mathcal{Y}^*} \right\} = 0. \tag{3.2}$$

Proof. Let $\rho \in (0,1)$ be given and let $\delta > 0$ be arbitrary. We let B be the closed unit ball in $\mathcal{Y}$. This set is compact in $L^{p'}(\Omega)$ by our assumption. Thus, we can find a set of finitely many step functions $\Theta \equiv \{\theta_i, 1 \le i \le r\}$, such that for any $y \in B$, there exists a $\theta_i \in \Theta$ satisfying

$$\|y - \theta_i\|_{L^{p'}(\Omega)} < \delta. \tag{3.3}$$

Since Θ is a finite set, we may let $\{K_j\}_{j=1}^m$ be a partition of Ω with $|K_j| > 0$ for each $1 \le j \le m$, such that

$$\theta_i(x) = \sum_{j=1}^m c_{ij}\chi_{K_j}(x), \qquad x \in \Omega, \quad 1 \le i \le r. \tag{3.4}$$

Then, for any $y \in B$, by choosing $\theta_i \in \Theta$ with the property (3.3), we have

$$\begin{aligned}
&\sum_{j=1}^m \int_{K_j} \left| y(x) - \frac{1}{|K_j|}\int_{K_j} y(\xi)d\xi \right|^{p'} dx \\
&\le 3^{p'-1} \sum_{j=1}^m \left\{ \int_{K_j} |y(x) - \theta_i(x)|^{p'}dx + \int_{K_j} \left| \theta_i(x) - \frac{1}{|K_j|}\int_{K_j} \theta_i(\xi)d\xi \right|^{p'} dx \right. \\
&\qquad \left. + \int_{K_j} \left| \frac{1}{|K_j|}\int_{K_j} (y(\xi) - \theta_i(\xi))d\xi \right|^{p'} dx \right\} \\
&\le 3^{p'-1} \left\{ \int_\Omega |y(x) - \theta_i(x)|^{p'}dx \right. \\
&\qquad \left. + \sum_{j=1}^m \frac{1}{|K_j|^{p'}} \int_{K_j} |K_j|^{p'/p} \int_{K_j} |y(\xi) - \theta_i(\xi)|^{p'}d\xi dx \right\} \\
&\le 3^{p'-1} \left\{ \delta^{p'} + \int_\Omega |y(\xi) - \theta_i(\xi)|^{p'}d\xi \right\} \le 2 \cdot 3^{p'-1}\delta^{p'}.
\end{aligned} \tag{3.5}$$

Here, we have used the fact that $\theta_i(x)$ is a constant on each set K_j. We have seen that the above estimate is uniform for $y \in B$.

Now, for any $y \in B$, let us define $\widetilde{y} : \Omega \to \mathbb{R}$ to be the following:

$$\widetilde{y}(x) = \frac{1}{|K_j|}\int_{K_j} y(\xi)d\xi, \qquad x \in K_j, \qquad 1 \le j \le m. \tag{3.6}$$

Then, (3.5) can be written as (by setting $\varepsilon^{p'} = 2 \cdot 3^{p'-1}\delta^{p'}$, which is still arbitrary)

$$\|y - \widetilde{y}\|_{L^{p'}(\Omega)} < \varepsilon, \qquad y \in B. \tag{3.7}$$

Next, on each K_j, we approximate the functions h^0 and h by step functions:

$$h_j^0(x) = \sum_{i=1}^{r_j} \alpha_{ij}\chi_{F_{ij}}(x), \quad h_j(x) = \sum_{i=1}^{r_j} \beta_{ij}\chi_{F_{ij}}(x), \qquad x \in \Omega, \tag{3.8}$$

with $\alpha_{ij}, \beta_{ij} \in \mathbb{R}$, $\{F_{ij}\}_{i=1}^{r_j}$ being a partition of K_j, $|F_{ij}| > 0$, and such that

$$\int_{K_j} |h^0(x) - h_j^0(x)|dx + \int_{K_j} |h(x) - h_j(x)|dx < \varepsilon|K_j|, \qquad 1 \le j \le m. \tag{3.9}$$

Let us take $E_\rho^{ij} \subset F_{ij}$, such that $|E_\rho^{ij}| = \rho|F_{ij}|$. Set $E_\rho^j = \bigcup_{i=1}^{r_j} E_\rho^{ij}$. Since $h_j^0(x)$ and $h_j(x)$ are simple functions, we have

$$\left\{\begin{aligned} &\int_{K_j} h_j^0(x)dx = \int_{K_j} \frac{1}{\rho}\chi_{E_\rho^j}(x)h_j^0(x)dx,\\ &\int_{K_j} h_j(x)dx = \int_{K_j} \frac{1}{\rho}\chi_{E_\rho^j}(x)h_j(x)dx, \end{aligned}\right. \qquad 1 \le j \le m. \tag{3.10}$$

Finally, we take $E_\rho = \bigcup_{j=1}^m E_\rho^j$. Then, $E_\rho \in \mathcal{E}_\rho$, and for any $y \in B$,

$$\begin{aligned} \left|\int_\Omega \left(1 - \frac{1}{\rho}\chi_{E_\rho}\right) h(x)y(x)dx\right| &\le \left|\int_\Omega \left(1 - \frac{1}{\rho}\chi_{E_\rho}\right) h(x)\widetilde{y}(x)dx\right| \\ &\quad + \int_\Omega \left(1 + \frac{1}{\rho}\chi_{E_\rho}\right) |h(x)|\,|y(x) - \widetilde{y}(x)|dx \\ &\le \left|\sum_{j=1}^m \int_{K_j} \left(1 - \frac{1}{\rho}\chi_{E_\rho^j}(x)\right) h(x)\widetilde{y}(x)dx\right| + \left(1 + \frac{1}{\rho}\right) \|h\|_{L^p(\Omega)}\|y - \widetilde{y}\|_{L^{p'}(\Omega)}. \end{aligned} \tag{3.11}$$

From (3.7) we see that

$$\left(1 + \frac{1}{\rho}\right) \|h\|_{L^p(\Omega)}\|y - \widetilde{y}\|_{L^{p'}(\Omega)} \le \varepsilon\left(1 + \frac{1}{\rho}\right) \|h\|_{L^p(\Omega)}. \tag{3.12}$$

On the other hand, we notice that $\widetilde{y}(x)$ is a constant on each set K_j (we denote this constant by $\widetilde{y}(K_j)$). Thus, by (3.9)–(3.10) and (3.6), we have

$$
\begin{aligned}
&\left|\sum_{j=1}^{m}\int_{K_j}\left(1-\frac{1}{\rho}\chi_{E_\rho^j}(x)\right)h(x)\widetilde{y}(x)dx\right| \\
&\le \sum_{j=1}^{m}|\widetilde{y}(K_j)|\int_{K_j}\left(1+\frac{1}{\rho}\chi_{E_\rho^j}(x)\right)|h(x)-h_j(x)|dx \\
&\quad+\left|\sum_{j=1}^{m}\widetilde{y}(K_j)\int_{K_j}\left(1-\frac{1}{\rho}\chi_{E_\rho^j}\right)h_j(x)dx\right| \\
&\le\left(1+\frac{1}{\rho}\right)\sum_{j=1}^{m}\varepsilon|K_j|\,|\widetilde{y}(K_j)|\le\varepsilon\left(1+\frac{1}{\rho}\right)\|y\|_{L^1(\Omega)}\le\varepsilon\left(1+\frac{1}{\rho}\right)C.
\end{aligned}
\tag{3.13}
$$

Here, $\|y\|_{L^1(\Omega)}\le C\|y\|_{\mathcal{Y}}\le C$, since $y\in B$. Thus, (3.11)–(3.13) imply that

$$
\left\|\left(1-\frac{1}{\rho}\chi_{E_\rho}\right)h\right\|_{\mathcal{Y}^*}\le\varepsilon\left(1+\frac{1}{\rho}\right)(C+\|h\|_{L^p(\Omega)}).
\tag{3.14}
$$

On the other hand by (3.9)–(3.10) again, we have

$$
\begin{aligned}
&\left|\int_{\Omega}\left(1-\frac{1}{\rho}\chi_{E_\rho}(x)\right)h^0(x)dx\right|\le\sum_{j=1}^{m}\left|\int_{K_j}\left(1-\frac{1}{\rho}\chi_{E_\rho^j}(x)\right)h^0(x)dx\right| \\
&\le\sum_{j=1}^{m}\left(1+\frac{1}{\rho}\right)\int_{K_j}|h^0(x)-h_j^0(x)|dx+\sum_{j=1}^{m}\left|\int_{K_j}\left(1-\frac{1}{\rho}\chi_{E_\rho^j}\right)h_j^0(x)dx\right| \\
&<\varepsilon\left(1+\frac{1}{\rho}\right).
\end{aligned}
\tag{3.15}
$$

Therefore, our conclusion follows. □

The above result was proved for $\mathcal{Y}=W_0^{1,p}(\Omega)$ in [5] using different methods. The proof given here is inspired by a personal communication of the second author with E. Casas.

Now we consider the equation

$$
\begin{cases}
\zeta_t-(a_{ij}(x,t)\zeta_{x_i})_{x_j}+c_\rho(x,t)\zeta(x,t)=\left(1-\frac{1}{\rho}\chi_{E_\rho}(x,t)\right)h(x,t), \text{ in } \Omega_T, \\
\zeta\big|_{\partial_p\Omega_T}=0,
\end{cases}
\tag{3.16}
$$

where a_{ij} satisfies (A1) and $\partial_p\Omega_T=(\partial\Omega\times[0,T])\bigcup(\bar{\Omega}\times\{0\})$ is the parabolic boundary of Ω_T. It is clear that the solution $\zeta=\zeta_{E_\rho}(x,t)$ is uniquely determined by the choice of the coefficient c_ρ and the set E_ρ.

LEMMA 3.2. *Suppose that $\frac{n+2}{2}<p<\infty$. Then, there exists a $\theta\in(0,1)$ such that for each $h^0\in L^1(\Omega_T)$, $h\in L^p(\Omega_T)$, $\mathcal{K}=\{c(x,t);\|c\|_{L^\infty(\Omega_T)}\le K\}$ $(K>0)$, and any $\rho\in(0,1)$,*

$$
\inf_{E_\rho\in\mathcal{E}_\rho}\sup_{c_\rho\in\mathcal{K}}\left\{\left|\int_{\Omega_T}\left(1-\frac{1}{\rho}\chi_{E_\rho}(x,t)\right)h^0(x,t)dxdt\right|+\|\zeta_{E_\rho}\|_{C^{\theta,\theta/2}(\bar{\Omega}_T)}\right\}=0.
\tag{3.17}
$$

Proof. By the assumption, $h \in L^p(\Omega_T)$, $p > \frac{n+2}{2}$. Therefore, the right-hand side of the equation (3.16) is in $L^p(\Omega_T)$ (although the $L^p(\Omega_T)$ norm may blow up as $\rho \to 0$). It follows from the parabolic estimates [12, Chap. III, §10] that there exists $\beta \in (0,1)$, such that

$$\sup_{E_\rho \in \mathcal{E}_\rho} \sup_{c_\rho \in \mathcal{K}} \|\zeta_{E_\rho}\|_{C^{\beta,\beta/2}(\Omega_T)} \le C_\rho. \tag{3.18}$$

We claim that, for any $0 < \theta < \beta$, (3.17) holds. In fact, we first note that the identity mapping $C^{\beta,\beta/2}(\bar{\Omega}_T) \hookrightarrow L^{p'}(\Omega_T)$ is compact. Therefore, if we take $\mathcal{Y} = C^{\beta,\beta/2}(\bar{\Omega}_T)$ in Lemma 3.1, then

$$\begin{aligned} &\inf_{E_\rho \in \mathcal{E}_\rho} \sup_{c_\rho \in \mathcal{K}} \left\{ \left| \int_{\Omega_T} \left(1 - \frac{1}{\rho}\chi_{E_\rho}\right) h^0 dxdt \right| + \left| \int_{\Omega_T} \left(1 - \frac{1}{\rho}\chi_{E_\rho}\right) h \cdot \zeta_{E_\rho} dxdt \right| \right\} \\ &\le \inf_{E_\rho \in \mathcal{E}_\rho} \left\{ \left| \int_{\Omega_T} \left(1 - \frac{1}{\rho}\chi_{E_\rho}\right) h^0 dxdt \right| + C_\rho \left\| \left(1 - \frac{1}{\rho}\chi_{E_\rho}\right) h \right\|_{\mathcal{Y}^*} \right\} = 0. \end{aligned} \tag{3.19}$$

Using a change of variable $\zeta(x,t) = \phi(x,t)e^{Kt}$ if necessary, we may assume without loss of generality that $c_\rho(x,t) \ge 0$. Multiplying equation (3.16) with ζ_{E_ρ} and integrating over Ω_T, we immediately obtain

$$\begin{aligned} &\int_\Omega (\zeta_{E_\rho})^2(x,T)dx + \int_{\Omega_T} |\nabla \zeta_{E_\rho}(x,\tau)|^2 dxd\tau \\ &\qquad \le C \left| \int_0^T \int_\Omega \left(1 - \frac{1}{\rho}\chi_{E_\rho}\right) h \cdot \zeta_{E_\rho} dxdt \right|. \end{aligned} \tag{3.20}$$

Notice that $\zeta_{E_\rho} = 0$ on $\partial\Omega \times \{t\}$, for each $t \in (0,T)$. Therefore, $\int_\Omega (\zeta_{E_\rho})^2(x,t)dx \le C \int_\Omega |\nabla \zeta_{E_\rho}|^2(x,t)dx$, by Poincaré's inequality. Integrating over $t \in [0,T]$, and taking (3.20) into account, yields

$$\int_{\Omega_T} (\zeta_{E_\rho})^2(x,t)dxdt \le C \left| \int_{\Omega_T} \left(1 - \frac{1}{\rho}\chi_{E_\rho}(x,t)\right) h(x,t)\zeta_{E_\rho}(x,t)dxdt \right|. \tag{3.21}$$

By the interpolation theorem (see Lemma 3.4 below), for any $\varepsilon > 0$, there exists $C_\varepsilon > 0$, such that

$$\|\zeta\|_{C^{\theta,\theta/2}(\bar{\Omega}_T)} \le \varepsilon \|\zeta\|_{C^{\beta,\beta/2}(\bar{\Omega}_T)} + C_\varepsilon \|\zeta\|_{L^2(\Omega_T)}, \quad \forall \zeta \in C^{\beta,\beta/2}(\bar{\Omega}_T). \tag{3.22}$$

Using (3.19), (3.21) and (3.22), we obtain

$$\inf_{E \in \mathcal{E}_\rho} \sup_{c_\rho \in \mathcal{K}} \left\{ \left| \int_{\Omega_T} \left(1 - \frac{1}{\rho}\chi_{E_\rho}(x,t)\right) h^0(x,t)dxdt \right| + \|\zeta_{E_\rho}\|_{C^{\theta,\theta/2}(\bar{\Omega}_T)} \right\} \le \varepsilon C_\rho.$$

Since ε can be arbitrarily small, the lemma follows. □

Now, for any feasible pair (y,u), we define

$$\begin{cases} c(x,t) = -f_y(x,t,y(x,t),u(x,t)), \\ c^0(x,t) = f_y^0(x,t,y(x,t),u(x,t)), \end{cases} \tag{3.23}$$

and for given $v \in \mathcal{U}$,

$$
\left\{\begin{array}{l}
h(x,t)=f(x,t,y(x,t),v(x,t))-f(x,t,y(x,t),u(x,t)),\\
h^0(x,t)=f^0(x,t,y(x,t),v(x,t))-f^0(x,t,y(x,t),u(x,t)).
\end{array}\right. \tag{3.24}
$$

Consider the following problem:

$$
\left\{\begin{array}{ll}
z_t-\sum\limits_{i,j=1}^{n}(a_{ij}(x,t)z_{x_i})_{x_j}+c(x,t)z=h(x,t), & \text{in } \Omega_T,\\
z\,\big|_{\partial_p\Omega_T}=0.
\end{array}\right. \tag{3.25}
$$

Clearly, since $h \in L^\infty(\Omega_T)$, this problem admits a unique solution $z \in C^{\beta,\beta/2}(\bar{\Omega}_T)\bigcap L^2(0,T;H_0^1(\Omega))$, as in Proposition 2.1.

Our main result of this section is the following.

THEOREM 3.3. *Let (y,u) be a given feasible pair and $v \in \mathcal{U}$ be fixed. Then, for any $\rho \in (0,1)$, there exists a measurable set $E_\rho \subset \Omega_T$, with property $|E_\rho| = \rho|\Omega_T|$, such that if we define u_ρ by*

$$
u_\rho(x,t)=\left\{\begin{array}{ll}
u(x,t), & \text{if } (x,t)\in\Omega_T\setminus E_\rho,\\
v(x,t), & \text{if } (x,t)\in E_\rho,
\end{array}\right. \tag{3.26}
$$

and let y_ρ be the state corresponding to u_ρ, then the following hold:

$$
\left\{\begin{array}{l}
y_\rho=y+\rho z+r_\rho,\\
\lim\limits_{\rho\to0}\dfrac{1}{\rho}\|r_\rho\|_{C^{\theta,\theta/2}(\bar{\Omega}_T)}=0,
\end{array}\right. \tag{3.27}
$$

for some $\theta \in (0,1)$, and

$$
\left\{\begin{array}{l}
J(u_\rho)=J(u)+\rho z^0+r_\rho^0,\\
\lim\limits_{\rho\to0}\dfrac{1}{\rho}|r_\rho^0|=0,
\end{array}\right. \tag{3.28}
$$

where z is the solution of (3.25) *and z^0 is given by*

$$
z^0=\int_\Omega[c^0(x,t)z(x,t)+h^0(x,t)]dx. \tag{3.29}
$$

Proof. First, we recall the so-called Ekeland distance. For any $u, v \in \mathcal{U}$, we let

$$
\bar{d}(u,v)=|\{(x,t)\in\Omega \mid u(x,t)\neq v(x,t)\}|. \tag{3.30}
$$

It is standard that $(\mathcal{U},\bar{d}(\cdot,\cdot))$ is a complete metric space (see [8]). Clearly, $\bar{d}(u,u_\rho)\le |E_\rho|$.

Now, we set

$$
z_\rho(x,t)=\frac{y_\rho(x,t)-y(x,t)}{\rho}, \qquad x\in\Omega. \tag{3.31}
$$

Then, z_ρ satisfies the following:

$$\tag{3.32}\begin{cases}(z_\rho)_t - \sum_{i,j=1}^n (a_{ij}(x,t)(z_\rho)_{x_i})_{x_j} + c_\rho(x,t)z_\rho = \frac{1}{\rho}\chi_{E_\rho}(x,t)h(x,t),\\ z_\rho \,\big|_{\partial_p\Omega_T} = 0,\end{cases}$$

where

$$\tag{3.33} c_\rho(x,t) = -\int_0^1 f_y(x,t,y(x,t)+\tau(y_\rho(x,t)-y(x,t)),u_\rho(x,t))d\tau.$$

We see that (note (2.9) and (2.3)) $c_\rho(x,t)$ and $h(x,t)$ are uniformly bounded (with the bounds independent of E_ρ, the controls u and v). The function $h(x,t)$ is actually independent of the set E_ρ; we shall use this fact when we apply Lemma 3.2. Since $h \in L^\infty(\Omega_T) \subset L^p(\Omega_T)$ for any $p > 1$, the parabolic Hölder's estimate implies that $y_\rho - y = \rho z_\rho$ satisfies, for fixed $p > \frac{n+2}{2}$,

$$\tag{3.34} \|y_\rho - y\|_{C^{\beta,\beta/2}(\bar\Omega_T)} \le C\|\chi_{E_\rho}\|_{L^p(\Omega_T)} \equiv \omega(C\rho) \to 0, \qquad \text{as } \rho \to 0,$$

where the constant C is independent of E_ρ, and ω is a modulus of continuity. It follows that

$$\tag{3.35} c_\rho(x,t) \to c(x,t) \equiv -f_y(x,t,y(x,t),u(x,t)), \qquad \text{in } L^p(\Omega_T),\ 1 \le p < \infty.$$

By recalling z, the solution of (3.25), we have the following:

$$\tag{3.36}\begin{cases}(z_\rho - z)_t - \sum_{i,j=1}^n \left(a_{ij}(x,t)(z_\rho - z)_{x_i}\right)_{x_j} + c_\rho(x,t)(z_\rho - z)\\ \qquad = -(c_\rho(x,t) - c(x,t))z - \left(1 - \frac{1}{\rho}\chi_{E_\rho}(x,t)\right)h(x,t),\\ (z_\rho - z)\big|_{\partial_p\Omega_T} = 0,\end{cases}$$

We note that the above equation is linear in $(z_\rho - z)$. Thus, we may write $z_\rho - z = \bar\zeta_\rho + \zeta_\rho$ with $\bar\zeta_\rho$ and ζ_ρ satisfying the following:

$$\tag{3.37}\begin{cases}(\bar\zeta_\rho)_t - \sum_{i,j=1}^n \left(a_{ij}(x,t)(\bar\zeta_\rho)_{x_i}\right)_{x_j} + c_\rho(x,t)\bar\zeta_\rho = -(c_\rho(x,t) - c(x,t))z\\ \bar\zeta_\rho \,\big|_{\partial_p\Omega_T} = 0,\end{cases}$$

and

$$\tag{3.38}\begin{cases}(\zeta_\rho)_t - \sum_{i,j=1}^n (a_{ij}(x,t)(\zeta_\rho)_{x_i})_{x_j} + c_\rho(x,t)\zeta_\rho = -\left(1 - \frac{1}{\rho}\chi_{E_\rho}(x,t)\right)h(x,t),\\ \zeta_\rho \,\big|_{\partial_p\Omega_T} = 0.\end{cases}$$

By Hölder estimates, again (notice that $z \in C^{\beta,\beta/2}(\bar\Omega_T) \subset L^\infty(\Omega_T)$, and p is fixed with $p > \frac{n+2}{2}$), we have

$$\tag{3.39} \|\bar\zeta_\rho\|_{C^{\beta,\beta/2}(\bar\Omega_T)} \le C\|(c_\rho - c)z\|_{L^p(\Omega_T)} = o(1) \quad \text{as } \rho \to 0;$$

all the constants involved in the above are independent of the choices of E_ρ.

Now we fix $\theta \in (0, \beta)$ as in Lemma 3.2. Then we can choose $E_\rho \subset \Omega_T$ with the property $|E_\rho| = \rho|\Omega_T|$, such that the solution ζ_ρ of (3.38) satisfies

$$\left|\int_{\Omega_T}\left(1-\frac{1}{\rho}\chi_{E_\rho}(x,t)\right)h^0(x,t)dxdt\right| + \|\zeta_\rho\|_{C^{\theta,\theta/2}(\bar\Omega_T)} \le \rho. \tag{3.40}$$

This proves (3.27). The proof of (3.28) is similar but simpler. □

The above result will play a very important role in the proof of our main result (Theorem 2.2). Conclusion (3.27) gives a "Taylor expansion" (of first order) in the space $C^{\theta,\theta/2}(\bar\Omega_T) \subset C(\bar\Omega_T)$. This will be sufficient for us to deal with the pointwise state constraint. It is not hard for us to see that the stronger the topology under which (3.27) holds, the harder for us to prove it. Thus, for example, if in (3.27), $C^{\theta,\theta/2}(\bar\Omega_T)$ is replaced by $L^p(\Omega_T)$, then it will be much easier to prove it. In another word, an $L^p(\Omega_T)$ constraint of the state is much easier to treat than a $C(\bar\Omega_T)$ constraint.

We now give a proof for the interpolation theorem used in the proof of Lemma 3.2. The identity mappings $C^{\beta,\beta/2}(\bar\Omega_T) \hookrightarrow C^{\theta,\theta/2}(\bar\Omega_T) \hookrightarrow L^2(\Omega_T)$ are continuous and compact. Therefore, the interpolation follows from a compactness argument. However, the compactness argument does not give us the exact form of the constants C_ε. Lemma 3.4 below is a stronger statement.

The interpolation involves *different types* of spaces. Nonetheless, the proof is similar to that in, for example, [11].

LEMMA 3.4. *Suppose that $\partial\Omega$ is Lipschitz continuous, $0 \le \theta < \beta < 1$ and $0 < p \le \infty$. Then there exists a constant C, depending only on Ω and T, such that*

$$\begin{gathered}\|\zeta\|_{C^{\theta,\theta/2}(\bar\Omega_T)} \le 4\varepsilon[\zeta]_{C^{\beta,\beta/2}(\bar\Omega_T)} + \frac{3C^{1/p}}{\varepsilon^\mu}\|\zeta\|_{L^p(\Omega_T)},\\ \forall\zeta \in C^{\beta,\beta/2}(\bar\Omega_T), \quad \forall 0<\varepsilon\le 1,\end{gathered} \tag{3.41}$$

where $\mu = \frac{\theta}{(\beta-\theta)} + \frac{n+2}{(\beta-\theta)p}$, *and*

$$\begin{cases} [\zeta]_{C^{\theta,\theta/2}(\bar\Omega_T)} = \sup\left\{\dfrac{|\zeta(x,t)-\zeta(\bar x,\bar t)|}{(\sqrt{|x-\bar x|^2+|t-\bar t|}\,)^\theta}; \quad (x,t) \ne (\bar x,\bar t) \in \bar\Omega_T\right\},\\ \|\zeta\|_{C^{\theta,\theta/2}(\bar\Omega_T)} = \|\zeta\|_{C(\bar\Omega_T)} + [\zeta]_{C^{\theta,\theta/2}(\bar\Omega_T)} \qquad \textit{when } 0<\theta<1,\\ \|\zeta\|_{C^{\theta,\theta/2}(\bar\Omega_T)} = \|\zeta\|_{C(\bar\Omega_T)} \qquad \textit{when } \theta = 0.\end{cases} \tag{3.42}$$

(*The* $\|\cdot\|_{L^p(\Omega_T)}$ *should be understood in the usual sense*; *it should be noted that it is not a norm when* $0<p<1$.)

Proof. We let $\delta = \varepsilon^{1/(\beta-\theta)}$; then $0 < \delta \le 1$. Splitting the *sup* in (3.42) into two sets $\{\sqrt{|x-\bar x|^2+|t-\bar t|} \le \delta\}$ and $\{\sqrt{|x-\bar x|^2+|t-\bar t|} > \delta\}$ immediately gives us

$$\|\zeta\|_{C^{\theta,\theta/2}(\bar\Omega_T)} \le \delta^{\beta-\theta}[\zeta]_{C^{\beta,\beta/2}(\bar\Omega_T)} + \left(1+\frac{2}{\delta^\theta}\right)\|\zeta\|_{C(\bar\Omega_T)}, \tag{3.43}$$

from which the case $p=\infty$ follows.

Now consider the case $0<p<\infty$. Since ζ is continuous on $\bar\Omega_T$, $\|\zeta\|_{C(\bar\Omega_T)} = |\zeta(\tilde x,\tilde t)|$, for some $(\tilde x,\tilde t) \in \bar\Omega_T$. Now let $B_\delta = \{(x,t) \in \bar\Omega_T;\ \sqrt{|x-\tilde x|^2+|t-\tilde t|} \le \delta\}$; then by the mean value theorem,

$$\left(\frac{1}{|B_\delta\bigcap\bar\Omega_T|}\int_{B_\delta\bigcap\bar\Omega_T}|\zeta(x,t)|^p dxdt\right)^{1/p} = |\zeta(x^*,t^*)|$$

for some $(x^*,t^*) \in B_\delta \bigcap \bar{\Omega}_T$. Since $\partial\Omega$ is Lipschitz continuous, $|B_\delta \bigcap \bar{\Omega}_T| \geq \delta^{n+2}/C$ for some generic constant $C > 0$. It follows that

$$\begin{aligned} \|\zeta\|_{C(\bar{\Omega}_T)} &= |\zeta(\widetilde{x},\widetilde{t})| \leq |\zeta(\widetilde{x},\widetilde{t}) - \zeta(x^*,t^*)| + |\zeta(x^*,t^*)| \\ &\leq \delta^\beta[\zeta]_{C^{\beta,\beta/2}(\bar{\Omega}_T)} + \left(\frac{C}{\delta^{n+2}}\right)^{1/p} \|\zeta\|_{L^p(\Omega_T)}; \end{aligned} \tag{3.44}$$

the case $\theta = 0$ follows immediately. Now substituting (3.45) into (3.44), we obtain

$$\|\zeta\|_{C^{\theta,\theta/2}(\bar{\Omega}_T)} \leq 4\delta^{\beta-\theta}[\zeta]_{C^{\beta,\beta/2}(\bar{\Omega}_T)} + \frac{3}{\delta^\theta}\left(\frac{C}{\delta^{n+2}}\right)^{1/p} \|\zeta\|_{L^p(\Omega_T)},$$

and the general case $0 < \theta < \beta < 1$, $0 < p < \infty$ follows. □

4. Proof of Theorem 2.2. In this section, we are going to prove the main theorem of this paper.

Proof of Theorem 2.2. Let $(\bar{y},\bar{u})$ be an optimal pair. For any $u \in \mathcal{U}$, let $y(\cdot,\cdot\,;\,u)$ be the corresponding state, emphasizing the dependence of it on the control. For any $\varepsilon > 0$, we define

$$J_\varepsilon(u) = \{[(J(u) - J(\bar{u}) + \varepsilon)^+]^2 + d_Q(G(y(\cdot,\cdot\,;\,u)))^2\}^{1/2}. \tag{4.1}$$

Clearly, this functional is continuous on the (complete) metric space $(\mathcal{U},\bar{d})$ (recall that $\bar{d}$ is the Ekeland distance; see (3.30)). Also, we have

$$J_\varepsilon(u) > 0, \qquad \forall u \in \mathcal{U}, \tag{4.2}$$

$$J_\varepsilon(\bar{u}) = \varepsilon \leq \inf_{\mathcal{U}} J_\varepsilon(u) + \varepsilon. \tag{4.3}$$

Hence, by Ekeland's variational principle [6], we can find a $u^\varepsilon \in \mathcal{U}$, such that

$$\bar{d}(\bar{u},u^\varepsilon) \leq \sqrt{\varepsilon}, \tag{4.4}$$

$$J_\varepsilon(u^\varepsilon) \leq J_\varepsilon(\bar{u}), \tag{4.5}$$

$$J_\varepsilon(\widehat{u}) - J_\varepsilon(u^\varepsilon) \geq -\sqrt{\varepsilon}\,d(\widehat{u},u^\varepsilon), \qquad \forall \widehat{u} \in \mathcal{U}. \tag{4.6}$$

We let $v \in \mathcal{U}$ and $\varepsilon > 0$ be fixed and let $y^\varepsilon = y(\cdot,\cdot\,;u^\varepsilon)$. By Theorem 3.3, we know that for any $\rho \in (0,1)$, there exists a measurable set $E^\varepsilon_\rho \subset \Omega_T$ with the property $|E^\varepsilon_\rho| = \rho|\Omega_T|$, such that if we define

$$u^\varepsilon_\rho(x,t) = \begin{cases} u^\varepsilon(x,t), & \text{if } (x,t) \in \Omega_T \setminus E^\varepsilon_\rho, \\ v(x,t), & \text{if } (x,t) \in E^\varepsilon_\rho, \end{cases} \tag{4.7}$$

and let $y^\varepsilon_\rho = y(\cdot,\cdot\,;\,u^\varepsilon_\rho)$ be the corresponding state, then

$$\begin{cases} y^\varepsilon_\rho = y^\varepsilon + \rho z^\varepsilon + r^\varepsilon_\rho, \\ J(u^\varepsilon_\rho) = J(u^\varepsilon) + \rho z^{0,\varepsilon} + r^{0,\varepsilon}_\rho, \end{cases} \tag{4.8}$$

where z^ε and $z^{0,\varepsilon}$ satisfy the following:

$$\begin{cases} z^\varepsilon_t - \displaystyle\sum_{i,j}^{n} (a_{ij}(x,t) z^\varepsilon_{x_i})_{x_j} - f_y(x,t,y^\varepsilon(x,t),u^\varepsilon(x,t)) z^\varepsilon = h^\varepsilon(x,t), & \text{in } \Omega_T, \\ z^\varepsilon \,\big|_{\partial_p \Omega_T} = 0. \end{cases} \tag{4.9}$$

$$z^{0,\varepsilon} = \int_{\Omega_T} [f^0_y(x,t,y^\varepsilon(x,t),u^\varepsilon(x,t)) z^\varepsilon(x,t) + h^{0,\varepsilon}(x,t)]dx, \tag{4.10}$$

with

$$\begin{cases} h^\varepsilon(x,t) = f(x,t,y^\varepsilon(x,t),v(x,t)) - f(x,t,y^\varepsilon(x,t),u^\varepsilon(x,t)), \\ h^{0,\varepsilon}(x,t) = f^0(x,t,y^\varepsilon(x,t),v(x,t)) - f^0(x,t,y^\varepsilon(x,t),u^\varepsilon(x,t)). \end{cases} \tag{4.11}$$

and for some $\theta \in (0,1)$,

$$\lim_{\rho\to0} \frac{1}{\rho} \|r^\varepsilon_\rho\|_{C^{\theta,\theta/2}(\bar\Omega_T)} = \lim_{\rho\to0} \frac{1}{\rho} |r^{0,\varepsilon}_\rho| = 0. \tag{4.12}$$

Now, we take $\widehat{u} = u^\varepsilon_\rho$ in (4.6). Then, it follows that

$$\begin{aligned} -\sqrt{\varepsilon}|\Omega_T| &\le \frac{J_\varepsilon(u^\varepsilon_\rho) - J_\varepsilon(u^\varepsilon)}{\rho} \\ &= \frac{1}{J_\varepsilon(u^\varepsilon_\rho) + J_\varepsilon(u^\varepsilon)} \left\{ \frac{[(J(u^\varepsilon_\rho) - J(\bar u) + \varepsilon)^+]^2 - [(J(u^\varepsilon) - J(\bar u) + \varepsilon)^+]^2}{\rho} \right. \\ &\qquad \left. + \frac{d_Q(G(y^\varepsilon_\rho))^2 - d_Q(G(y^\varepsilon))^2}{\rho} \right\} \\ &\to \frac{(J(u^\varepsilon) - J(\bar u) + \varepsilon)^+}{J_\varepsilon(u^\varepsilon)} z^{0,\varepsilon} + \left\langle \frac{d_Q(G(y^\varepsilon))\xi_\varepsilon}{J_\varepsilon(u^\varepsilon)}, G'(y^\varepsilon) z^\varepsilon \right\rangle, \qquad (\rho \to 0), \end{aligned} \tag{4.13}$$

where

$$\xi_\varepsilon = \begin{cases} \nabla d_Q(G(y^\varepsilon)), & \text{if } G(y^\varepsilon) \notin Q, \\ 0, & \text{if } G(y^\varepsilon) \in Q. \end{cases} \tag{4.14}$$

We note that since $G : C_0(\bar\Omega_T) \to Z$, to obtain the convergence in (4.13), the expansion (4.8) *in the space* $C_0(\bar\Omega_T)$ is necessary.

Next, we define $(\varphi^{0,\varepsilon}, \varphi^\varepsilon) \in [0,1] \times \mathcal{M}(\bar\Omega_T)$ as follows:

$$\begin{cases} \varphi^{0,\varepsilon} = \dfrac{(J(u^\varepsilon) - J(\bar u) + \varepsilon)^+}{J_\varepsilon(u^\varepsilon)}, \\ \varphi^\varepsilon = \dfrac{d_Q(G(y^\varepsilon))\xi_\varepsilon}{J_\varepsilon(u^\varepsilon)}. \end{cases} \tag{4.15}$$

Then we see that (4.13) becomes

$$-\sqrt{\varepsilon}|\Omega_T| \le \varphi^{0,\varepsilon} z^{0,\varepsilon} + \langle \varphi^\varepsilon, G'(y^\varepsilon) z^\varepsilon \rangle. \tag{4.16}$$

By (2.8) and (4.1), we have

$$|\varphi^{0,\varepsilon}|^2 + \|\varphi^\varepsilon\|^2_{Z^*} = 1. \tag{4.17}$$

On the other hand, by the definition of subdifferential, we have

$$\langle \varphi^\varepsilon, \eta - G(y^\varepsilon) \rangle \le 0, \qquad \forall \eta \in Q. \tag{4.18}$$

Next, by (4.4), we have

$$\|y^\varepsilon - \bar{y}\|_{C^{\beta,\beta/2}(\bar{\Omega}_T)} \to 0, \qquad (\varepsilon \to 0). \tag{4.19}$$

Thus, (4.18) implies

$$\langle \varphi^\varepsilon, \eta - G(\bar{y}) \rangle \le \langle \varphi^\varepsilon, G(y^\varepsilon) - G(\bar{y}) \rangle \le \|G(y^\varepsilon) - G(\bar{y})\|_Z \to 0, \qquad \forall \eta \in Q. \tag{4.20}$$

Since Q is finite codimensional in Z, from [15, Lem. 3.2], we know that by extracting some subsequence, still denoted by itself, one has

$$(\varphi^{0,\varepsilon}, \varphi^\varepsilon) \overset{*}{\rightharpoonup} (\varphi^0, \varphi) \ne 0. \tag{4.21}$$

On the other hand, from (4.19) and the equations (4.9), (4.10), we have

$$\begin{cases} z^\varepsilon \longrightarrow z, & \text{in } C^{\beta,\beta/2}(\bar{\Omega}_T), \\ z^{0,\varepsilon} \longrightarrow z^0, & \end{cases} \qquad (\varepsilon \to 0), \tag{4.22}$$

where z is the solution of the following variational system:

$$\begin{cases} z_t - \displaystyle\sum_{i,j=1}^n (a_{ij}(x,t) z_{x_i})_{x_j} = f_y(x,t,\bar{y}(x,t),\bar{u}(x,t)) z \\ \qquad\qquad + f(x,t,\bar{y}(x,t),v(x,t)) - f(x,t,\bar{y}(x,t),\bar{u}(x,t)), \qquad \text{in } \Omega_T, \\ z \,\big|_{\partial_p \Omega_T} = 0 \end{cases} \tag{4.23}$$

and

$$\begin{aligned} z^0 = & \int_{\Omega_T} f^0_y(x,t,\bar{y}(x,t),\bar{u}(x,t)) z(x,t)\,dxdt \\ & + \int_{\Omega_T} [f^0(x,t,\bar{y}(x,t),v(x,t)) - f^0(x,t,\bar{y}(x,t),\bar{u}(x,t))]\,dxdt. \end{aligned} \tag{4.24}$$

We note that the solution z of (4.23) and the quantity z^0 defined by (4.24) depend on the choice of $v \in \mathcal{U}$. Thus, we denote them by $z(\cdot,\cdot\,;\, v)$ and $z^0(v)$, respectively. Then, taking limits in (4.16), we obtain

$$\varphi^0 z^0(v) + \langle \varphi, G'(\bar{y}) z(\cdot,\cdot\,; v) \rangle \ge 0, \qquad \forall v \in \mathcal{U}. \tag{4.25}$$

Now, we let

$$\psi^0 = -\varphi^0 \in [-1, 0]. \tag{4.26}$$

Then, (2.11) follows from (4.21). Also, we obtain (2.13) by taking limits in (4.20) (along the above-mentioned subsequence). Furthermore, (4.25) can be written as

$$\psi^0 z^0(v) - \langle G'(\bar y)^*\varphi, z(\cdot,\cdot\,;v)\rangle \le 0, \qquad \forall v \in \mathcal{U}. \tag{4.27}$$

We note that $G'(\bar y)^*\varphi \in \mathcal{M}(\bar\Omega_T)$, and by our compatible condition (2.10), we see that

$$\begin{aligned}\langle G'(\bar y)^*\varphi, z\rangle &= \langle (G'(\bar y)^*\varphi)\big|_{\Omega_T}, z\rangle_{\mathcal{M}(\Omega_T),C(\bar\Omega_T)}\\ &\quad + \langle (G'(\bar y)^*\varphi)\big|_{\Omega\times\{T\}}, z\big|_{t=T}\rangle_{\mathcal{M}(\Omega),C(\bar\Omega)}.\end{aligned} \tag{4.28}$$

By [1], we know that (2.12) admits a solution in $L^q(0,T,W_0^{1,q}(\Omega))$ for any $1 < q < \frac{n+2}{n+1}$. However, unlike the elliptic equations, the function $z(x,t;v)$ is not smooth enough (in the t direction) to be a test function for the equation (2.12). We shall get around this problem by approximating the equations for z and the equations (2.12).

We consider the following approximation for $z(x,t;v)$:

$$\begin{cases} z_t^\delta - \displaystyle\sum_{i,j=1}^n (a_{ij}^\delta(x,t) z_{x_i}^\delta)_{x_j} = f_y(x,t,\bar y(x,t),\bar u(x,t)) z^\delta \\ \qquad\qquad + f(x,t,\bar y(x,t),v(x,t)) - f(x,t,\bar y(x,t),\bar u(x,t)), \qquad \text{in } \Omega_T, \\ z^\delta\,\big|_{\partial_p\Omega_T} = 0, \end{cases} \tag{4.29}$$

where $a_{ij}^\delta \in C^{2,1}(\bar\Omega_T)$, a_{ij}^δ satisfies (A1), and

$$a_{ij}^\delta \to a_{ij} \qquad \text{in } L^p(\Omega_T), \text{ as } \delta \to 0, \tag{4.30}$$

for any $1 < p < \infty$. By L^p estimates for the parabolic equations, $z^\delta = z^\delta(\cdot,\cdot\,;v) \in W_p^{2,1}(\Omega_T)$ for any $1 < p < \infty$. As before, we have the estimates

$$\|z^\delta\|_{C^{\beta,\beta/2}(\bar\Omega_T)} + \|\nabla_x z^\delta\|_{L^2(\Omega_T)} \le C, \tag{4.31}$$

where the constants C and β are independent of δ and $v \in \mathcal{U}$. Thus, by compactness and the uniqueness of the equation (4.23), one can easily derive that

$$\|z^\delta - z\|_{C(\bar\Omega_T)} \to 0, \qquad \text{as } \delta \to 0. \tag{4.32}$$

Clearly, (4.27) and (4.32) imply that

$$\lim_{\delta\to 0} \{\psi^0 z^0(v) - \langle G'(\bar y)^*\varphi, z^\delta(\cdot,\cdot\,;v)\rangle\} \le 0, \qquad \forall v \in \mathcal{U}. \tag{4.33}$$

By [1], if we replace a_{ij} with a_{ij}^δ, then (2.12) has a solution $\psi^\delta \in L^q(0,T,W_0^{1,q}(\Omega))$, where $1 < q < \frac{n+2}{n+1}$. (ψ^δ is actually unique.) Furthermore,

$$\|\psi^\delta\|_{L^q(0,T,W_0^{1,q}(\Omega))} \le C, \tag{4.34}$$

where the constant C is independent of δ. Clearly, v is not involved in the definition of ψ^δ. By passing to a subsequence if necessary, we have, as $\delta \to 0$,

$$\psi^\delta \overset{w}{\rightharpoonup} \psi \quad \text{in } L^q(\Omega_T), \qquad \psi_{x_j}^\delta \overset{w}{\rightharpoonup} \psi_{x_j} \quad \text{in } L^q(\Omega_T), \tag{4.35}$$

for some function ψ. It follows that ψ is a solution of (2.12).

Since $z^\delta \in W_p^{2,1}(\Omega_T)$ for $p > q/(q-1)$, we can use z^δ as a test function in the equation for ψ^δ.

Then, by some direct computation, we can reduce (4.33) to the following:

$$\begin{aligned}\int_{\Omega_T} &\{\psi^0[f^0(x,t,\bar{y}(x,t),\bar{u}(x,t)) - f^0(x,t,\bar{y}(x,t),v(x,t))] \\ &+ \psi^\delta(x,t)[f(x,t,\bar{y}(x,t),\bar{u}(x,t)) - f(x,t,\bar{y}(x,t),v(x,t))]\}dx \\ \geq\ & o(1), \qquad \text{as } \delta \to 0.\end{aligned} \tag{4.36}$$

Now letting $\delta \to 0$ and recalling (4.35), we obtain

$$\begin{aligned}&\int_{\Omega_T} \{\psi^0[f^0(x,t,\bar{y}(x,t),\bar{u}(x,t)) - f^0(x,t,\bar{y}(x,t),v(x,t))] \\ &\qquad + \psi(x,t)[f(x,t,\bar{y}(x,t),\bar{u}(x,t)) - f(x,t,\bar{y}(x,t),v(x,t))]\}dx \\ &\equiv \int_\Omega [H(x,t,\bar{y}(x,t),\bar{u}(x,t),\psi^0,\psi(x,t)) - H(x,t,\bar{y}(x,t),v(x,t),\psi^0,\psi(x,t))]dx, \\ &\geq 0, \qquad\qquad \forall v \in \mathcal{U}.\end{aligned} \tag{4.37}$$

Then, by the separability of U and the continuity of the Hamiltonian H in the variable v, noticing also that $v \in \mathcal{U}$ is arbitrary, we obtain the maximum condition (2.14) (see [5]). □

5. Applications. In this section, we would like to discuss some special cases which are covered by our main result.

We first consider the following case. Let $Z = C_0(\bar{\Omega}_T)$ with some norm $|\cdot|_0$ which is equivalent to $\|\cdot\|_{C(\bar{\Omega}_T)}$ and whose dual, still denoted by $\mathcal{M}_0(\bar{\Omega}_T)$, is strictly convex. We let $Q \subset Z$ be defined as in (1.3) and $g : \bar{\Omega}_T \times \mathbb{R} \to \mathbb{R}$ be continuous, with $g_y(x,t,y)$ also being continuous. Moreover,

$$\begin{cases} g(x,t,0) < 0, & \forall (x,t) \in \partial\Omega \times [0,T], \\ g(x,0,y_0(x)) < 0, & \forall x \in \bar{\Omega}, \end{cases} \tag{5.1}$$

where $y_0 \in C_0(\bar{\Omega})$. We let $G(\eta)(x,t) = g(x,t,\eta(x,t))$, for any $\eta \in C_0(\bar{\Omega}_T)$. Then, the following result holds.

PROPOSITION 5.1. *For the above Q, G, and y_0, condition* (2.10) *holds.*

Proof. By (5.1), we see that for any $\varepsilon > 0$, there exists a $\delta > 0$, such that

$$\begin{aligned} g(x,t,y) \leq -\varepsilon, \quad & \text{if } t \in [\delta,T], |y| < \delta, d(x,\partial\Omega) < \delta, \\ & \text{or } t \in [0,\delta], x \in \Omega, |y - y_0(x)| < \delta. \end{aligned} \tag{5.2}$$

Thus, for any $\eta \in C_0(\bar{\Omega}_T)$ with $G(\eta) \in Q$ and $\eta\big|_{t=0} = y_0(x)$, we have the following: For any $\varepsilon > 0$, there exists a $\delta > 0$, such that

$$g(x,t,\eta(x,t)) \leq -\varepsilon, \qquad \text{if } d((x,t),\partial_p\Omega_T) < \delta. \tag{5.3}$$

Now, for any $\varphi \in C(\bar{\Omega}_T)$, if for some $\delta > 0$, it holds that

$$\operatorname{supp}\varphi \subset \{(x,t) \in \bar{\Omega}_T \bigm| d((x,t),\partial_p\Omega_T) \leq \delta\}; \tag{5.4}$$

then, for all small enough $\sigma > 0$, we have

$$g(x,t,\eta(x,t)) + \sigma\varphi(x,t) \le 0, \qquad (x,t) \in \bar{\Omega}_T. \tag{5.5}$$

This means $G(\eta) + \sigma\varphi \in Q$ for all small $\sigma > 0$. Hence, by the definition of the generalized gradient, we obtain

$$\langle \zeta, \varphi \rangle = 0, \qquad \forall \zeta \in \partial d_Q(G(\eta)) \subset \mathcal{M}(\bar{\Omega}_T). \tag{5.6}$$

In other words, we have (for the above η)

$$\operatorname{supp}\zeta \subset \{(x,t) \in \bar{\Omega}_T \mid d((x,t), \partial_p\Omega_T) \ge \delta\}. \tag{5.7}$$

Since $G'(\eta) = g_y(x,t,\eta(x,t))I$, with I being the identity on Z, we see that (2.10) holds. □

We already know that Q has a nonempty interior in Z, hence it is of codimension 0 in Z. Then, our main result is applicable to this case. Let us state the corresponding result.

THEOREM 5.2. *Let $(\bar{y}, \bar{u})$ be an optimal pair. Then, there exists a constant $\psi^0 \le 0$, a function $\psi \in L^q(0,T; W_0^{1,q}(\Omega))$ $(q < \frac{n+2}{n+1})$, and a $\varphi \in \mathcal{M}_0(\bar{\Omega}_T)$, such that*

$$|\psi^0| + \|\varphi\|_{\mathcal{M}_0(\bar{\Omega}_T)} > 0, \tag{5.8}$$

$$\begin{cases} \psi_t + \displaystyle\sum_{i,j=1}^n (a_{ij}(x,t)\psi_{x_j})_{x_i} = -f_y(x,t,\bar{y}(x,t),\bar{u}(x,t))\psi \\ \qquad\qquad - \psi^0 f_y^0(x,t,\bar{y}(x,t),\bar{u}(x,t)) + g_y(x,t,\bar{y}(x,t))^*\varphi\big|_{\Omega_T}, \qquad \text{in } \Omega_T, \\ \psi \big|_{\partial\Omega} = 0, \\ \psi \big|_{t=T} = g_y(x,T,\bar{y}(x,T))\varphi\big|_{\Omega\times\{T\}}, \end{cases} \tag{5.9}$$

$$\int_{\bar{\Omega}_T} \Big[z(x,t) - g(x,t,\bar{y}(x,t))\Big] d\varphi(x,t) \le 0, \qquad \forall z \in Q. \tag{5.10}$$

$$\begin{aligned} H(x,t,\bar{y}(x,t),&\bar{u}(x,t),\psi^0,\psi(x,t)) = \max_{v\in U} H(x,t,\bar{y}(x,t),v,\psi^0,\psi(x,t)), \\ &a.e.(x,t) \in \Omega \times [0,T], \end{aligned} \tag{5.11}$$

where H is the Hamiltonian defined by (2.15).

Let us make some further remark on the above result. We set

$$\Omega_T^0 = \{(x,t) \in \bar{\Omega}_T \mid g(x,t,\bar{y}(x,t)) = 0\}. \tag{5.12}$$

Then, by our condition (5.1), we see that

$$\Omega_T^0 \subset \Omega_T \bigcup (\Omega \times \{T\}). \tag{5.13}$$

The set Ω_T^0 is called the active set for the optimal state $\bar{y}$. We have that

$$\operatorname{supp}\varphi \subset \Omega_T^0. \tag{5.14}$$

In fact, for any $\eta \in C(\bar{\Omega}_T)$ with $\operatorname{supp}\eta \subset \bar{\Omega}_T \setminus \Omega_T^0$, $g(\cdot,\cdot,\bar{y}(\cdot,\cdot)) \pm \varepsilon\eta \in Q$ if ε is small enough. By the transversality condition (5.10), we see immediately that

$$\int_{\bar{\Omega}_T} \eta(x,t)d\varphi(x,t) = 0. \tag{5.15}$$

This gives (5.14).

The above situation is comparable with the case discussed in [5] for quasilinear elliptic equations.

Next, let us look at another important case. Let $Z = C_0(\bar{\Omega}_T)$ as before and let $(x_i,t_i) \in \Omega_T \bigcup(\Omega \times \{T\})$ $(1 \le i \le m)$ be given m (different) points and also let $b_i \in \mathbb{R}, 1 \le i \le m$. We define Q as in (1.5). Then, we see that Q is a finite codimensional convex and closed subset of Z. Also, it is not hard to see that for any $\eta \in Q$ and any $\zeta \in \partial d_Q(\eta)$, we have

$$\zeta = \sum_{i=1}^{m} \lambda_i \delta_{(x_i t_i)}, \tag{5.16}$$

where $\lambda_i \in \mathbb{R}$ and $\delta_{(x_i,t_i)}$ is the Dirac measure concentrated at point (x_i,t_i) with mass 1. Thus, we see that condition (2.10) holds (in the present case, $G = I$, the identity). Hence, our result applies to this situation. Let us state the corresponding result below.

THEOREM 5.3. *Let $(\bar{y},\bar{u})$ be an optimal pair. Then, there exists a constant $\psi^0 \le 0$, a function $\psi \in L^q(0,T;W_0^{1,q}(\Omega))$ $(q < \frac{n+2}{n+1})$, and real numbers $\lambda_i, 1 \le i \le m$, such that*

$$|\psi^0| + \sum_{i=1}^{m} |\lambda_i| > 0, \tag{5.17}$$

$$\begin{cases} \psi_t + \displaystyle\sum_{i,j=1}^{n} (a_{ij}(x,t)\psi_{x_j})_{x_i} = -f_y(x,t,\bar{y}(x,t),\bar{u}(x,t))\psi \\ \qquad - \psi^0 f_y^0(x,t,\bar{y}(x,t),\bar{u}(x,t)) + \displaystyle\sum_{t_i<T} \lambda_i\delta_{(x_i,t_i)}, \qquad \textit{in } \Omega_T, \\ \psi \mid_{\partial\Omega} = 0, \\ \psi \mid_{t=T} = \displaystyle\sum_{t_i=T} \lambda_i\delta_{(x_i,t_i)}, \end{cases} \tag{5.18}$$

$$\begin{aligned} H(x,t,\bar{y}(x,t),\bar{u}(x,t),\psi^0,\psi(x,t)) &= \max_{v\in U} H(x,t,\bar{y}(x,t),v,\psi^0,\psi(x,t)), \\ &\qquad a.e.(x,t) \in \Omega \times [0,T]. \end{aligned} \tag{5.19}$$

Next, let us point out some other state constraints, which are covered by our general result.

1°. Let $Z = L^p(\Omega_T)$, $1 < p < \infty$. $F : \Omega_T \times \mathbb{R} \to \mathbb{R}$ and

$$\begin{cases} Q = \left\{ z \in L^p(\Omega_T) \mid \displaystyle\int_{\Omega_T} z(x,t)dxdt \le 0 \right\}, \\ G(\eta)(x,t) = F(x,t,\eta(x,t)), \qquad \forall \eta \in C_0(\bar{\Omega}_T). \end{cases} \tag{5.20}$$

Then, the corresponding state constraint is

$$\int_{\Omega_T} F(x,t,\eta(x,t))dxdt \le 0. \tag{5.21}$$

2°. Let $Z = L^p(\Omega_T)^m$, $1 < p < \infty$, $F_i : \Omega_T \times \mathbb{R} \to \mathbb{R}$, $1 \le i \le m$,

$$\begin{cases} Q = \left\{ z \in L^p(\Omega_T)^m \mid \int_{\Omega_T} z(x,t)dxdt = 0 \right\}, \\ G(\eta)(x,t) = (F_1(x,t,\eta(x,t)), \dots, F_m(x,t,\eta(x,t))), \qquad \forall \eta \in C_0(\bar{\Omega}_T). \end{cases} \tag{5.22}$$

Then, the state constraint is

$$\int_{\Omega_T} F_i(x,t,y(x,t))dxdt = 0, \qquad 1 \le i \le m. \tag{5.23}$$

3°. Let $Z = C_0(\bar{\Omega}_T)$, $F : \bar{\Omega}_T \times \mathbb{R} \to \mathbb{R}$ such that

$$F(x_i,t_i,y) = g_i(y), \qquad 1 \le i \le m, \tag{5.24}$$

where $(x_i,t_i) \in \Omega_T \bigcup (\Omega \times \{T\})$ are given different points. Let

$$\begin{cases} Q = \{ z \in C_0(\bar{\Omega}_T) \mid z(x_i,t_i) = b_i, \ 1 \le i \le m \}, \\ G(\eta)(x,t) = F(x,t,\eta(x,t)), \qquad \forall \eta \in C_0(\bar{\Omega}_T). \end{cases} \tag{5.25}$$

Then, the state constraint is

$$g_i(y(x_i,t_i)) = b_i, \qquad 1 \le i \le m. \tag{5.26}$$

This is a generalization of (1.6).

4°. Let $Z = C_0(\bar{\Omega}_T)$, $F : \bar{\Omega}_T \times \mathbb{R} \to \mathbb{R}$ such that (5.24) holds. Let

$$\begin{cases} Q = \{ z \in C_0(\bar{\Omega}_T) \mid z(x_i,t_i) = z(x_j,t_i), \ 1 \le i,j \le m \}, \\ G(\eta)(x,t) = F(x,t,\eta(x,t)), \qquad \forall \eta \in C_0(\bar{\Omega}_T). \end{cases} \tag{5.27}$$

Then, the state constraint is

$$g_i(y(x_i,t_i)) = g_j(y(x_j,t_j)), \qquad 1 \le i,j \le m. \tag{5.28}$$

In particular, if we take $F(x,t,y) = y$, then (5.29) means

$$y(x_i,t_i) = y(x_j,t_j), \qquad 1 \le i,j \le m. \tag{5.29}$$

Physically, this means that we want the temperatures, say, at points (x_i,t_i) to be the same.

There are many other examples, but we prefer to omit them here. We should point out that the pointwise constraint, like (1.6), is actually an approximation of the constraint, like

$$|y(x_i,t_i) - b_i| \le \varepsilon, \qquad 1 \le i \le m, \tag{5.30}$$

with $\varepsilon > 0$ being very small. Physically, this means, for example, that the temperature at point (x_i, t_i) has to be controlled near b_i with an accuracy ε.

6. Quasilinear parabolic equations. Finally, let us remark that the result extends to the quasilinear parabolic equations as those considered in [5] for elliptic cases. The Hölder estimates for the gradient (in the x direction) of the solution are required. These are available [16] when the leading coefficients are assumed to be Hölder continuous in the x direction. After a careful examination of the proof of Lemma 3.2, we conclude that a_{ij} can actually be allowed to depend on ρ.

We consider the equations

$$\begin{cases} \zeta_t - \sum\limits_{i,j=1}^{n} \left(a_{ij}^{\rho}(x,t)\zeta_{x_i}\right)_{x_j} + c_\rho(x,t)\zeta(x,t) = \left(1 - \frac{1}{\rho}\chi_{E_\rho}(x,t)\right) h(x,t), \\ \zeta = 0, \qquad (x,t) \in \partial_p \Omega_T. \end{cases} \tag{6.1}$$

It is clear that the solution $\zeta = \zeta_{E_\rho}(x,t)$ is uniquely determined by the choice of the coefficients a_{ij}^{ρ}, c_ρ and the set E_ρ. Let $K > 0$, $0 < \lambda < \Lambda$, and

$$\begin{aligned} \mathcal{K} = \Bigg\{ &(a_{ij}(x,t), c(x,t)); \ \ \|c\|_{L^\infty(\Omega_T)} \le K, \\ &\lambda|\xi|^2 \le \sum_{i,j=1}^{n} a_{ij}(x,t)\xi_i\xi_j \le \Lambda|\xi|^2, \ \forall (x,t) \in \bar{\Omega}_T, \xi \in \mathbb{R}^n \Bigg\}. \end{aligned} \tag{6.2}$$

LEMMA 6.1. *Suppose that $\frac{n+2}{2} < p < \infty$. Then, there exists $\theta \in (0,1)$ such that for each $h^0 \in L^1(\Omega_T)$, $h \in L^p(\Omega_T)$, $K > 0$, and any $\rho \in (0,1)$*

$$\begin{aligned} \inf_{E_\rho \in \mathcal{E}_\rho} \sup_{(a_{ij}^\rho, c_\rho) \in \mathcal{K}} \Bigg\{ &\left| \int_0^T \int_\Omega \left(1 - \frac{1}{\rho}\chi_{E_\rho}(x,t)\right) h^0(x,t)\,dx\,dt \right| \\ &+ \|\zeta_{E_\rho}\|_{C^{\theta,\theta/2}(\bar{\Omega}\times[0,T])} \Bigg\} = 0. \end{aligned} \tag{6.3}$$

We now consider the following parabolic equation:

$$\begin{cases} y_t - \nabla_x \left(a(x,t,\nabla_x y)\right) = f(x,t,y,u(t,x)), & \text{in } \Omega_T, \\ y \,\big|_{\partial\Omega} = 0, \\ y \,\big|_{t=0} = y_0(x), & x \in \Omega, \end{cases} \tag{6.4}$$

where a satisfies the follwoing assumption.

($\overline{\text{A1}}$) The functions $a : \bar{\Omega}_T \times \mathbb{R}^n \to \mathbb{R}^n$ and $\frac{\partial a_i}{\partial p_j} : \bar{\Omega}_T \times \mathbb{R}^n \to \mathbb{R}$ are continuous. There exist $\Lambda > \lambda > 0$, $\Lambda_1 > 0$, $\alpha \in (0,1)$, and $m > 1$ such that

$$\begin{cases} \dfrac{\partial a_i}{\partial p_j}(x,t,p)\xi_i\xi_j \ge \lambda(1+|p|)^{m-2}|\xi|^2, & \forall (x,t,p) \in \bar{\Omega} \times [0,T] \times \mathbb{R}^n, \\ \left|\dfrac{\partial a_i}{\partial p_j}(x,t,p)\right| \le \Lambda(1+|p|)^{m-2}, & \forall (x,t,p) \in \bar{\Omega} \times [0,T] \times \mathbb{R}^n, \end{cases} \tag{6.5}$$

and

$$
\begin{aligned}
(6.6)\qquad & |a(x,t,p)-a(\widetilde{x},t,p)| \le \Lambda_1(1+|p|)^{m-1}|x-\widetilde{x}|^{\alpha}, \\
& \qquad\qquad \forall (x,t),\ (\widetilde{x},t) \in \bar{\Omega}\times[0,T],\ p\in \mathbb{R}^n.
\end{aligned}
$$

Under the assumptions $(\overline{\text{A1}})$ and (A2), (6.4) has a unique solution $y(x,t)$ such that $y, y_{x_j} \in C^{\beta,\beta/m}(\bar{\Omega}_T)$ for any $y_0 \in C^{1+\alpha}(\bar{\Omega})\bigcap C_0(\bar{\Omega})$ (see [16]). The a priori estimates for y and y_{x_j} in the space $C^{\beta,\beta/m}(\bar{\Omega}_T)$ are valid uniformly for $u \in \mathcal{U}$. If $u, \widehat{u} \in \mathcal{U}$ and $y, \widehat{y}$ are the corresponding states, then by using an interpolation theorem of the $C^{k+\alpha}$ spaces, we have, for any $0<\theta<\beta$,

$$
(6.7)\qquad \|y-\widehat{y}\|_{C(\bar{\Omega}_T)} + \|\nabla_x(y-\widehat{y})\|_{C^{\theta,\theta/m}(\bar{\Omega}_T)} \to 0, \qquad \text{uniformly as } \bar{d}(u,\widehat{u}) \to 0.
$$

Lemma 6.1 and (6.7) are the essential estimates needed to establish the variation theorem similar to Theorem 3.3. We have the following lemma.

LEMMA 6.2. *The same statement of Theorem* 3.3 *is still valid for the quasilinear case, with the $a_{ij}(x,t)$ in* (3.25) *being replaced by $\bar{a}_{ij}(x,t) \equiv \frac{\partial a_i}{\partial p_j}(x,t,\nabla_x y(x,t))$.*

Sketch of Proof. The proof of Theorem 3.3 also goes through, where we shall replace $a_{ij}(x,t)$ in (3.32) with

$$
a^{\rho}_{ij}(x,t) \equiv \int_0^1 \frac{\partial a_i}{\partial p_j}(x,t,\nabla_x y(x,t) + \tau\nabla_x(y_\rho(x,t)-y(x,t))d\tau.
$$

Using (6.7), we can easily obtain

$$
(6.8)\qquad a^{\rho}_{ij} \to \bar{a}_{ij}, \qquad \text{in } C(\bar{\Omega}_T).
$$

Write $z_\rho - z \equiv (z_\rho - \widehat{z}_\rho) + (\widehat{z}_\rho - z)$, where $\widehat{z}_\rho$ satisfies the equation

$$
(6.9)\qquad \begin{cases} (\widehat{z}_\rho)_t - \displaystyle\sum_{i,j=1}^n (a^{\rho}_{ij}(x,t)(\widehat{z}_\rho)_{x_i})_{x_j} + c(x,t)\widehat{z}_\rho = h(x,t), & \text{in } \Omega_T, \\ \widehat{z}_\rho \,\big|_{\partial_p\Omega_T} = 0. \end{cases}
$$

We can treat $z_\rho - \widehat{z}_\rho$ the same way as before, apply Lemma 6.1 instead of Lemma 3.2, and then obtain

$$
(6.10)\qquad \|z_\rho - \widehat{z}_\rho\|_{C^{\theta,\theta/2}(\bar{\Omega}_T)} \to 0 \qquad \text{as } \rho \to 0,
$$

for a special choice of E_ρ with $|E_\rho| = \rho|\Omega_T|$. Clearly, by Hölder's estimates [12, Chap. III, §10], for $\widehat{\zeta}_\rho = \widehat{z}_\rho - z$,

$$
(6.11)\qquad \|\widehat{\zeta}_\rho\|_{C^{\beta,\beta/2}(\bar{\Omega}_T)} \le \|\widehat{z}_\rho\|_{C^{\beta,\beta/2}(\bar{\Omega}_T)} + \|z\|_{C^{\beta,\beta/2}(\bar{\Omega}_T)} \le C,
$$

and $\widehat{\zeta}_\rho$ satisfies the equation

$$
(6.12)\qquad \begin{cases} (\widehat{\zeta}_\rho)_t - \displaystyle\sum_{i,j=1}^n \Big(a^{\rho}_{ij}(x,t)(\widehat{\zeta}_\rho)_{x_i}\Big)_{x_j} + c(x,t)\widehat{\zeta}_\rho \\ \qquad = \Big([a^{\rho}_{ij}(x,t) - \bar{a}_{ij}(x,t)]z_{x_i}(x,t)\Big)_{x_j}, \\ \widehat{\zeta}_\rho \,\big|_{\partial_p\Omega_T} = 0. \end{cases}
$$

Multiplying the above equation with $\widehat{\zeta}_\rho$ and integrating over Ω_T give us the usual energy estimates:

$$\begin{aligned}\|\widehat{\zeta}_\rho\|_{L^2(\Omega_T)} &\le \|\nabla_x\widehat{\zeta}_\rho\|_{L^2(\Omega_T)}\\ &\le C\|a^\rho_{ij}-\bar{a}_{ij}\|_{C(\bar{\Omega}_T)}\|\nabla_x z\|_{L^2(\Omega_T)} \to 0, \qquad \text{as } \rho\to 0,\end{aligned} \tag{6.13}$$

where we have used the fact that $\|\nabla_x z\|_{L^2(\Omega_T)} \le C$, which is an easy consequence of the energy estimates for the solution $z(x,t)$.

Using the interpolation (Lemma 3.4) and the estimates (6.11) and (6.13), we immediately obtain, for any $0<\theta<\beta$,

$$\|\widehat{z}_\rho - z\|_{C^{\theta,\theta/2}(\bar{\Omega}_T)} = \|\widehat{\zeta}_\rho\|_{C^{\theta,\theta/2}(\bar{\Omega}_T)} \to 0, \qquad \text{as } \rho\to 0. \tag{6.14}$$

Combining (6.10) and (6.14), we get a variation theorem like Theorem 3.3. □

Therefore, with the same argument as in §4 (where we used (6.7) for the convergence of the leading coefficients), we have the following maximum principle.

THEOREM 6.3. *Consider the same problem except that the governing equation* (1.1) *is replaced by* (6.4). *Let* $(\overline{A1})$, $(A2)$, *and* $(A3)$ *be in force. Suppose that* $y_0 \in C^{1+\alpha}(\bar{\Omega})\bigcap C_0(\bar{\Omega})$ *and that* (2.10) *holds. Let* $(\bar{y},\bar{u})$ *be an optimal pair of Problem C corresponding to the state equation* (6.4). *Then, there exists a constant* $\psi^0 \le 0$, *a function* $\psi \in L^q(0,T;W^{1,q}_0(\Omega))$ $(1<q<\frac{n+2}{n+1})$, *and a* $\varphi \in \partial d_Q(G(\bar{y})) \subset Z^*$, *such that*

$$|\psi^0| + \|\varphi\|_{Z^*} > 0, \tag{6.15}$$

$$\begin{cases}\psi_t + \displaystyle\sum_{i,j=1}^n \left(\frac{\partial a_i}{\partial p_j}(x,t,\nabla_x\bar{y}(x,t))\psi_{x_j}\right)_{x_i} = -f_y(x,t,\bar{y}(x,t),\bar{u}(x,t))\psi \\ \qquad - \psi^0 f^0_y(x,t,\bar{y}(x,t),\bar{u}(x,t)) + (G'(\bar{y})^*\varphi)\big|_{\Omega_T}, \qquad \text{in } \Omega_T,\\ \psi\,\big|_{\partial\Omega} = 0,\\ \psi\,\big|_{t=T} = (G'(\bar{y})^*\varphi)\big|_{\Omega\times\{T\}},\end{cases} \tag{6.16}$$

$$\langle z - G(\bar{y}), \varphi\rangle \le 0, \qquad \forall z \in Q. \tag{6.17}$$

$$\begin{aligned}H(x,t,\bar{y}(x,t),\bar{u}(x,t),\psi^0,\psi(x,t)) &= \max_{v\in U} H(x,t,\bar{y}(x,t),v,\psi^0,\psi(x,t)),\\ &\qquad a.e.\ (x,t)\in\Omega\times[0,T],\end{aligned} \tag{6.18}$$

where

$$\begin{aligned}H(x,t,y,u,\psi^0,\psi) &= \psi^0 f^0(x,t,y,u) + \psi f(x,t,y,u),\\ &\qquad \forall (x,t,y,u,\psi^0,\psi)\in\Omega\times[0,T]\times\mathbb{R}\times U\times\mathbb{R}\times\mathbb{R}.\end{aligned} \tag{6.19}$$

Remark 6.4. The semilinear case is not a special case of the quasilinear case, since the Hölder continuity of a_{ij} (in the x direction) is not assumed in the semilinear case.

Acknowledgment. The authors thank Dr. Hong-Ming Yin for a stimulating discussion on this problem.

REFERENCES

[1] L. BOCCARDO AND T. GALLOUET, *Non-linear elliptic and parabolic equations involving measure data*, J. Funct. Anal., 87 (1989), pp. 149–169.

[2] J. F. BONNANS AND E. CASAS, *Un principe de Pontryagine pour le contrôle des systèmes semilinéaires elliptiques*, J. Differential Equations, 90 (1991), pp. 288–303.

[3] ———, *A boundary Pontryagin's principle for the optimal control of state-constrained elliptic systems*, Internat. Ser. Numer. Math., 107 (1992), pp. 241–249.

[4] ———, *An extension of Pontryagin's principle for state-constrained optimal control of semilinear elliptic equations and variational inequalities*, SIAM J. Control Optim., 33 (1995), pp. 274–298.

[5] E. CASAS AND J. YONG, *Maximum principle for state-constrained optimal control problems governed by quasilinear elliptic equations*, Differential Integral Equations, 8 (1995), pp. 1–18.

[6] F. H. CLARKE, *Optimization and Nonsmooth Analysis*, Society for Industrial and Applied Mathematics, Philadelphia, 1990.

[7] J. DIESTEL, *Geometry of Banach Spaces — Selected Topics*, Lecture Notes in Math. 485, Springer-Verlag, Berlin, 1975.

[8] I. EKELAND, *Nonconvex minimization problems*, Bull. Amer. Math. Soc. (New Series), 1 (1979), pp. 443–474.

[9] H. O. FATTORINI, *A unified theory of necessary conditions for nonlinear nonconvex control systems*, Appl. Math. Optim., 15 (1987), pp. 141–185.

[10] H. O. FATTORINI AND H. FRANKOWSKA, *Necessary conditions for infinite dimensional control problems*, Math. Control Signal Systems, 4 (1991), pp. 41–67.

[11] D. GILBARG AND N. S. TRUDINGER, *Elliptic Partial Differential Equations of Second Order*, 2nd ed., Springer-Verlag, Berlin, 1983.

[12] O. A. LADYŽENSKAJA, V. A. SOLONNIKOV, AND N. N. URALČEVA, *Linear and Quasi-linear Equations of Parabolic Type*, Transl. Math. Monographs., Vol. 23, American Mathematical Society, Providence, RI, 1968.

[13] X. LI, *Vector–valued measure and the necessary conditions for the optimal control problems of linear systems*, Proc. IFAC 3rd Symposium on Control of Distributed Parameter Systems, Toulouse, France, 1982.

[14] X. LI AND Y. YAO, *Maximum principle of distributed parameter systems with time lags*, Lecture Notes in Control and Inform. Sci., Vol. 75, Springer-Verlag, Berlin, 1985, pp. 410–427.

[15] X. LI AND J. YONG, *Necessary conditions of optimal control for distributed parameter systems*, SIAM J. Control Optim., 29 (1991), pp. 895–908.

[16] G. M. LIEBERMAN, *Boundary and initial regularity for solutions of degenerate parabolic equations*, Nonlinear Anal., TMA, 29 (1993), pp. 551–569.

[17] J. L. LIONS, *Optimal Control of Systems Governed by Partial Differential Equations*, Springer-Verlag, New York, 1971

[18] L. PAN AND J. YONG, *Optimal control for quasilinear retarded parabolic systems*, Math. Systems, Estimations & Control, to appear.

[19] L. S. PONTRYAGIN, V. G. BOLTYANSKII, R. V. GAMKRELIDZE AND E. F. MISCHENKO, *Mathematical Theory of Optimal Processes*, Wiley, New York, 1962.

[20] J. YONG, *Pontryagin maximum principle for semilinear second order elliptic partial differential equations and variational inequalities with state constraints*, Differential Integral Equations, 5 (1992), pp. 1307–1334.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 33, No. 6, pp. 1881–1915, November 1995

© 1995 Society for Industrial and Applied Mathematics

014

RISK-SENSITIVE CONTROL ON AN INFINITE TIME HORIZON*

WENDELL H. FLEMING† AND WILLIAM M. McENEANEY‡

Abstract. Stochastic control problems on an infinite time horizon with exponential cost criteria are considered. The Donsker–Varadhan large deviation rate is used as a criterion to be optimized. The optimum rate is characterized as the value of an associated stochastic differential game, with an ergodic (expected average cost per unit time) cost criterion. If we take a small-noise limit, a deterministic differential game with average cost per unit time cost criterion is obtained. This differential game is related to robust control of nonlinear systems.

Key words. risk-sensitive control, H^∞ control, differential games, viscosity solutions, Hamilton–Jacobi equations, Isaacs equations

AMS subject classifications. 93E20, 93B36, 93C10, 90D25, 60F10, 49L25, 35B37

1. Introduction.

There are various approaches to treating disturbances in control systems. In stochastic control, disturbances are modelled as stochastic processes (random noise). On the other hand, in robust control theory, disturbances are modelled deterministically. The theory of risk-sensitive optimal control provides a link between stochastic and deterministic approaches.

For linear systems with quadratic cost criteria, H^∞-optimization provides a method for robust control design. The disturbance attenuation problem is one of those considered in robust H^∞-control theory. If a state space formulation is used, an associated "soft constrained" differential game arises naturally; see Basar and Bernhard [2]. The stochastic control counterpart is a linear exponential quadratic regulator (LEQR) problem, introduced by Jacobson [24]. This analysis of the LEQR problem leads to the same differential game. Glover and Doyle [20] gave a further connection between the LEQR problem and H^∞-control via a minimum entropy principle.

An interesting question is to find, for nonlinear systems or nonquadratic cost criteria, similar connections between stochastic and robust control approaches to disturbance attenuation problems. Whittle [40], [41] introduced an interesting approach to this question, using large-deviations ideas. Whittle considered problems on a finite-time horizon $0 \leq t \leq T$ and used Freidlin–Wentzell-type "small-noise" asymptotics. In [11] and [25] Whittle's formula for the optimal large-deviations rate was obtained using partial differential equation (PDE)–viscosity solution methods in a special case when the process being controlled is governed by a stochastic differential equation (SDE).

In this paper we are concerned with infinite-horizon risk-sensitive control problems with state-feedback control laws (the "complete state information case.") For these problems a different kind of large-deviations principle, of Donsker–Varadhan type, is needed. Runolfsson [32], [33] used Donsker–Varadhan-type large-deviations ideas to obtain a corresponding stochastic differential game for which the game payoff is an ergodic (expected average cost per unit time) criterion.

* Received by the editors November 18, 1993; accepted for publication (in revised form) July 5, 1994.

† Division of Applied Mathematics, Brown University, Providence, Rhode Island 02912. The research of this author was supported by Air Force Office of Scientific Research grant F49620-92-J-0081, Army Research Office grant DAAL03-92-G-0115, and National Science Foundation grant DMS-9300048.

‡ Department of Mathematics, Carnegie Mellon University, Pittsburgh, Pennsylvania 15213.

As outlined in [12], we approach these problems with different methods. To illustrate the ideas we will consider the following "model problem." Let x_t and u_t denote respectively the state and control at time $t \geq 0$, with $x_t \in I\!R^n, u_t \in U$. ($U$ is the control space.) The state dynamics are

$$dx_t = f(x_t, u_t)dt + \left(\frac{\epsilon}{2\gamma^2}\right)^{\frac{1}{2}} db_t, \tag{1.1}$$

with initial state $x_0 = x$, where $b_.$ is an n-dimensional Brownian motion. Here γ is a given constant and ϵ is a parameter related to the noise intensity. As a criterion to be minimized we take a long run expected rate of exponential growth criterion:

$$\lambda^\epsilon = \epsilon \lim_{T\to\infty} \frac{1}{T} \log E_x \exp\left[\epsilon^{-1}\int_0^T L(x_t, u_t)dt\right], \tag{1.2}$$

where $L \geq 0$. Further assumptions on f, L, and U will be stated later (see (7.1), (7.2)). The risk-sensitive stochastic control problem is to minimize λ^ϵ among all U-valued progressively measurable control processes $u_.$.

To postpone discussions involving stochastic differential games, we begin in §§2–5 with the conceptually simpler situation when the control u_t is absent. For the model corresponding to (1.1)–(1.2), x_t is then a Markov diffusion process satisfying an SDE

$$dx_t = g(x_t)dt + \left(\frac{\epsilon}{2\gamma^2}\right)^{\frac{1}{2}} db_t, \tag{1.3}$$

and the long run expected rate of growth is

$$\lambda^\epsilon = \epsilon \lim_{T\to\infty} \frac{1}{T} \log E_x \exp\left[\epsilon^{-1}\int_0^T \ell(x_t)dt\right]. \tag{1.4}$$

The functions g and ℓ are assumed to satisfy assumptions (3.1). For given $\epsilon > 0$, formula (1.4) is just a particular case of a Donsker–Varadhan large-deviations formula. We wish to give a stochastic control interpretation of λ^ϵ and afterward (§5) take a deterministic limit as $\epsilon \longrightarrow 0$.

If the expectation in (1.4) is denoted by $\phi^\epsilon(T, x)$, then ϕ^ϵ satisfies a PDE of parabolic type related to this expectation via the Feynman–Kac formula. If one formally writes for large T that $\phi^\epsilon \sim \exp[\epsilon^{-1}(\lambda^\epsilon T + W^\epsilon(x))]$, then $\lambda^\epsilon, W^\epsilon(x)$ satisfy formally the dynamic programming equation for a stochastic control problem with expected average cost per unit time criterion. This formalism is explained in §2. Then the model (1.3)–(1.4) is treated rigorously in §§3 and 4. A stochastic control interpretation of λ^ϵ was first given by Holland [21] in a slightly different formulation. The assumptions (3.1) on g and ℓ are stronger than necessary in order to obtain the Donsker–Varadhan formula (1.4). However, they are used to obtain the "cost potential" function $W^\epsilon(x)$ and also in passing to the deterministic limit in §5. Bounds are obtained for λ^ϵ and $|\nabla W^\epsilon|$ which do not depend on ϵ. In the limit, λ^0 and $W^0(x)$ are obtained, with W^0 a viscosity solution of the first-order PDE (5.3) which is the dynamic programming equation for a deterministic average cost per unit time control problem. For applications to robust nonlinear control, a key question is whether $\lambda^0 = 0$ or $\lambda^0 > 0$. If $\lambda_0 = 0$, then a dissipation inequality which is familiar in robust control holds. See (5.15).

In §§6 and 7 we return to considering risk-sensitive stochastic control problems, with the goal of finding a control which minimizes the long-term growth rate λ^ϵ in (1.2). If the control enters (1.1) via a stationary Markov control policy $\underline{u}(x)$, namely, $u_t = \underline{u}(x_t)$, then (1.1)–(1.2) correspond to (1.3)–(1.4) with

$$g(x) = f(x, \underline{u}(x)),\ \ell(x) = L(x, \underline{u}(x)).$$

The corresponding growth rate $\lambda^\epsilon = \lambda^\epsilon(\underline{u})$ depends on the control policy $\underline{u}$. One would like to find a policy $\underline{u}^*$ which minimizes $\lambda^\epsilon(\underline{u})$. In §6 we describe a formalism by which dynamic programming leads to a kind of nonlinear eigenvalue problem for a pair Λ^ϵ, Ψ^ϵ. See (6.3). By making a logarithmic transformation of the positive eigenfunction Ψ^ϵ, equation (6.3) becomes the Isaacs equation (6.10) for a stochastic differential game with an expected average cost per unit time payoff criterion. The number Λ^ϵ has two interpretations: first as the optimal growth rate and second as the value of the stochastic differential game. These results are treated rigorously in §7, for the model problem (1.1)–(1.2). In §8, we consider the deterministic limit $\epsilon \to 0$. As $\epsilon \to 0$, Λ^ϵ tends to a limit Λ^0, which is the value of the corresponding deterministic differential game. Connections with nonlinear robust control are also discussed in §§5 and 8.

Recently Dupuis and Ellis [7] introduced a different technique, based on stochastic control ideas, which is applicable to a wide variety of large-deviations problems.

We consider only risk-sensitive control problems with complete state information in which the current state of the process being controlled is known. For problems with partial state information see [26], [33], [39], [41] and references cited there.

2. Logarithmic transformations and exponential growth. In this section we recall a large-deviations formula (2.1) of Donsker and Varadhan [6] and reformulate it in terms of a stochastic control interpretation. The discussion in the present section is formal, without proofs. In §§3 and 4 we will put things on a rigorous basis in a particular case of interest for disturbance attenuation control problems.

Let x_t be a time-homogeneous Markov process with state space Σ. Thus $x_t \in \Sigma$ for $t \geq 0$. Let ℓ be a bounded, continuous function on Σ, and $\epsilon > 0$ a parameter. Under suitable assumptions on x_t, the following limit exists:

$$\lambda = \epsilon \lim_{T\to\infty} \frac{1}{T} \log E_x \exp\left[\epsilon^{-1} \int_0^T \ell(x_t)dt\right], \tag{2.1}$$

where the subscript x indicates the initial state $x_0 = x$. Among the assumptions a sufficiently strong kind of ergodicity is needed. Moreover, one anticipates that $\epsilon^{-1}\lambda$ can be interpreted as the dominant eigenvalue of the linear operator $G + \epsilon^{-1}\ell$, where G is the generator of the Markov process x_t. See [6, part I] in case Σ is compact and [6, part III] for noncompact Σ. Moreover, considered as a function of ℓ, $\epsilon^{-1}\lambda(\ell)$ is dual (in the sense of convex duality) to the Donsker–Varadhan entropy function $I(\mu)$. This viewpoint is well developed in Stroock [36].

Let

$$\phi(T, x) = E_x \exp\left[\epsilon^{-1} \int_0^T \ell(x_t)dt\right]. \tag{2.2}$$

Then (2.1) implies that $\phi(T, x)$ grows exponentially at rate $\epsilon^{-1}\lambda$ as $T \to \infty$. To obtain heuristically the eigenvalue interpretation of $\epsilon^{-1}\lambda$, we proceed formally as follows. Under suitable assumptions, ϕ satisfies the linear evolution equation

$$\frac{\partial \phi}{\partial T} = G_x \phi + \epsilon^{-1} \ell(x) \phi, \tag{2.3}$$

where $G_x \phi = G\phi(T, \cdot)$. If we formally separate variables, namely,

$$\phi \sim \exp(\epsilon^{-1} \lambda T) \psi(x),$$

then ψ should be a positive eigenfunction corresponding to the dominant eigenvalue $\epsilon^{-1}\lambda$:

$$\epsilon^{-1} \lambda \psi = G\psi + \epsilon^{-1} \ell \psi. \tag{2.4}$$

The Donsker–Varadhan formula (2.1) involves λ but not the eigenfunction ψ. Extra assumptions are generally needed to ensure that ψ exists. However, for compact state space Σ, the eigenfunction ψ exists under the original assumptions in [6, part I]. See Fleming, Sheu, and Soner [14, §4]. The eigenfunction ψ, or more precisely its log transform $W = \epsilon \log \psi$, plays an essential role for the stochastic control interpretation which we use.

In §5, we will let x_t depend on the parameter ϵ in such a way that x_t is nearly deterministic for small ϵ. Thus $G = G^\epsilon, \lambda = \lambda^\epsilon$. The scaling is such that λ^ϵ tends to a limit λ^0 as $\epsilon \to 0$.

Logarithmic transformations. Let $W = \epsilon \log \psi$. Then (2.4) is changed into a nonlinear equation

$$\lambda = \mathcal{H}(W) + \ell, \quad \text{where} \tag{2.5}$$

$$\mathcal{H}(W) = \epsilon \exp(-\epsilon^{-1} W) G[\exp(\epsilon^{-1} W)].$$

For a broad class of generators G, (2.5) has an interpretation as the dynamic programming equation for a stochastic control problem with average cost per unit time cost criterion. Then λ is the optimal expected average cost per unit time, and $W(x)$ is an associated cost potential function. Heuristically, $\epsilon \log \phi(x,T) \approx \lambda T + W(x)$ for large T.

In the stochastic control problem arising from the logarithmic transformation, we will denote by ξ_t the state of the process being controlled and by v_t the control acting at time t. The goal is to find a control process $v_{\cdot}$ which maximizes an average cost per unit time criterion

$$J = \limsup_{T \to \infty} \frac{1}{T} E_x \int_0^T k(\xi_t, v_t) dt, \tag{2.6}$$

where k is a suitably chosen "running cost" function. We will describe the choice of k and the state dynamics for ξ_t only for nondegenerate diffusions in $\mathbb{R}^n$. For other classes of processes (for example, Markov chains or jump Markov processes) see [15, Chaps. 3 and 6] and Sheu [34].

Consider a Markov diffusion x_t with generator G of the following form. Sufficiently strong ergodicity properties will be needed in order for the results outlined formally above to hold.

For $\psi \in C^2(I\!R^n), x = (x_1, \dots, x_n)$, let

$$G\psi(x) = \frac{\epsilon}{2} \sum_{i,j=1}^{n} a_{ij}(x)\psi_{x_i x_j}(x) + g(x) \cdot \nabla\psi(x).$$

In this case

$$\mathcal{H}(W) = GW + \frac{1}{2}a(x)\nabla W \cdot \nabla W,$$

where for any vector $p = (p_1, \dots, p_n)$

$$ap \cdot p = \sum_{i,j=1}^{n} a_{ij} p_i p_j.$$

If the symmetric matrices $a(x) = (a_{ij}(x))$ are positive definite, then

$$\frac{1}{2}ap \cdot p = \max_{v} \left[-\frac{a^{-1}v \cdot v}{2} + v \cdot p \right], \tag{2.7}$$

where the maximum is taken over $\mathcal{V} = I\!R^n$. In this case, (2.5) takes the form

$$\begin{aligned} \lambda &= \frac{\epsilon}{2} \sum_{i,j=1}^{n} a_{ij}(x) W_{x_i x_j} \\ &+ \max_{v}[(g(x) + v) \cdot \nabla W + k(x, v)], \end{aligned} \tag{2.8}$$

where the running cost function is

$$k(x, v) = \ell(x) - \frac{1}{2}a^{-1}(x)v \cdot v. \tag{2.9}$$

If $a(x) = \sigma(x)\sigma^{'}(x)$, then the dynamics of the controlled Markov process ξ_t are governed by the Ito-sense SDE

$$d\xi_t = [g(\xi_t) + v_t]dt + \epsilon^{\frac{1}{2}}\sigma(\xi_t)dw_t, \tag{2.10}$$

with w_t a Brownian motion. We admit any bounded, progressively measurable control process $v_.$, associated with some reference probability system [15, p. 160]. If the control is feedback via a Markov control policy $\underline{v}$ ($v_t = \underline{v}(\xi_t)$), then under suitable technical assumptions on $\underline{v}$ the process ξ_t is Markov. We anticipate that an optimal Markov control policy $\underline{v}^*$ can be found by taking arg max in (2.8), namely, $\underline{v}^* = a\nabla W$.

For the class of problems to be considered in §§3–5, we take $a = (2\gamma^2)^{-1}I$, with $\gamma > 0$ a constant and I the identity matrix. The "drift" coefficient $g(x)$ satisfies assumption (3.1) which ensures a sufficiently strong form of ergodicity of the Markov diffusion x_t needed for the analysis.

3. Ergodic stochastic control problem. In this section and §4 we will put the formalism in §2 on a rigorous basis in the following special case. In (2.8) we take the matrix $a = (a_{ij})$ to be constant and positive definite. By a linear change of coordinates in $I\!R^n$, we may then assume that $a = (2\gamma^2)^{-1}I$, where I is the identity matrix and $\gamma > 0$ is a constant which has a prominent role in robust, deterministic control theory (§5). Assume that $\ell \in C^1(I\!\!R^n), g \in C^1(I\!\!R^n)$. Let g_x denote the matrix of partial derivatives of g. Additionally we assume the following.

$$\text{(3.1)}\qquad \begin{array}{ll} \text{(a)} & \ell,\ \nabla\ell \text{ are bounded and } \ell \geq 0; \\ \text{(b)} & g_x \text{ is bounded;} \\ \text{(c)} & \text{there exists } c > 0 \text{ such that, for all } x, y \in I\!\!R^n, \\ & (x-y)\cdot[g(x)-g(y)] \leq -c|x-y|^2. \end{array}$$

These assumptions are considerably stronger than what is needed to prove the results that follow. However, these strong assumptions will make the proofs much less technical. Proofs of similar results under weaker assumptions can be found in McEneaney [30]. In (3.1a), the assumption that ℓ is bounded can be omitted. See Remark 4.4 below. The assumption that $\nabla\ell$ is bounded allows linear growth of $\ell(x)$ as $|x| \to \infty$ but not quadratic growth. Some issues concerning quadratically growing $\ell(x)$ are discussed at the end of §5. Assumption (3.1b) can be replaced by a weaker assumption, involving one-sided bounds [30]. However, assumption (3.1c) plays a crucial role in the proof of Theorem 3.3 and in passage to a deterministic limit in §5. By the mean value theorem, (3.1c) has the equivalent form

$$\text{(3.2)}\qquad z \cdot g_x(x) z \leq -c|z|^2 \text{ for all } x, z \in I\!\!R^n.$$

For the results of the present section to hold, it may suffice to require (3.2) only for x outside some bounded set. See Remark 4.5. However, to obtain the dissipation inequality (5.15) of deterministic robust control theory, we need (3.2) for all $x \in I\!\!R^n$.

We consider the Markov diffusion process x_t governed by the SDE

$$\text{(3.3)}\qquad dx_t = g(x_t)dt + \left(\frac{\epsilon}{2\gamma^2}\right)^{\frac{1}{2}} db_t, \qquad t \geq 0,$$

with the initial data

$$\text{(3.4)}\qquad x_0 = x.$$

This SDE has a strong (pathwise) solution for any reference probability system [28] $\nu = (\Omega, \{\mathcal{F}_t\}, P, b_.)$, where Ω is a sample space, $\{\mathcal{F}_t\}$ a filtration, P a probability measure, and $b_.$ a P-Brownian motion adapted to $\{\mathcal{F}_t\}$. The generator G of x_t is

$$\text{(3.5)}\qquad G\psi = \frac{\epsilon}{4\gamma^2}\Delta\psi + g \cdot \nabla\psi.$$

Assumptions (3.1b) and (3.1c) imply that x_t is ergodic. More than that, we will show that they insure that the Donsker–Varadhan large-deviations formula (2.1) holds and that the positive eigenfunction ψ in (2.4) exists.

Let us now introduce controlled processes ξ_t, already discussed in an imprecise way in §2. Let $\mu = (\Omega, \{\mathcal{F}_t\}, P, w_.)$ be some reference probability system, which may or may not be the same as the reference probability system ν for (3.3). We admit as control processes all $I\!\!R^n$-valued $\mathcal{F}_t$-progressively measurable processes $v_.$ such that v_t is bounded. Let $\mathcal{W}_M$ denote the set of all such $v_.$ for which $|v_t| \leq M$ for all $t \geq 0$. Of course, $\mathcal{W}_M$ depends on μ but our notation does not make this explicit. Then

$$\mathcal{W} = \bigcup_{M>0} \mathcal{W}_M$$

is the set of all admissible control processes. Given $v_{.}$, let ξ_t be the solution to

$$d\xi_t = [g(\xi_t) + v_t]dt + \left(\frac{\epsilon}{2\gamma^2}\right)^{\frac{1}{2}} dw_t \tag{3.6}$$

with $\xi_0 = x$. The goal is to choose $v_{.} \in \mathcal{W}$ to maximize the criterion J in (2.6).

LEMMA 3.1. *Let $F(x) = \exp[K\sqrt{1+|x|^2}]$, where $K > 0$ is any constant. Then there exists C_1 (depending on x and previously introduced constants) such that*

$$E_x F(\xi_t) \le C_1 \text{ for all } t < \infty.$$

The proof is rather standard and is postponed to Appendix A.

Remark 3.2. Comparing (3.3) and (3.6), we see that we also have $E_x F(x_t) \le C_1$ for all $t < \infty$. Furthermore, this bound implies a bound for all moments of ξ_t and x_t for all $t < \infty$.

We proceed to state the main results of this section. The dynamic programming equation (2.5) now takes the form

$$\lambda = \frac{\epsilon}{4\gamma^2}\Delta W + g(x)\cdot\nabla W + \frac{1}{4\gamma^2}|\nabla W|^2 + \ell(x), \tag{3.7}$$

or equivalently (see (2.8)–(2.9))

$$\lambda = \frac{\epsilon}{4\gamma^2}\Delta W + \max_{v\in I\!R^n}[(g(x)+v)\cdot\nabla W + k(x,v)], \tag{3.7$'$}$$

where

$$k(x,v) = \ell(x) - \gamma^2|v|^2. \tag{3.8}$$

Let $\|\cdot\|$ denote the sup norm.

THEOREM 3.3. *There exist $\lambda \in I\!R$, $W \in C^2(I\!R^n)$ such that (3.7) holds. Moreover, there exists B (not depending on ϵ) such that $|\nabla W| \le B$ and $0 \le \lambda \le \|\ell\|$.*

This theorem is similar to one by Bensoussan [3, Thm. 7.1]. Unfortunately, the presence of the (maximizing) controller taking values in all of $\mathcal{R}^n$ and the unbounded cost (at least in the control but also possibly in the state as in Remark 4.4) prevents a direct application of the Bensoussan result. The addition of a second (minimizing) controller in §§6–8 will further complicate matters.

The proof will give $B = c^{-1}\|\nabla\ell\|$, with c the constant in assumption (3.1c). We postpone the proof of Theorem 3.3 to §4. In (2.6) let us write $J = J(x, v_{.})$ to indicate dependence on the initial state x and the control process $v_{.}$. The maximum in (3.7$'$) is attained at $\underline{v}^*(x) = (2\gamma^2)^{-1}\nabla W(x)$. This gives an optimal Markov control policy for the control problem with the average cost per unit time criterion J, as is seen by the following corollary to Theorem 3.3. Define ξ_t^* as the solution to

$$d\xi_t^* = [g(\xi_t^*) + \underline{v}^*(\xi_t^*)]dt + \left(\frac{\epsilon}{2\gamma^2}\right)^{\frac{1}{2}} dw_t$$

with $\xi_0^* = x$. The control process $v_{.}^*$ defined by

$$v_t^* = \underline{v}^*(\xi_t^*) = (2\gamma^2)^{-1}\nabla W(\xi_t^*)$$

belongs to $\mathcal{W}$ since ∇W is bounded.

COROLLARY 3.4 (a) $J(x, v_.) \le \lambda$ *for all* $x \in I\!\!R^n$, $v_. \in \mathcal{W}$.

(b) $J(x, v_.^*) = \lambda$ *for all* $x \in I\!\!R^n$.

Proof of (a). We apply Ito's rule to obtain

$$E_x W(\xi_T) = W(x) + E_x \int_0^T \left\{ [g(\xi_t) + v_t] \cdot \nabla W(\xi_t) + \frac{\epsilon}{4\gamma^2} \Delta W(\xi_t) \right\} dt.$$

Note that since $\ell(x)$ and $\nabla W(x)$ are bounded, $|\Delta W(x)|$ grows at most linearly with $|x|$ as $|x| \to \infty$, by (3.1b) and (3.7). Therefore, by (3.7′)

$$E_x W(\xi_T) \le W(x) - E_x \int_0^T k(\xi_t, v_t) dt + \lambda T, \tag{3.9}$$

which yields

$$\lambda \ge J(x, v_.) + \limsup_{T \to \infty} \left[\frac{E_x W(\xi_T) - W(x)}{T} \right]. \tag{3.10}$$

Since ∇W is bounded,

$$|W(\xi_T)| \le k_1 |\xi_T| + k_2$$

for suitable k_1, k_2. Since $E_x |\xi_T|$ is bounded, by Remark 3.2, the last term on the right side of (3.10) is 0. This proves (a).

Proof of (b). In (3.9) equality now holds when $v_. = v_.^*$, $\xi_. = \xi_.^*$. We repeat the argument for part (a). □

Without referring to the controlled process ξ_t, we also obtain from Theorem 3.3 that λ is the Donsker–Varadhan large-deviations rate in formula (2.1). Also note that the function $\psi = \exp(\epsilon^{-1} W)$ satisfies the eigenfunction equation (2.4), with generator G as in (3.5).

THEOREM 3.5. *If* λ, W *are as in Theorem* 3.3, *then*

$$\lambda = \lim_{T \to \infty} \frac{\epsilon}{T} \log E_x \exp \left[\frac{1}{\epsilon} \int_0^T \ell(x_t) dt \right].$$

Proof. Let $\nu = (\Omega, \{\mathcal{F}_t\}, P, b_.)$ be a reference probability system, and let $x_.$ be the corresponding solution to (3.3)–(3.4). Let $v_t = (2\gamma^2)^{-1} \nabla W(x_t)$, which is $\mathcal{F}_t$-progressively measurable and bounded since ∇W is bounded. From (3.3)–(3.4)

$$\begin{aligned} x_t &= x + \int_0^t [g(x_r) + v_r] dr + \left(\frac{\epsilon}{2\gamma^2} \right)^{\frac{1}{2}} b_t - \int_0^t v_r dr \\ &= x + \int_0^t [g(x_r) + v_r] dr + \left(\frac{\epsilon}{2\gamma^2} \right)^{\frac{1}{2}} b_t^0, \end{aligned} \tag{3.11}$$

which defines b^0.

Since $v_\cdot$ is bounded, we may use Girsanov's theorem [27] to obtain a probability measure P^0 under which $b^0_\cdot$ is a Brownian motion. In fact,

$$\begin{aligned} P^0(d\omega) &= \exp\left[\sqrt{\frac{2\gamma^2}{\epsilon}}\int_0^t v_r \cdot db_r - \frac{\gamma^2}{\epsilon}\int_0^t |v_r|^2 dr\right] P(d\omega) \\ &= \exp\left[\sqrt{\frac{2\gamma^2}{\epsilon}}\int_0^t v_r \cdot db^0_r + \frac{\gamma^2}{\epsilon}\int_0^t |v_r|^2 dr\right] P(d\omega). \end{aligned}$$

Under this change of measure

$$\begin{aligned} E_x \exp\left[\frac{1}{\epsilon}\int_0^T \ell(x_t)dt\right] = E^0_x \exp\Biggl\{\frac{1}{\epsilon}\biggl[\int_0^T (\ell(x_t) - \gamma^2|v_t|^2)dt \\ - \sqrt{2\gamma^2\epsilon}\int_0^T v_t \cdot db^0_t\biggr]\Biggr\}, \end{aligned} \tag{3.12}$$

where E^0_x indicates expectation with respect to P^0.

On the other hand, by (3.7), (3.11), the definition of v_t, and Ito's rule

$$\begin{aligned} W(x_T) - W(x) = \int_0^T [\lambda - \ell(x_t) + \gamma^2|v_t|^2]dt \\ + \sqrt{2\gamma^2\epsilon}\int_0^T v_t \cdot db^0_t. \end{aligned} \tag{3.13}$$

Combining (3.12) and (3.13) yields

$$E_x \exp\left[\frac{1}{\epsilon}\int_0^T \ell(x_t)dt\right] = \exp\left[\frac{\lambda T + W(x)}{\epsilon}\right] E^0_x \exp\left[\frac{-W(x_T)}{\epsilon}\right],$$

which implies

$$\lim_{T\to\infty} \frac{\epsilon}{T}\log E_x \exp\left[\frac{1}{\epsilon}\int_0^T \ell(x_t)dt\right] = \lambda + \lim_{T\to\infty}\frac{\epsilon}{T}\log E^0_x \exp\left[\frac{-W(x_T)}{\epsilon}\right] \tag{3.14}$$

(where we anticipate the existence of the limit which follows below).

Now note that by assumption, there exists $K < \infty$ such that

$$-K\sqrt{1+|z|^2} \le \frac{-W(z)}{\epsilon} \le K\sqrt{1+|z|^2} \quad \forall z \in I\!R^n.$$

(The constant K depends on ϵ, which is fixed throughout this section.)

By Lemma 3.1 and Jensen's inequality $(E^0(Z^{-1}) \ge (E^0 Z)^{-1})$ this implies that there exists $C_1 < \infty$ such that

$$\frac{1}{C_1} \le E^0_x \exp\left[\frac{-W(x_T)}{\epsilon}\right] \le C_1 \quad \text{for all } T < \infty.$$

Employing this in (3.15) yields the result. □

4. Proof of Theorem 3.3. In this section we prove existence of $\lambda, W(x)$ satisfying the dynamic programming PDE for the average cost per unit time control problem governed by (3.6) with running cost function (3.8). This is done by a rather standard technique. We consider the corresponding infinite-horizon discounted-cost problem, with small discount factor $\rho > 0$. Let $W_\rho(x)$ be the value function for this problem. Once an a priori bound for $\nabla W_\rho(x)$ is found, Theorem 3.3 is obtained by letting $\rho \to 0$ and using standard estimates for elliptic PDEs.

We begin by considering finite-horizon discounted-cost problems. Let

$$J_\rho(T,x;v_.) = E_x \int_0^T e^{-\rho t} k(\xi_t, v_t)dt, \tag{4.1}$$

where ξ_t is the solution to (3.6) with $\xi_0 = x$.

LEMMA 4.1. *For every $T > 0, x, y \in I\!R^n$, and $v_. \in \mathcal{W}$*

(a) $\rho J_\rho(T,x;v_.) \leq \|\ell\|$,

(b) $|J_\rho(T,x;v_.) - J_\rho(T,y;v_.)| \leq \frac{\|\nabla \ell\|}{\rho + c}$.

Proof. Part (a) is immediate from (4.1). To prove (b), let $\xi^x_.$ and $\xi^y_.$ denote the solutions to (3.6) with initial data $\xi^x_0 = x$, $\xi^y_0 = y$. By applying Ito's rule to $|\xi^x_t - \xi^y_t|^2$, one easily obtains, using (3.1c),

$$|\xi^x_t - \xi^y_t|^2 \leq |x-y|^2 - 2c\int_0^t |\xi^x_s - \xi^y_s|^2 ds.$$

Gronwall's inequality gives

$$|\xi^x_t - \xi^y_t|^2 \leq e^{-2ct}|x-y|^2.$$

Since

$$|J_\rho(T,x;v_.) - J_\rho(T,y;v_.)| \leq \|\nabla \ell\| E_x \int_0^T e^{-\rho t}|\xi^x_t - \xi^y_t|dt,$$

we get part (b). □

LEMMA 4.2. *For $T^{'} > T$,*

$$|J_\rho(T^{'},x;v_.) - J_\rho(T,x;v_.)| \leq \rho^{-1}(\|\ell\| + \gamma^2\|v_.\|^2)(e^{-\rho T} - e^{-\rho T^{'}}).$$

Lemma 4.2 is immediate from (4.1). Let us fix a bound $|v_t| \leq M$ for the controls, and denote by $\mathcal{W}_M$ the corresponding class of admissible control processes. Later we will show that the choice of M is arbitrary if M is large enough (in fact, if $M \geq (2\gamma^2)^{-1}B$ with B the bound for ∇W in Theorem 3.3). Consider the value function

$$V_\rho(T,x) = \sup_{\mathcal{W}_M} J_\rho(T,x;v_.). \tag{4.2}$$

By results about parabolic PDEs and a verification theorem, V_ρ is a solution to the dynamic programming equation

$$\begin{aligned} &\frac{\partial V_\rho}{\partial T} + \rho V_\rho = \frac{\epsilon}{4\gamma^2}\Delta_x V_\rho + \max_{|v|\leq M}[(g(x)+v)\cdot \nabla_x V_\rho + k(x,v)], \\ &V_\rho(0,x) = 0. \end{aligned} \tag{4.3}$$

Moreover, V_ρ and the partial deviatives $V_{\rho T}, V_{\rho x_i}$, $V_{\rho x_i x_j}$, $i, j = 1, \ldots, n$, are continuous. See Appendix B. Since $\ell \geq 0, J_\rho(T, x; 0) \geq 0$. Hence, $V_\rho \geq 0$. Lemma 4.1 implies the estimates

$$\text{(4.4)} \qquad \begin{aligned} &\text{(a)} \quad 0 \leq \rho V_\rho \leq \|\ell\|, \\ &\text{(b)} \quad |\nabla_x V_\rho| \leq \frac{\|\nabla \ell\|}{\rho + c}. \end{aligned}$$

Lemma 4.2 implies that

$$\text{(4.5)} \qquad |V_\rho(T', x) - V_\rho(T, x)| \leq \rho^{-1}(\|\ell\| + \gamma^2 M^2)(e^{-\rho T} - e^{-\rho T'}).$$

By (4.5) the limit

$$\text{(4.6)} \qquad W_\rho(x) = \lim_{T \to \infty} V_\rho(T, x)$$

exists. Moreover, $W_\rho(x)$ is the value function for the corresponding infinite-horizon discounted control problem. From (4.4), for each $x, y \in I\!R^n$

$$\text{(4.7)} \qquad \begin{aligned} &\text{(a)} \quad 0 \leq \rho W_\rho(x) \leq \|\ell\|, \\ &\text{(b)} \quad |W_\rho(x) - W_\rho(y)| \leq \frac{\|\nabla \ell\|}{\rho + c}|x - y|. \end{aligned}$$

Since the right side of (4.7b) is less than $c^{-1}\|\nabla \ell\||x - y|$, the bounds in (4.7) do not depend on ρ. (They also do not depend on ϵ.) If in (4.5) we set $T' = T + h$ and let $h \to 0$, we obtain

$$\text{(4.8)} \qquad |(V_\rho)_T| \leq \rho^{-1}(\|\ell\| + \gamma^2 M^2)e^{-\rho T}.$$

Thus $(V_\rho)_T$ tends uniformly to 0 as $T \to \infty$. Since V_ρ satisfies the PDE (4.3), $\Delta_x V_\rho$ is also bounded uniformly on compact sets. (This bound may depend on ϵ but not on ρ.) From (4.6)–(4.8) and standard estimates for semilinear parabolic PDEs (see Appendix B) $W_\rho \in C^2(I\!R^n)$ and satisfies the steady-state form of (4.3):

$$\text{(4.9)} \qquad \begin{aligned} \rho W_\rho = &\frac{\epsilon}{4\gamma^2}\Delta W_\rho \\ &+ \max_{|v| \leq M}[(g(x) + v) \cdot \nabla W_\rho + k(x, v)]. \end{aligned}$$

Let $B = c^{-1}\|\nabla \ell\|$. By (4.7b), $|\nabla W_\rho(x)| \leq B$ for each $x \in I\!R^n$. We take $M \geq (2\gamma^2)^{-1}B$. Then the maximum in (4.9) is an interior max, achieved at $\underline{v}_\rho^*(x) = (2\gamma^2)^{-1}|\nabla W_\rho(x)|$. Thus, W_ρ satisfies the same equation as (4.9), with max over $|v| \leq M$ replaced by max over $I\!R^n$.

Proof of Theorem 3.3. Fix any "reference point" $x_0 \in I\!R^n$. By (4.7), $\rho W_\rho(x_0)$ is uniformly bounded and the functions $W_\rho(x) - W_\rho(x_0)$ are equicontinuous. If we use Ascoli's theorem there is a sequence ρ_m tending to 0 as $m \to \infty$ such $\rho_m W_{\rho_m}(x_0)$ tends to a limit λ and $W_{\rho_m}(x) - W_{\rho_m}(x_0)$ tends to a limit $W(x)$ uniformly on compact sets. We have

$$\text{(4.10)} \qquad \begin{aligned} &\text{(a)} \quad 0 \leq \lambda \leq \|\ell\|, \\ &\text{(b)} \quad |W(x) - W(y)| \leq \tfrac{\|\nabla \ell\|}{c}|x - y| \end{aligned}$$

for all $x, y \in I\!R^n$. Since $\rho_m W_{\rho_m}$ and ∇W_{ρ_m} are bounded independent of m, ΔW_{ρ_m} is also bounded on compact sets independent of m. This implies a Hölder estimate for ∇W_{ρ_m} which is uniform on compact sets. Standard arguments for elliptic PDEs [19] then give that $W \in C^2(I\!R^n)$ and that λ, W satisfy (3.7). This proves Theorem 3.3. □

Remark 4.3. According to Corollary 3.4 (or Theorem 3.5), the number λ in Theorem 3.3 is unique. We anticipate that $W(x)$ is unique up to an additive constant, i.e. that the corresponding solution $\psi = \exp(\epsilon^{-1} W)$ to equation (2.4) is unique up to a positive multiplicative constant. (This has since been proved. See [42, Thm. 3.1].)

Remark 4.4. In all of the results of §3, the assumption that $\ell(x)$ is bounded in (3.1a) can actually be omitted. The remaining assumptions in (3.1a) imply that, for suitable C,

$$0 \le \ell(x) \le C(1 + |x|).$$

The fact that there exists $K(x)$ such that $E_x|\xi_t| \le K(x)$ for any solution ξ_t of (3.6) corresponding to $v_. \in \mathcal{W}_M$ depends on assumptions (3.1b) and (3.1c). In the definition (4.1) of J_ρ let us replace ℓ by ℓ_ρ, where $\ell_\rho(x)$ is bounded:

$$\|\nabla \ell_\rho\| \le \|\nabla \ell\|,$$

$$0 \le \ell_\rho(x) \le C(1 + |x|),$$

$$\ell_\rho(x) = \ell(x) \text{ if } |x| \le \rho^{-1}.$$

In Lemma 4.1, inequality (a) is replaced by

$$\rho J_\rho(T, x; v_.) \le C(1 + K(x)),$$

and inequality (b) is unchanged. In (4.7), inequality now becomes

$$0 \le \rho W_\rho(x) \le C(1 + K(x)),$$

and inequality (b) is unchanged. The proof of Theorem 3.3 proceeds as before, with (4.10a) replaced by

$$0 \le \lambda \le C(1 + K(x_0)).$$

The proofs of Corollary 3.4 and Theorem 3.5 relied on boundedness of ∇W, which still is true without boundedness of $\ell(x)$.

Remark 4.5. The crucial step in the proof of Theorem 3.3 was to get a priori bounds for ρW_ρ and ∇W_ρ (see (4.7)). The bound $0 \le \rho W_\rho \le \|\ell\|$ was immediate. We conjecture that it suffices to assume (3.2) only outside some bounded set in order for a bound $|\nabla W_\rho| \le B$ to hold. This is seen to be correct in dimension $n = 1$ by the following argument. Assume that $g_x(x) \le -c < 0$ if $|x| \ge r$. Since $W_\rho(x)$ is bounded on $I\!R^1$ there exist sequences tending to $\pm\infty$ for which $(W_\rho)_x(x)$ tends to 0. To obtain an a priori bound for $(W_\rho)_x$ it suffices to do so at inflection points (local maxima and minima of $(W_\rho)_x$.) At an inflection point x_0 with $|x_0| \le r$, we have $(W_\rho)_{xx}(x_0) = 0$. Since

$$\rho W_\rho = \frac{\epsilon}{4\gamma^2}(W_\rho)_{xx} + g(x)(W_\rho)_x + \frac{1}{4\gamma^2}(W_\rho)_x^2 + \ell(x) \tag{4.11}$$

and ρW_ρ is bounded, the quadratic formula gives a bound $|(W_\rho)_x(x_0)| \leq B_1$. If $|x_0| > r$, we differentiate (4.11). Then $Z = (W_\rho)_x$ satisfies

$$\rho Z = \frac{\epsilon}{4\gamma^2} Z_{xx} + g_x Z + g Z_x + \frac{1}{2\gamma^2} Z Z_x + \ell_x. \tag{4.12}$$

If Z has a positive local maximum at x_0, then $Z_x(x_0) = 0$, $Z_{xx}(x_0) \leq 0$. Hence $(\rho+c)Z(x_0) \leq (\rho - g_x(x_0))Z(x_0) \leq \|\ell_x\|$. A similar estimate holds if Z has a negative local minimum at x_0. Thus

$$|(W_\rho)_x| \leq \frac{\|\ell_x\|}{\rho + c} \quad \text{if } |x| \geq r. \tag{4.13}$$

This is the same bound obtained in (4.7b). When combined with the bound above for $|x| \leq r$, we obtain $|(W_\rho)_x| \leq B$ for some B which does not depend on ρ.

5. Limiting deterministic control problem. We now let the noise intensity in (3.6) tend to 0. Thus, in (3.6) we consider $\epsilon \to 0$ with $\gamma > 0$ fixed. The formal limit of the stochastic control problem in §3 is then a deterministic problem with average cost per unit time criterion. The deterministic analogue of (3.7) is the first-order PDE (5.3), and the cost potential function $W^0(x)$ turns out to be a solution of (5.3) in the viscosity sense [4], [15].

Let us indicate the dependence of $\lambda, W(x)$ on ϵ in §§3 and 4 by relabelling them as $\lambda^\epsilon, W^\epsilon(x)$. According to (4.10)

$$0 \leq \lambda^\epsilon \leq \|\ell\|, \ |\nabla W^\epsilon(x)| \leq \frac{\|\nabla \ell\|}{c}. \tag{5.1}$$

By Ascoli's theorem, there is a sequence $\epsilon_m \to 0$ such that as $m \to \infty$

$$\lambda^{\epsilon_m} \to \lambda^0, \ W^{\epsilon_m}(x) \to W^0(x). \tag{5.2}$$

The convergence of W^{ϵ_m} is uniform on compact subsets of $I\!R^n$. It is easy to show that $W^0(x)$ is a viscosity solution to the corresponding first-order PDE

$$\lambda^0 = \max_{v \in I\!R^n} [(g(x) + v) \cdot \nabla W^0(x) + k(x,v)]. \tag{5.3}$$

See [15, Chap. 2]. Thus we have the following theorem.

THEOREM 5.1. *There exist $\lambda^0 \geq 0$ and Lipschitz continuous W^0 such that W^0 is a viscosity solution of* (5.3).

By (3.8) the PDE (5.3) can be rewritten in the form

$$\lambda^0 = g(x) \cdot \nabla W^0(x) + \frac{1}{4\gamma^2} |\nabla W^0(x)|^2 + \ell(x). \tag{5.4}$$

If B is a Lipschitz constant for W^0, then in (5.3) the max over $v \in I\!R^n$ can be replaced by the max over $|v| \leq M$, if $M \geq (2\gamma^2)^{-1}B$. For W^0 obtained as the limit of W^{ϵ_m} as above, we can take $B = c^{-1}\|\nabla \ell\|$ according to (5.1). Since W^0 is a Lipschitz continuous viscosity solution, the gradient $\nabla W^0(x)$ exists and (5.3) holds for almost all $x \in I\!R^n$ However, W^0 may not have continuous first-order partial derivatives, and thus W^0 need not be a solution to (5.3) in the classical sense. By Corollary 5.3 below, λ^0 is unique.

Let us now consider the deterministic control problem which is the formal limit of the stochastic control problem in §3 as $\epsilon \to 0$. The state dynamics are

$$\frac{d\xi_t^0}{dt} = g(\xi_t^0) + v_t \tag{5.5}$$

with $\xi_0^0 = x$. The control $v_.$ is a bounded Lebesgue-measurable function of $[0, \infty)$, such that $|v_t| \leq M$ for some fixed (sufficiently large) M. As in the discussion above, we require that $M \geq (2\gamma^2)^{-1}B$, with B a Lipschitz constant for W^0. Let $\mathcal{W}_M^0$ denote the set of all such admissible deterministic control functions $v_.$. The running cost function is $k(x, v)$ given by (3.8).

THEOREM 5.2. *If W^0 is a Lipschitz continuous solution to* (5.3), *then for every* $T < \infty$

$$W^0(x) = \sup_{v_. \in \mathcal{W}_M^0} \left[\int_0^T k(\xi_t^0, v_t)dt + W^0(\xi_T^0) \right] - \lambda^0 T. \tag{5.6}$$

Proof. Denote the right side by $\tilde{W}(T, x)$. Then $\tilde{W}$ is the value function of the finite-time-horizon control problem on $0 \leq t \leq T$, with running cost function $k - \lambda^0$ and terminal cost function W^0. Moreover, $\tilde{W}$ is Lipschitz continuous on $[0, T_0] \times \mathbb{R}^n$ for any $T_0 < \infty$. See [15, §4.8]. Therefore, by standard results [15, Chap. 2], $\tilde{W}$ is a viscosity solution of the time-dependent PDE

$$\tilde{W}_T = \max_{|v| \leq M} [(g(x) + v) \cdot \nabla_x \tilde{W} + k(x, v) - \lambda^0] \tag{5.7}$$

with initial data

$$\tilde{W}(0, x) = W^0(x), \; x \in \mathbb{R}^n. \tag{5.8}$$

However, W^0 is also a (stationary) Lipschitz continuous viscosity solution to (5.7)–(5.8). By a uniqueness theorem for viscosity solutions, (see [5, Thm. 7.2] among others) $\tilde{W} = W^0$. □

Since M is arbitrary, subject only to $M \geq (2\gamma^2 c)^{-1} \|\nabla \ell\|$, in (5.6) we can replace $\mathcal{W}_M^0$ by the class $\mathcal{W}^0$ of all bounded measurable $v_.$. Let us consider the finite-time control problem, with running cost $k(x, v)$ and zero terminal cost. Let $V^0(T, x)$ be the value function

$$V^0(T, x) = \sup_{v_. \in \mathcal{W}_M^0} \int_0^T k(\xi_t^0, v_t)dt. \tag{5.9}$$

COROLLARY 5.3.

$$\lambda^0 = \lim_{T \to \infty} \frac{1}{T} V^0(T, x) \tag{5.10}$$

uniformly for x in any compact set.

Proof. Let $|x| \leq R$ and $v_. \in \mathcal{W}_M^0$. By assumption (3.1c) there exists $R_1 \geq R$ such that $|\xi_t^0| \leq R_1$ for all $T \geq 0$. In (5.6) we divide by T and let $T \to \infty$. Then $T^{-1}W^0(x)$ tends to 0, and $T^{-1}W^0(\xi_t^0)$ tends to 0 uniformly with respect to $v_. \in \mathcal{W}_M^0$ and $x = \xi_0$ in the ball $\{|x| \leq R\}$. □

Let us next show that λ^0 is the maximum of the infinite-horizon average cost per unit time criterion $J^0(x; v_.)$, which is the deterministic version of (2.6). Let

$$J^0(x; v_.) = \limsup_{T \to \infty} \frac{1}{T} \int_0^T k(\xi^0, v_t)dt. \tag{5.11}$$

We have the following theorem.

THEOREM 5.4. *For any $x \in I\!\!R^n$*

$$\lambda^0 = \sup_{v_. \in \mathcal{W}^0_M} J^0(x; v_.). \tag{5.12.}$$

Proof. For any $v_. \in \mathcal{W}^0_M$ and $T < \infty$ (5.6) implies that

$$W^0(\xi^0_T) - W^0(x) \leq - \int_0^T k(\xi^0_t, v_t)dt + \lambda^0 T. \tag{5.13}$$

By dividing by T and letting $T \to \infty$, we obtain as in the proof of Corollary 5.3 that λ^0 is no less than the right side of (5.12). It remains to show that $v_.$ can be chosen so that $J^0(x, v_.)$ is arbitrarily close to λ^0. Fix x, with $|x| \leq R$. As in the proof of Corollary 5.3, $|\xi^0_t| \leq R_1$ for some $R_1 \geq R$. Given $\delta > 0$, by Corollary 5.3 there exists T_0 such that

$$|V^0(T_0, y) - \lambda^0 T_0| < \frac{\delta T_0}{2} \text{ for } |y| \leq R_1. \tag{5.14}$$

On $[0, T_0)$ we choose v_t such that

$$\int_0^T k(\xi^0_t, v_t)dt > V^0(T_0, x) - \frac{\delta T_0}{2},$$

which by (5.14) exceeds $\lambda^0 T_0 - \delta T_0$. Proceeding by induction on $N = 1, 2, \ldots$, we choose v_t on $[NT_0, (N+1)T_0)$ such that

$$\int_{NT_0}^{(N+1)T_0} k(\xi^0_t, v_t)dt > V^0(T_0, \xi^0_{NT_0}) - \frac{\delta T_0}{2} > \lambda^0 T_0 - \delta T_0.$$

By summing from 0 to $N - 1$,

$$\frac{1}{NT_0} \int_0^{NT_0} k(\xi^0_t, v_t)dt > \lambda^0 - \delta.$$

Since $k(x, v)$ is bounded, this implies that

$$J^0(x; v_.) \geq \lambda^0 - \delta$$

as required. □

Robust regulation of nonlinear systems. Let us now relate these results to recent work on robust control approaches to the disturbance attenuation problem for nonlinear systems. See [22], [23], [37], for example. As mentioned earlier, we consider in this paper only the complete state observation case. We claim that, from the viewpoint of robust regulation, a key question is whether $\lambda^0 = 0$ or $\lambda^0 > 0$. Since we have assumed that the running cost function $\ell(x)$ is bounded, our results are not immediately comparable with those in robust control in which $\ell(x)$ is quadratic. In order to make this comparison, some "cut-off" argument is needed. See the comments at the end of the section. By (3.8), when $\lambda^0 = 0$ the inequality (5.13) becomes

$$\int_0^T [\ell(\xi_t^0) - \gamma^2|v_t|^2]dt + W^0(\xi_T^0) \leq W^0(x) \tag{5.15}$$

for every deterministic control function $v_.$ and $T < \infty$. Inequality (5.15) is called a *dissipation inequality* in robust control theory. Whether $\lambda^0 = 0$ or $\lambda^0 > 0$ depends on the parameter γ. We note that if $W^0(x)$ has a minimum at some point x^*, then for the initial data $\xi_0^0 = x^*$ the dissipation inequality (5.15) implies that for each $v_.$ and T

$$\int_0^T \ell(\xi_t^0)dt \leq \int_0^T \gamma^2|v_t|^2dt. \tag{5.16}$$

For quadratic $\ell(x)$, (5.16) is a similar condition in robust control theory corresponding to bounding an L^2-operator norm by γ. See [2]. Typically, in robust control theory x^* is a point about which the system $dx_t^0 = g(x_t^0)dt$ without disturbances is globally asymptotically stable. Let us assume that $x^* = 0$ and that $\ell(0) = 0$, $\ell(x) > 0$ for $x \neq 0$. We recall that the PDE (5.3) can be rewritten as

$$\lambda^0 = g(x) \cdot \nabla W^0 + \frac{1}{4\gamma^2}|\nabla W^0|^2 + \ell(x). \tag{5.4}$$

Since W^0 is a Lipschitz continuous viscosity solution to (5.4), it satisfies (5.4) in the usual sense for almost all x (in fact, at each x where W^0 is differentiable). This implies the following necessary condition for $\lambda^0 = 0$:

$$\ell(x) \leq \gamma^2|g(x)|^2 \text{ for all } x \in I\!R^n. \tag{5.17}$$

If (5.4) holds in the usual sense at x, then (5.17) is immediate from the inequality

$$\left|\gamma g(x) + \frac{1}{2\gamma}\nabla W^0(x)\right|^2 \geq 0.$$

By continuity of g and ℓ, (5.17) then holds for all x.

In dimension $n = 1$, (5.17) implies that (5.4) has a classical solution $W^0 \in C^1(I\!R^1)$ with $\lambda^0 = 0$. In fact,

$$W_x^0(x) = -2\gamma^2 g(x) - 2\ (\operatorname{sgn} x)(\gamma^4 g^2(x) - \gamma^2\ell(x))^{\frac{1}{2}}. \tag{5.18}$$

Since $g(0) = \ell(0) = 0$, we have $W_x^0(0) = 0$. By (3.1), $\ell(x) << g^2(x)$ as $|x| \to \infty$ and hence $W_x^0(x)$ is bounded. In fact, $W_x^0(x) \to 0$ as $|x| \to \infty$. For $x > 0$

$$W_x^0(x) = 2\gamma^2|g(x)|\left[1 - \left(1 - \frac{\ell(x)}{\gamma^2 g^2(x)}\right)^{\frac{1}{2}}\right]$$

$$\sim \frac{\ell(x)}{|g(x)|} \quad \text{as} \quad |x| \to \infty.$$

Since $\ell(x)$ is bounded and $|g(x)|x^{-1}$ is bounded below, the last term tends to 0. Similarly, $W_x(x) \to 0$ as $x \to -\infty$.

If $\lambda^0 = 0$, then (5.4) is just the steady-state form of the time-dependent PDE (5.7) satisfied by the value function $V^0(T,x)$. Corollary 5.2 says in case $\lambda^0 = 0$ that $T^{-1}V^0(T,x)$ tends to 0 uniformly on compact sets as $T \to \infty$. Under suitable further assumptions it should be true that $V^0(T,x)$ tends to a Lipschitz continuous viscosity solution $W^0(x)$ as $T \to \infty$. However, we have not proved any result of that kind.

If $\lambda^0 > 0$, then according to Theorem 5.3 there is a control $v_.$ for which average cost per unit time accumulates at a rate arbitrarily close to λ^0. The dissipation property is violated by the appearance of the term $\lambda^0 T$ on the right side of (5.13). If $W^0(x)$ has a minimum at the reference point $x^* = 0$ and $\lambda^0 > 0$, then W^0 cannot be differentiable at 0. Indeed, if W^0 is differentiable at 0, then $\nabla W^0(0) = 0$. Since $\ell(0) = 0$, this implies $\lambda^0 = 0$ by (5.4).

Example 5.5. Let $n = 1, g(-x) = -g(x), \ell(-x) = \ell(x)$ with $\ell(0) = 0$, $\ell(x) > 0$ for $x \neq 0$. Suppose that, for $x > 0$, $\ell(x) - \gamma^2 g^2(x)$ has a positive maximum at a unique point x_1. Let

$$\lambda^0 = \max_x[\ell(x) - \gamma^2 g^2(x)] = \ell(x_1) - \gamma^2 g^2(x_1). \tag{5.19}$$

We define $W^0(x)$ by

$$W^0(0) = 0, \qquad W^0(-x) = W^0(x),$$

and for $x > 0$

$$\begin{aligned} W^0_x(x) &= -2\gamma^2 g(x) + 2\alpha(x)(\gamma^4 g^2(x) - \gamma^2\tilde{\ell}(x))^{\frac{1}{2}}, \\ \alpha(x) &= \begin{cases} 1 & \text{if } 0 < x \le x_1 \\ -1 & \text{if } x > x_1, \end{cases} \\ \tilde{\ell}(x) &= \ell(x) - \lambda^0. \end{aligned} \tag{5.20}$$

Then W^0_x is continuous except at $x = 0$ and satisfies (5.4). At $x = 0, W^0_x$ jumps, with

$$W^0_x(0^-) = -2\gamma(\lambda^0)^{\frac{1}{2}}, \qquad W^0_x(0^+) = 2\gamma(\lambda^0)^{\frac{1}{2}}.$$

Let us rewrite (5.4) as

$$\lambda^0 + H(x, \nabla W^0) = 0,$$

$$H(x,p) = -g(x)\cdot p - \frac{1}{4\gamma^2}|p|^2 - \ell(x).$$

Since $H(x,p)$ is concave in p, in dimension 1 the derivative W^0_x of a viscosity solution $W^0(x)$ can have positive jumps but not negative jumps. See [15, §2.8]. Thus, $W^0(x)$ defined above is a viscosity solution. As in the discussion following (5.18), $W^0_x(x)$ is bounded. In fact, since $\alpha(x) = -\text{sgn } x$ for $|x| > x_1$, $W^0_x(x) \to 0$ as $|x| \to \infty$.

Remark 5.6. As already noted in Remark 4.4, the assumption that $\ell(x)$ is bounded in (3.1a) can be omitted. In dimension $n = 1$, if $\ell_x(x)$ is bounded but not $\ell(x)$, then in (5.18) one cannot expect that $W^0_x(x) \to 0$ as $|x| \to \infty$. If $|g(x)|^{-1}\ell(x)$ tends to limits as $x \to \pm\infty$, then the elementary argument below (5.18) shows that $W^0_x(x)$ tends to the same limits.

In robust control theory, typically $\ell(x)$ is taken to be quadratic. Of course, a quadratic ℓ does not have $\nabla\ell(x)$ bounded; hence a different result must be expected. This is seen from the following simple example.

Example 5.7. Let $n = 1$, $g(x) = -cx, \ell(x) = Kx^2$. The value function $V^0(T,x)$ for the finite-time-horizon problem (with $\mathcal{W}_M^0$ replaced by $\mathcal{W}^0$ in (5.9)) is quadratic in x. If $c^2\gamma^2 > K$, then $V^0(T,x)$ tends to a limit $W^0(x) = kx^2$ as $T \to \infty$. Moreover, $W^0(x)$ satisfies (5.4) with $\lambda^0 = 0$. On the other hand, if $c^2\gamma^2 < K$, then $V^0(T,x)$ explodes at some time $T^* < \infty$, and there is no infinite-horizon, average cost per unit time control problem.

One can reintroduce average cost per unit time control problems by a "cutoff" procedure, in which quadratic $\ell(x)$ is replaced by bounded functions which agree with $\ell(x)$ on large bounded sets. Of course, the optimal average cost per unit time must tend to infinity as the cutoff is relaxed. We illustrate this idea in Example 5.7. For $R > 0$ let

$$\ell_R(x) = \begin{cases} Kx^2 & \text{if } |x| \le R, \\ KR^2 & \text{if } |x| > R. \end{cases}$$

When $g(x) = -cx$,

$$\ell_R(x) - \gamma^2 g^2(x) = \begin{cases} (K - c^2\gamma^2)x^2 & \text{if } |x| \le R, \\ KR^2 - c^2\gamma^2 x^2 & \text{if } |x| > R. \end{cases}$$

When $c^2\gamma^2 < K$, this is maximum when $|x| = R$. By (5.19), $\lambda^0 = \lambda_R^0 = (K - c^2\gamma^2)R^2$, which tends to infinity as $R \to \infty$. On the other hand, if $c^2\gamma^2 > K$ then $\lambda_R^0 = 0$ for every $R > 0$.

This example suggests the following kind of result. If $\ell(x)$ is unbounded, with $\ell(x)$ growing quadratically as $|x| \to \infty$, then one should look for a critical value γ_1 of the parameter γ with the following properties. If $\gamma > \gamma_1$, then a dissipation inequality like (5.15) should hold with $W^0(x)$ a quadratically growing viscosity solution to (5.15) for $\lambda^0 = 0$. On the other hand, if $\gamma < \gamma_1$ then the optimal average cost per unit time λ_R^0 for cutoff problems should tend to infinity as $R \to \infty$.

6. Infinite-horizon risk-sensitive control. In the remainder of the paper, we return to consider risk-sensitive control problems of the kind mentioned in the introduction. The format will be similar to that of §§3–5. The present section is concerned with a rather general problem formulation, without proofs. The method of dynamic programming leads, in a formal way, to a nonlinear analogue of the eigenvalue problem in §2. By making a logarithmic transformation we arrive (again formally) at an Isaacs equation for a stochastic differential game, with average cost per unit time payoff criterion. In §7, we put these ideas on a rigorous basis for a particular class of risk-sensitive control problems governed by SDEs. Then in §8 we consider deterministic limits.

We now consider controlled Markov processes x_t with state space Σ, with the dynamics of x_t affected by a control process u_t. We require that $u_t \in U$, where U is a given "control space." Proceeding formally, for each constant control $u \in U$, the process x_t should be Markov with generator G^u. More generally, given a stationary Markov control policy $\underline{u}(x)$ belonging to some admissible class, x_t is a Markov process with generator $G^{\underline{u}}$. For further discussion of controlled Markov processes in this general setting, see [15, Chap. 3].

Let $L(x,u)$ be a running cost criterion, and $\epsilon > 0$ a parameter (in this section and §7, ϵ is fixed.) Given a stationary Markov control policy $\underline{u}$, we may refer to the formal results in §2 with $G = G^{\underline{u}}$ and

$$\ell^{\underline{u}}(x) = L(x, \underline{u}(x)).$$

From (2.1) we anticipate that (under suitable hypotheses) the limit

$$\lambda^{\underline{u}} = \epsilon \lim_{T\to\infty} \frac{1}{T} \log E_x \exp\left[\epsilon^{-1} \int_0^T \ell^{\underline{u}}(x_t)dt\right] \tag{6.1}$$

exists and that $\epsilon^{-1}\lambda^{\underline{u}}$ can be interpreted as the dominant eigenvalue of the linear operator $G^{\underline{u}} + \epsilon^{-1}\ell^{\underline{u}}$.

The goal is to find a stationary Markov control policy $\underline{u}^*$ which minimizes $\lambda^{\underline{u}^*}$. Again proceeding formally, we let

$$\Lambda = \inf_{\underline{u}} \lambda^{\underline{u}}. \tag{6.2}$$

Then dynamic programming and the separation of variables technique in §2 lead formally to a nonlinear eigenvalue problem for Λ and a positive eigenfunction $\Psi(x)$:

$$\epsilon^{-1}\Lambda\Psi(x) = \min_{u\in U}[G^u\Psi(x) + \epsilon^{-1}L(x,u)\Psi(x)]. \tag{6.3}$$

This is done via the heuristic $\Phi(T,x) \sim \exp(\epsilon^{-1}\Lambda T)\Psi(x)$, where $\Phi(T,x)$ is the value function for a finite-time-horizon control problem of minimizing the exponential cost criterion

$$E_x \exp\left[\epsilon^{-1} \int_0^T L(x_t, u_t,)dt\right].$$

The dynamic programming equation for the finite-time problem is

$$\Phi_T = \min_{u\in U}[G^u\Phi + \epsilon^{-1}L\Phi]. \tag{6.4}$$

As in §2, we make the logarithmic transformation $W = \epsilon \log \Psi$, and let

$$\mathcal{H}^u(W) = \epsilon \exp(-\epsilon^{-1}W)G^u[\exp(\epsilon^{-1}W)]. \tag{6.5}$$

Then (6.3) becomes

$$\Lambda = \min_{u\in U}[\mathcal{H}^u(W) + \ell]. \tag{6.6}$$

By introducing another (maximizing) control v in the same way as in §2, (6.6) becomes the Isaacs equation for a stochastic differential game. The game payoff has the average cost per unit time form

$$J = \limsup_{T\to\infty} \frac{1}{T} E_x \int_0^T K(\xi_t, u_t, v_t)dt, \tag{6.7}$$

with suitably defined running cost function $K(x,u,v)$. Here ξ_t denotes the state of the game at time t.

Let us describe explicitly the game and the form of the Isaacs equation (6.6) only for controlled nondegenerate Markov diffusions in $\mathbb{R}^n$. Proceeding in a way similar to §2, we suppose that x_t is governed by an SDE

$$dx_t = f(x_t, u_t)dt + \epsilon^{\frac{1}{2}}\sigma(x_t, u_t)db_t, \tag{6.8}$$

with initial data $x_0 = x$. Here $b_.$ is an n-dimensional Brownian motion, and the matrix $a = \sigma\sigma^{'}$ is positive definite for all $x \in I\!R^n$, $u \in U$. (In the rigorous treatment in §7 we will take U compact, $a = (2\gamma^2)^{-1}I$ with assumptions on $f(x,u), L(x,u)$ corresponding to those made for $g(x), \ell(x)$ in §3.)

For the dynamics (6.8), the family G^u of generators becomes (see [15, p. 173]) for $\psi \in C^2(I\!R^n)$

$$G^u\psi(x) = \frac{\epsilon}{2}\sum_{i,j=1}^{n} a_{ij}(x,u)\psi_{x_ix_j} + f(x,u)\cdot\nabla\psi(x). \tag{6.9}$$

By arguing as in §2 (see (2.7)–(2.9)) we obtain for $W = \epsilon \log \psi$ the PDE

$$\Lambda = \min_{u\in U}\max_{v\in I\!R^n}\left[\frac{\epsilon}{2}\sum_{i,j=1}^{n} a_{ij}(x,u)W_{x_ix_j}(x) \right. \tag{6.10}$$
$$\left. +(f(x,u)+v)\cdot\nabla W(x) + K(x,u,v)\right],$$

$$K(x,u,v) = L(x,u) - \frac{a^{-1}(x,u)}{2}v\cdot v. \tag{6.11}$$

Equation (6.10) is the Isaacs equation for a stochastic differential game with state ξ_t governed by the dynamics

$$d\xi_t = [f(\xi_t,u_t)+v_t]dt + \epsilon^{\frac{1}{2}}\sigma(\xi_t,u_t)dw_t, \tag{6.12}$$

with w_t some Brownian motion and with payoff J as in (6.7). Note that (6.12) is just a "controlled" version of (2.10). Formally, Λ corresponds to the upper game value, since min max and max min cannot in general be exchanged in (6.10). However, if $a = a(x)$, then the minimizing control u and the maximizing control v appear separately in (6.10). In that case, the Isaacs condition min max = max min holds. In particular, this is true when a is constant, as in §7.

7. Ergodic stochastic differential game. In this section we will put the formalism in §6 on a rigorous basis, in a special case in which the matrix σ in (6.10) is constant. In fact, we assume as in §3 that $a = \sigma\sigma^{'} = (2\gamma^2)^{-1}I$, with assumptions on $f(x,u), L(x,u)$ similar to the assumptions (3.1) made earlier. Then in §8 we will consider the deterministic differential game obtained in the limit $\epsilon \to 0$ and its relation to some questions in robust nonlinear control theory. The results of these two sections are analogous to those of §§3–5, and we will simply adapt those proofs whenever possible.

We make the following assumptions.

(7.1) The control space U is compact.

The functions L, f are continuous on $I\!R^n \times U$. Moreover, $L(\cdot,u), f(\cdot,u)$ are of class $C^1(I\!R^n)$ for all $u \in U$, and

(7.2)
(a) $L, \nabla_x L$ are bounded and $L \geq 0$;

(b) f_x is bounded;

(c) There exists $c > 0$ such that for all $x, y \in I\!R^n, u \in U$, $(x-y)\cdot[f(x,u)-f(y,u)] \leq -c|x-y|^2$.

As in the discussion following (3.1) these assumptions are considerably stronger than needed to prove the results that follow. However, they make the proofs much less technical. In particular, the assumption that L is bounded can be omitted, using arguments similar to Remark 4.4. The assumption that f, L are of class $C^1(\mathbb{R}^n)$ can be weakened to assume Lipschitz conditions for $f(x,\cdot), L(x,\cdot)$ which are uniform in u.

Equation (6.3) for Λ and $\Psi(x)$ now has the form

$$\begin{aligned}\epsilon^{-1}\Lambda\Psi(x) &= \frac{\epsilon}{4\gamma^2}\Delta\Psi(x) \\ &+ \min_{u\in U}[f(x,u)\cdot\nabla\Psi(x) + \epsilon^{-1}L(x,u)\Psi(x)].\end{aligned} \tag{7.3}$$

Equivalently, equation (6.6) for Λ and $W(x)$ now takes the form

$$\begin{aligned}\Lambda &= \frac{\epsilon}{4\gamma^2}\Delta W + \frac{1}{4\gamma^2}|\nabla W(x)|^2 \\ &+ \min_{u\in U}[f(x,u)\cdot\nabla W(x) + L(x,u)].\end{aligned} \tag{7.4}$$

This is the same as (6.10) with $a = (2\gamma^2)^{-1}I$ and

$$K(x,u,v) = L(x,u) - \gamma^2|v|^2. \tag{7.5}$$

THEOREM 7.1. *There exist $\Lambda \in \mathbb{R}$, $W \in C^2(\mathbb{R}^n)$ such that* (7.4) *holds. Moreover, there exists B (not depending on ϵ) such that $|\nabla W| \le B$ and $0 \le \Lambda \le B$.*

We defer the proof of Theorem 7.1 to later in this section. The proof will give $B = c^{-1}\|\nabla_x L\|$, with constant c as in assumption (7.2c). Instead, we will first demonstrate that Λ has an interpretation as the minimum cost for the risk-sensitive control problem introduced in §6. We recall from §3 the definition of reference probability system. Given any reference probability system $\nu = (\Omega, \{\mathcal{F}_t\}, P, b_\cdot)$, let $\mathcal{U}_\nu$ be the set of $\mathcal{F}_t$-progressively measurable processes with values in U.

We consider the dynamics given by (1.1):

$$\begin{aligned} dx_t &= f(x_t,u_t)dt + \left(\frac{\epsilon}{2\gamma^2}\right)^{\frac{1}{2}} db_t, \\ x_0 &= x, \end{aligned} \tag{7.6}$$

where b_t is the Brownian motion of the reference probability system ν and $u_\cdot \in \mathcal{U}_\nu$. This system has a strong solution [28].

A candidate for an optimal control policy is defined as follows. By a measurable selection theorem [11, Appendix B], there exists a Borel measurable U-valued function $\underline{u}^*$ on $\mathbb{R}^n$ such that

$$\underline{u}^*(x) \in \arg\min_{u\in U}[f(x,u)\cdot\nabla W(x) + L(x,u)] \tag{7.7}$$

for almost all $x \in \mathbb{R}^n$. By a result of Veretennikov [38] as extended in [30], the SDE

$$\begin{aligned} dx_t^* &= f(x_t^*, \underline{u}^*(x_t^*))dt + \left(\frac{\epsilon}{2\gamma^2}\right)^{\frac{1}{2}} db_t, \\ x_0^* &= x \end{aligned} \tag{7.8}$$

has a strong solution x_t^*.

THEOREM 7.2. *Let Λ, W be as in Theorem 7.1, and let ν be any reference probability system. Then*

(a) *For every $u_. \in \mathcal{U}_\nu$,*

$$\Lambda \leq \epsilon \liminf_{T\to\infty} \frac{1}{T} \log E_x \exp\left[\epsilon^{-1}\int_0^T L(x_t,u_t)dt\right].$$

(b) *Let $u_t^* = \underline{u}^*(x_t^*)$, with $\underline{u}^*, x_t^*$ as in (7.7), (7.8). Then*

$$\Lambda = \epsilon \lim_{T\to\infty} \frac{1}{T} \log E_x \exp\left[\epsilon^{-1}\int_0^T L(x_t^*,u_t^*)dt\right].$$

Proof. The proof of (a) is nearly identical to the proof of Theorem 3.5. However, in place of (3.13) we have

$$\begin{aligned} W(x_T) - W(x) \geq & \int_0^T [\Lambda - L(x_t,u_t) + \gamma^2|v_t|^2]dt \\ & + \sqrt{2\gamma^2\epsilon}\int_0^T v_t \cdot db_t^0. \end{aligned} \tag{7.9}$$

Thus, we have inequality rather than equality due to the minimization in (7.4). The inequality passes transparently through the remainder of the proof.

For part (b), equality holds in (7.9) with probability 1, and this gives the desired equality exactly as in the proof of Theorem 3.5. □

Remark 7.3. Given any Borel measurable, U-valued Markov contol policy $\underline{u}$, let $u_t = \underline{u}(x_t)$ where x_t is the solution to the SDE (7.8) with $\underline{u}^*$ replaced by $\underline{u}$ and with initial data $x_0 = x$. If the limit $\lambda^{\underline{u}}$ in (6.1) exists, then by Theorem 7.2(a), $\Lambda \leq \lambda^{\underline{u}}$. Without knowing that the limit exists in (6.1) Theorem 7.2(a) provides a slightly weaker result in which limit is replaced by lim inf. Theorem 7.2(b) asserts that $\lambda^{\underline{u}^*}$ exists and that $\lambda^{\underline{u}^*} = \Lambda$. This justifies calling $\underline{u}^*$ an optimal Markov control policy.

We now turn to the proof of Theorem 7.1. The proof will be structurally similar to the proof of Theorem 3.3 given in §4. However, the addition of the minimizing controller $u_.$ implies that we must now work with differential games which have two controllers (or players), $u_.$ and $v_.$. In contrast, in §4 we had a control problem with one (maximizing) controller.

In parallel with §4, we start with the finite-time-horizon, discounted-cost differential game with dynamics

$$\begin{aligned} & d\xi_t = [f(\xi_t,u_t) + v_t]dt + \left(\frac{\epsilon}{2\gamma^2}\right)^{\frac{1}{2}} dw_t, \\ & \xi_0 = x \end{aligned} \tag{7.10}$$

and payoff

$$J_\rho(T,x;u_.,v_.) = E_x\int_0^T e^{-\rho t}K(\xi_t,u_t,v_t)dt \tag{7.11}$$

where K is given by (7.5). We use the Elliott–Kalton definition of value of a differential game [8] as extended to stochastic differential games by Fleming and Souganidis [16]. We summarize this definition. Given a reference probability system $\mu = (\Omega, \{\mathcal{F}_t\}, P, w_.)$ let $\mathcal{U}$ denote the set of all U-valued, $\mathcal{F}_t$-progressively measurable processes $u_.$. As in §4, let $\mathcal{W}_M$ denote the set of all $I\!R^n$-valued $\mathcal{F}_t$-progressively measurable process $v_.$ such that $|v_t| \leq M$. The processes u_t, v_t are defined for all $t \geq 0$, although in the present discussion only $0 \leq t \leq T$ is relevant. Of course, $\mathcal{U} = \mathcal{U}^\mu$ and $\mathcal{W}_M = \mathcal{W}_M^\mu$ depend on μ. A strategy for the minimizing controller is a mapping $\beta : \mathcal{W}_M \to \mathcal{U}$ with the following property. For each $t > 0, v_r = \hat{v}_r$ for almost all $r \in [0,t]$ almost surely implies $\beta[v_.]_r = \beta[\hat{v}_.]_r$ for almost all $r \in [0,t]$ almost surely. This set of strategies is denoted by Δ_M. Similarly, a strategy for the maximizing controller is a mapping $\alpha : \mathcal{U} \to \mathcal{W}_M$ with the corresponding property. This set of strategies is denoted by Θ_M. The lower value is defined to be

$$V_\rho^\ell(T,x) = \inf_{\Delta_M} \sup_{\mathcal{W}_M} J_\rho(T,x;\beta[v_.],u_.), \tag{7.12}$$

and the upper value is

$$V_\rho^u(T,x) = \sup_{\Theta_M} \inf_{\mathcal{U}} J_\rho(T,x;u_.,\alpha[u_.]). \tag{7.13}$$

The Isaacs PDE for this differential game is

$$\begin{aligned} \rho V + V_T = {} & \frac{\epsilon}{4\gamma^2}\Delta_x V + \min_{u\in U}[f(x,u)\cdot\nabla_x V + L(x,u)] \\ & + \max_{|v|\leq M}[v\cdot\nabla_x V - \gamma^2|v|^2] \end{aligned} \tag{7.14}$$

with the initial data

$$V(0,x) = 0. \tag{7.15}$$

Results about semilinear parabolic PDEs (see Appendix B) imply that (7.14)–(7.15) has a classical solution $\bar{V}(T,x)$, with all partial derivatives $\bar{V}_T, \bar{V}_{x_i}, \bar{V}_{x_i x_j}$, $i,j = 1,\ldots,n$, continuous. Moreover, $\nabla_x \bar{V}$ is bounded. In order to obtain a priori estimates which do not depend on ϵ, ρ, or T for small ρ and large T, similar to (4.4) and (4.5), we first show that $\bar{V}(T,x)$ equals the upper value $V_\rho^u(T,x)$.

LEMMA 7.4. $\bar{V} = V_\rho^u$.

Proof. We show first that $\bar{V} \leq V_\rho^u$ and then the opposite inequality. We consider the control policy $h(t,x)$ defined by

$$h(t,x) = \arg\max_{|v|\leq M}[v\cdot\nabla_x\bar{V}(T-t,x) - \gamma^2|v|^2]. \tag{7.16}$$

Then h is continuous on $[0,T]\times I\!R^n$, and $h(t,\cdot)$ satisfies a local Lipschitz condition uniformly for $0 \leq t \leq T$. See [13, p. 170]. For each $u_. \in U$ define $\alpha^*[u_.]$ by

$$\begin{aligned} d\xi_t^* &= [f(\xi_t^*,u_t) + h(t,\xi_t^*)]dt + \left(\frac{\epsilon}{2\gamma^2}\right)^{\frac{1}{2}} dw_t, \\ \xi_0^* &= x, \end{aligned} \tag{7.17}$$

$$\alpha^*[u_.]_t = h(t,\xi_t^*). \tag{7.18}$$

Since h is continuous, bounded, and locally Lipschitz in x, the strong solution to (7.17) exists. Then (7.14) and (7.16) imply, for any $u_. \in \mathcal{U}$,

$$0 \le -\bar{V}_T(T-t,\xi_t^*) - \rho\bar{V}(T-t,\xi_t^*) + \frac{\epsilon}{4\gamma^2}\Delta_x\bar{V}(T-t,\xi_t^*)$$

$$+f(\xi_t^*,u_t)\cdot\nabla_x\bar{V}(T-t,\xi_t^*) + L(\xi_t^*,u_t)$$

$$+h(t,\xi_t^*)\cdot\nabla_x\bar{V}(T-t,\xi_t^*) - \gamma^2|h(t,\xi_t^*)|^2.$$

Since $\nabla_x\bar{V}$ is bounded, $\bar{V}(T,x)$ grows at most linearly as $|x|\to\infty$. By using the Feynman–Kac formula and the fact that $\bar{V}(0,x)=0$, we conclude that

$$\bar{V}(T,x) \le \inf_{\mathcal{U}} J_\rho(T,x;u_.,\alpha^*[u_.]).$$

Therefore, $\bar{V}(T,x) \le V_\rho^u(T,x)$.

In order to prove the opposite inequality $\bar{V}(T,x) \ge V_\rho^u(T,x)$ it suffices to show that, given any strategy $\alpha \in \Theta_M$ and $\delta > 0$, there exists $u_. \in \mathcal{U}$ such that

$$J_\rho(T,x;u_.,\alpha[u_.]) \le \bar{V}(T,x) + \delta. \tag{7.19}$$

We will define u_t to be piecewise constant in time, via a discrete time Markov control policy $\vec{u}$ in a way similar to [15, p 183]. Let π be a partition of [0,T] into intervals $[t_{j-1},t_j], j = 1,\dots,j_0$ with $0 = t_0 < t_1 < \cdots < t_{j_0} = T$. A discrete-time Markov control policy is

$$\vec{u} = (\underline{u}_1,\dots,\underline{u}_{j_0}),$$

where $\underline{u}_j$ is a Borel measurable function from $\mathbb{R}^n$ into U. Given initial data $\xi_0 = x$, a strategy α and a discrete time Markov control policy $\vec{u}$, we define u_t inductively as follows. For $0 \le t < t_1$, $u_t = \underline{u}_1(x)$. Then (7.10) is solved on $[0,t_1]$ with $v_t = \alpha[u_.]_t$ to obtain ξ_t. Proceeding by induction, if u_t and the corresponding solution ξ_t have been defined on $[0,t_j]$ we let

$$u_t = \underline{u}_{j+1}(\xi_{t_j}) \text{ for } t_j \le t < t_{j+1}.$$

We then extend ξ_t to $[t_j,t_{j+1}]$ as the solution to (7.10) on $[0,t_{j+1}]$ with the initial data $\xi_0 = x$. By uniqueness of strong solutions to (7.10), this is consistent with the definition of ξ_t on $[0,t_j]$. We will choose $\vec{u}$ in the following way. In addition to a partition π of $[0,T]$ we will partition $\mathbb{R}^n$ into disjoint Borel sets $A_0,A_1,\dots,A_{k_0}$. We will also choose $u_{jk} \in U$ and let

$$u_j(y) = u_{jk} \text{ if } y \in A_k.$$

These partitions are chosen as follows. Since $\bar{V}(T,x)$ is a solution to (7.14),

$$-\rho\bar{V} - \bar{V}_T + \frac{\epsilon}{4\gamma^2}\nabla_x\bar{V} + \min_{u\in U}[f(x,u)\cdot\nabla_x\bar{V} + L(x,u)] \tag{7.20}$$

$$+v\cdot\nabla_x\bar{V} - \gamma^2|v|^2 \le 0$$

for all v such that $|v| \leq M$. As in the proof of [15, Lem. 4.7.1], we choose R sufficiently large and $A_1, \ldots, A_{k_0}$ a partition of the ball $\{|y| \leq R\}$ into disjoint Borel sets of sufficiently small diameter. Let $A_0 = \{|y| > R\}$. The points $u_{jk} \in U, j = 1, \ldots, j_0, k = 1, \ldots, k_0$ are chosen such that

$$f(y, u_{jk}) \cdot \nabla_x \bar{V}(T-t, y) + L(y, u_{jk}) \tag{7.21}$$

$$< \min_{u \in U}[f(y,u) \cdot \nabla_x \bar{V}(T-t,y) + L(y,u)] + \eta$$

for all $t \in [t_{j-1}, t_j], y \in A_k$, where $\eta > 0$. We choose $u_{j0} \in U$ arbitrarily. By (7.20) and (7.21)

$$-\rho \bar{V}(T-t, \xi_{t_j}) - \bar{V}_T(T-t, \xi_{t_j}) + \frac{\epsilon}{4\gamma^2} \Delta_x \bar{V}(T-t, \xi_{t_j})$$

$$+(f(\xi_{tj}, u_t) + v_t) \cdot \nabla \bar{V}(T-t, \xi_{t_j}) + K(\xi_{t_j}, u_t, v_t) < \eta,$$

where $v_t = \alpha[u_.]_t$ if $t_j \leq t \leq t_{j+1}$ and $|\xi_{t_j}| \leq R$. By an argument in [15, p. 184], if η and the lengths of the intervals $[t_j, t_{j+1}]$ are chosen small enough,

$$\bar{V}(T,x) \geq E_x \int_0^{\tau_R} e^{-\rho t} K(\xi_t, u_t, \alpha[u_.]_t) dt \tag{7.22}$$

$$+E_x e^{-\rho \tau_R} \bar{V}(T - \tau_R, \xi_{\tau_R}) - \frac{\delta}{2},$$

where τ_R is the smaller of T and the exit time of ξ_t from the ball $\{|y| \leq R\}$. Since U is compact and $|v_t| \leq M$, we have from Cauchy–Schwartz, together with

$$|\bar{V}(s,y)| \leq C(1 + |y|), \; V(0,y) = 0,$$

that for suitable C,

$$J_\rho(T, x; u_., \alpha[u_.]) \leq \bar{V}(T,x) + \frac{\delta}{2}$$

$$+T(\|L\| + \gamma^2 M^2) P_x(\tau_R < T) \tag{7.23}$$

$$+C[E_x(1 + |\xi_{\tau_R}|)^2]^{\frac{1}{2}} [P_x(\tau_R < T)]^{\frac{1}{2}}.$$

Now $\tau_R < T$ if and only if $\|\xi_.\| \geq R$, where $\| \; \|$ is the sup norm on $[0,T]$, Moreover, $|\xi_{\tau_R}| \leq \|\xi_.\|$. By standard estimates for SDEs (see, for example, [28, §1.5] or [15, Appendix D]), the sum of the last two terms in (7.23) is less than $\delta/2$, if R is large enough. (The choice of R does not depend on $u_.$ or α.) Then (7.19) holds, which completes the proof. □

The assumptions (7.2) are essentially the same as assumptions (3.1), with the exception that the bounds are now uniform in the new control variable $u_.$. Thus by the same methods as employed in §4 we obtain the same bounds on J_ρ and V_ρ^u. In particular, we have the following lemma.

LEMMA 7.5. *For every $T > 0, x, y \in I\!R^n, u_. \in \mathcal{U},\ v_. \in \mathcal{W}_M$ we have*

$$\begin{aligned}
&\text{(a)}\quad \rho J_\rho(T,x;u_.,v_.) \le \|L\|;\\
&\text{(b)}\quad |J_\rho(T,x;u_.v_.) - J_\rho(T,y;u_.,v_.)| \le \frac{\|\nabla_x L\|}{\rho+c};\\
&\text{(c)}\quad 0 \le \rho V_\rho^u(T,x) \le \|L\|;\\
&\text{(d)}\quad |\nabla_x V_\rho^u| \le \frac{\|\nabla_x L\|}{\rho+c};\\
&\text{(e)}\quad |(V_\rho^u)_T| \le \rho^{-1}e^{-\rho T}(\|L\| + \gamma^2 M^2).
\end{aligned} \tag{7.24}$$

Let $M^* = (2\gamma^2 c)^{-1}\|\nabla_x L\|$, and note that by (7.14) and (7.24d)

$$\max_{v\in I\!R^n}[v\cdot\nabla V_\rho^u - \gamma^2|v|^2] = \max_{|v|\le M}[v\cdot\nabla V_\rho^u - \gamma^2|v|^2].$$

Therefore, the solution $\bar{V} = V_\rho^u$ of (7.14)–(7.15) does not depend on M, if $M \ge M^*$. Moreover, in the definition (7.13) of upper value, Θ_M can be replaced by the set Θ of all bounded strategies α.

Since V_ρ^u is a classical solution to (7.14), its first-order partial derivative in T and first- and second-order partial derivatives in x are continuous. By (7.24e) the limit

$$W_\rho(x) = \lim_{T\to\infty} V_\rho^u(T,x)$$

exists. Furthermore, the same argument used for (4.9) shows that $W_\rho \in C^2(I\!R^n)$ satisfies

$$\begin{aligned}
\rho W_\rho\ (x) = {}& \frac{\epsilon}{4\gamma^2}\Delta W_\rho(x)\\
&+\min_{u\in U}[f(x,u)\cdot\nabla W_\rho(x) + L(x,u)]\\
&+\max_{v\in I\!R^n}[v\cdot\nabla W_\rho(x) - \gamma^2|v|^2].
\end{aligned} \tag{7.25}$$

In fact, $W_\rho(x)$ is the upper value of the infinite-horizon, discounted-cost game (with $T=\infty$ on the right side of (7.11)). Moreover, by (7.24c) and (7.24d)

$$\begin{aligned}
&\text{(a)}\quad 0 \le \rho W_\rho \le \|L\|,\\
&\text{(b)}\quad |\nabla W_\rho| \le \frac{\|\nabla_x L\|}{\rho+c}.
\end{aligned} \tag{7.26}$$

To complete the proof of Theorem 7.1, we proceed just as in §4. By Ascoli's theorem, there is a sequence $\rho_m \to 0$ as $m\to\infty$ such that $\rho_m W_{\rho_m}(x)$ tends to some number Λ and, for fixed x_0, $W_{\rho_m}(x) - W_{\rho_m}(x_0)$ tends to a limit $W(x)$ uniformly on compact sets. Moreover, $W \in C^2(I\!R^n)$ and Λ, W satisfy (7.4). Finally, (7.26) implies that $|\nabla W| \le B$ and $0 \le \Lambda \le B$ if

$$B = \max\{\|L\|, c^{-1}\|\nabla_x L\|\}. \qquad \square \tag{7.27}$$

8. Limiting deterministic differential game. Following the same procedure as in §5, we take deterministic limits as $\epsilon \to 0$. We indicate the dependence on ϵ of $\Lambda, W(x)$ in §7 by writing $\Lambda = \Lambda^\epsilon, W = W^\epsilon$. According to (7.26), the following inequalities corresponding to (5.1) hold:

$$0 \le \Lambda^\epsilon \le \|L\|, \qquad |\nabla W^\epsilon(x)| \le \frac{\|\nabla_x L\|}{c}. \tag{8.1}$$

We again use Ascoli's theorem to find a sequence $\epsilon_m \to 0$ such that as $m \to \infty$

$$\Lambda^{\epsilon_m} \to \Lambda^0, \qquad W^{\epsilon_m}(x) \to W^0(x) \tag{8.2}$$

uniformly on compact subsets of $I\!R^n$. Moreover, W^0 is a viscosity solution to the first-order PDE

$$\begin{aligned}\Lambda^0 &= \min_{u\in U}[f(x,u)\cdot\nabla W^0(x) + L(x,u)]\\ &\quad+\max_{v\in I\!R^n}[v\cdot\nabla W^0(x) - \gamma^2|v|^2].\end{aligned} \tag{8.3}$$

The analogue of Theorem 5.1 is therefore the following theorem.

THEOREM 8.1. *There exist $\Lambda^0 \ge 0$ and Lipschitz continuous W^0 such that W^0 is a viscosity solution of* (8.3).

The PDE (8.3) can alternatively be written as

$$\begin{aligned}\Lambda^0 &= \min_{u\in U}[f(x,u)\cdot\nabla W^0(x) + L(x,u)]\\ &\quad+\frac{1}{4\gamma^2}|\nabla W^0(x)|^2.\end{aligned} \tag{8.3$'$}$$

As noted in §5, the maximum over $v \in I\!R^n$ in (8.3) can be replaced by the maximum over $|v| \le M$, provided $M \ge (2\gamma^2)^{-1}B$ with B a Lipschitz constant for W^0.

Equation (8.3) is the Isaacs equation for a (deterministic) differential game, with average cost per unit time payoff. Before considering this game, let us first consider some corresponding finite-time-horizon games. There are two controllers (or players). The minimizing controller chooses a Lebesgue-measurable function $u_.$, with values in U. Let $\mathcal{U}^0$ denote the set of such $u_.$. The maximizing controller chooses $v_. \in \mathcal{W}_M^0$, with $\mathcal{W}_M^0$ the set of Lebesgue-measurable $v_.$ such that $|v_t| \le M$ as in §5. The state of the game at time t is ξ_t^0, and the state dynamics are

$$\frac{d\xi_t^0}{dt} = f(\xi_t^0, u_t) + v_t, \tag{8.4}$$

with $\xi_0^0 = x$. The running cost function is $K(x,u,v)$ (see (7.5)). The payoff for the finite-time-horizon game is

$$P = \int_0^T K(\xi_t^0, u_t, v_t)dt + Z(\xi_T^0), \tag{8.5}$$

where Z is a Lipschitz continuous terminal cost function. Later in the section we will consider two choices for Z, namely, $Z = W^0$ and $Z = 0$.

This differential game has a value in the Elliott–Kalton sense. The Elliott–Kalton value is defined in terms of strategies, as follows [8]. A strategy for the maximizing controller is a map α from $\mathcal{U}^0$ into $\mathcal{W}_M^0$ which is "progressive" in the following sense. For each $t > 0$ and $u_., \tilde{u}_. \in \mathcal{U}^0$, $u_r = \tilde{u}_r$ for almost all $r \in [0,t]$ implies that $\alpha[u_.]_r = \alpha[\tilde{u}_.]_r$ for almost all $r \in [0,t]$. This set of strategies is denoted by Δ_M^0. Similarly, a strategy for the minimizing controller is a progressive map $\beta : \mathcal{W}_M^0 \to \mathcal{U}^0$. Let Θ_M^0 denote this set of strategies. For a game with payoff P, let us denote the upper and lower values by $u -$ val P and $\ell -$ val P. Thus

$$(8.6)\qquad \begin{aligned} u - \text{ val } P &= \sup_{\Delta_M^0} \inf_{\mathcal{U}^0} P, \\ \ell - \text{ val } P &= \inf_{\Theta_M^0} \sup_{\mathcal{W}_M} P. \end{aligned}$$

If $u -$ val $P = \ell -$ val P, this is called the Elliott–Kalton value and will be denoted by val P. For the finite-time game (8.4)–(8.5) we also denote upper and lower values by W^u and W^ℓ.

The Isaacs PDE for the finite-time game formulated above is

$$(8.7)\qquad \begin{aligned} W_T = &\min_{u\in U}[f(x,u)\cdot \nabla_x W + L(x,u)) \\ &+ \max_{|v|\le M}[v\cdot \nabla_x W - \gamma^2|v|^2] \end{aligned}$$

with initial data $W(0,x) = Z(x)$. Considered as functions of (T,x), the upper and lower values W^u, W^ℓ are both Lipschitz continuous viscosity solutions to (8.7) with the initial data. This follows from a dynamic programming principle, as was proved by Evans and Souganidis [9]. Uniqueness of viscosity solutions (see [30]) then implies that the finite-time-horizon game has a value $W(T,x)$, where $W = W^\ell = W^u$ is the unique Lipschitz continuous viscosity solution to (8.7) with $W(0,x) = Z(x)$.

Several other definitions of value have been given for differential games; see, for example, [10], [18]. All "reasonable" definitions of value lead to the same (unique) viscosity solution of (8.7) with the initial data. Such results can be proved by a method of Souganidis [35].

THEOREM 8.2. *If W^0 is any Lipschitz continuous viscosity solution of* (8.3), *then for every $T < \infty$*

$$(8.8)\qquad W^0(x) = \text{ val } \left[\int_0^T K(\xi_t^0, u_t, v_t)dt + W^0(\xi_T^0)\right] - \Lambda^0 T.$$

Proof. We replace the running cost function K by $K - \Lambda^0$ and proceed as in the proof of Theorem 5.2. □

We next consider zero terminal cost, and let

$$(8.9)\qquad V^0(T,x) = \text{ val } \int_0^T K(\xi_t^0, u_t, v_t)dt.$$

By the same proof as for Corollary 5.3, we get Corollary 8.3.

COROLLARY 8.3.

$$(8.10)\qquad \Lambda^0 = \lim_{T\to\infty} \frac{1}{T} V^0(T,x)$$

uniformly for x in any compact set.

Let us next show that Λ^0 is the value of the infinite-time-horizon differential game with payoff

$$J^0 = \limsup_{T\to\infty} \frac{1}{T} \int_0^T K(\xi_t^0, u_t, v_t)dt. \tag{8.11}$$

THEOREM 8.4 *For every initial state $x \in I\!R^n, \Lambda^0 =$ val J^0.*

Proof. Let us show that the lower value $\ell -$ val J^0 equals Λ^0. The same technique shows that $u -$ val $J^0 = \Lambda^0$. Given a strategy β for the minimizing controller and an initial state $x = \xi_0^0$, we write in (8.11) $J^0 = J^0(x; \beta, v_.)$, where the control $u_.$ is chosen by $u_t = \beta[v_.]_t$ for all $t \geq 0$. We begin by showing that, given any initial state x, strategy β, and $\delta > 0$, there exists $v_. \in \mathcal{W}_M^0$ such that

$$J^0(x; \beta, v_.) \geq \Lambda^0 - \delta. \tag{8.12}$$

To do this, we proceed in a way similar to the proof of Theorem 5.4. We choose R_1 and T_0 such that (5.14) holds for $|y| \leq R_1$ (where λ^0 is replaced by Λ^0 and V^0 is given by (8.9)). Since $V^0(T, y)$ is the value of the finite-time game with initial state y, given a strategy $\bar\beta$ we have

$$\sup_{v_.} \int_0^{T_0} K(\eta_t^0, \bar\beta[v_.]_t, v_t)dt \geq V^0(T_0, y), \tag{8.13}$$

where η_t^0 is the solution to (8.4) with $\eta_0^0 = y$ and $u_t = \bar\beta[v_.]_t$. We first take $y = x, \bar\beta = \beta$, and choose v_t on $[0, T_0)$ such that

$$\int_0^{T_0} K(\xi_t^0, \beta[v_.]_t, v_t)dt > V^0(T_0, x) - \frac{\delta T_0}{2}.$$

Then we proceed inductively to define v_t on $[NT_0, (N+1)T_0)$ as follows. Suppose that v_t has already been defined for $0 \leq t < NT_0$. Define the strategy β_N as follows. For each $z_. \in \mathcal{W}_M$,

$$\beta_N[z_.] = \beta[z_.^N], \text{ where}$$

$$z_t^N = \begin{cases} v_t & \text{if } 0 \leq t < NT_0, \\ z_{t-NT_0} & \text{if } NT_0 \leq t. \end{cases}$$

In (8.13) we take $y = \xi_{NT_0}^0, \bar\beta = \beta_N$, and $v_.^N$ such that

$$\int_0^{T_0} K(\eta_t^0, \beta_N[v_.^N]_t, v_t^N)dt > V^0(T_0, \xi_{NT_0}^0) - \frac{\delta T_0}{2}.$$

Let $v_t = v_{t+NT_0}^N$ for $NT_0 \leq t < (N+1)T_0$. Then

$$\int_{NT_0}^{(N+1)T_0} K(\xi_t^0, \beta[v_.]_t, v_t)dt > V^0(T_0, \xi_{NT_0}^0) - \frac{\delta T_0}{2}.$$

As in the proof of Theorem 5.5, we obtain (8.12). Thus

$$\ell - \text{ val } J^0 = \inf_\beta \sup_v J^0 \geq \Lambda^0. \tag{8.14}$$

It remains to prove the opposite inequality. Given an initial state x and $\delta > 0$, we again consider R_1, T_0 such that (5.14) holds for $|y| \leq R_1$. For each y with $|y| \leq R_1$ there exists a strategy β^y such that

$$\sup_{v.} \int_0^{T_0} K(\eta_t^0, \beta^y[v.]_t, v_t)dt \leq V^0(T, y) + \frac{\delta T_0}{2}, \tag{8.15}$$

where η_t^0 is the solution of (8.4) with $\eta_0^0 = y$ and $u_t = \beta^y[v.]_t$. We define a strategy β as follows. For $0 \leq t < T_0$, $\beta[v.]_t = \beta^x[v.]_t$. Proceeding inductively, suppose that $\beta[v.]_t$ is defined for $0 \leq t < NT_0$. In (8.14) let $y = y_N = \xi_{NT_0}^0$ and for $NT_0 \leq t < (N+1)T_0$ let

$$\beta[v.]_t = \beta^{y_N}[v_{N}.]_t,$$

$$v_{Nt} = v_t - NT_0, t \geq NT_0.$$

Then

$$\sup_{v.} \int_{NT_0}^{(N+1)T_0} K(\xi_t^0, \beta[v.]_t, v_t)dt \leq V^0(T_0, \xi_{NT_0}^0) + \frac{\delta T_0}{2}.$$

As before we obtain

$$\ell - \text{ val } J^0 = \inf_{\beta} \sup_{v.} J^0 \leq \Lambda^0. \qquad \square \tag{8.16}$$

Robust regulation of nonlinear systems (continued). At the end of §5, we related our results on risk-sensitive control to robust control approaches to disturbance attenuation. The results in §5 apply if a control policy $\underline{u}$ for the minimizing controller is given, and we take $g = g^{\underline{u}}, \ell = \ell^{\underline{u}}$ with

$$g^{\underline{u}}(x) = f(x, \underline{u}(x)), \ \ell^{\underline{u}}(x) = L(x, \underline{u}(x)).$$

If $g^{\underline{u}}$ and $\ell^{\underline{u}}$ satisfy (3.1), then according to Theorem 5.1 there is a pair $\lambda^{0,\underline{u}}$, $W^{0,\underline{u}}$ satisfying the PDE (5.3). Moreover, $\lambda^{0,\underline{u}}$ is the limit of the corresponding long-term growth rate $\lambda^{\epsilon,\underline{u}}$ in (2.1) as $\epsilon \to 0$. From Theorem 7.2(a), $\lambda^{\epsilon,\underline{u}} \geq \Lambda^\epsilon$ and hence $\lambda^{0,\underline{u}} \geq \Lambda^0$. If $\lambda^{0,\underline{u}} = \Lambda^0$ for $\underline{u} = \underline{u}^*$, then we call $\underline{u}^*$ an optimal Markov control policy.

A formal application of dynamic programming suggests that one should take

$$\underline{u}^*(x) \in \arg\min_{u \in U}[f(x,u) \cdot \nabla W^0(x) + L(x,u)]. \tag{8.17}$$

However, this formalism encounters evident, well-known difficulties. For example, the gradient $\nabla W^0(x)$ of the viscosity solution W^0 to (8.3) may not exist for all x.

In view of these difficulties, let us merely state and prove some results which hold in a "classical" setting. Suppose that $W^0 \in C^2(\mathbb{R}^n)$ is a solution to (8.3), with ∇W^0 bounded. Moreover, suppose that $\underline{u}^*$ is Lipschitz continuous on each compact subset of $\mathbb{R}^n$ and satisfies (8.17) for all $x \in \mathbb{R}^n$. Let

$$g^*(x) = g(x, \underline{u}^*(x)), \qquad \ell^*(x) = L(x, \underline{u}^*(x)), k^*(x,v) = \ell^*(x) - \gamma^2|v|^2.$$

By (8.3) and (8.17),

$$\Lambda^0 = g^*(x) \cdot \nabla W^0(x) + \max_{v \in \mathbb{R}^n}[v \cdot \nabla W^0(x) + k^*(x,v)],$$

which is just equation (5.3) with $g = g^*, k = k^*$. Let ξ_t^0 satisfy (5.5) with $g = g^*$. Formula (5.6) with $k = k^*, \lambda^0 = \Lambda^0$ is now elementary, since W^0 is a classical solution. An optimal $v^*_\cdot$ in (5.6) is given via the Markov control policy $\underline{v}^*(x) = (2\gamma^2)^{-1}\nabla W^0(x)$, namely, $v_t^* = \underline{v}^*(\xi_t^*)$, where ξ_t^* is the corresponding solution to (5.5) with $\xi_0^* = x$. We have not assumed that g^* satisfies (3.1c). However, the corresponding assumption (7.2) implies that given $R > 0$ there exists R_1 such that $|\xi_t^0| \le R_1$ whenever $|\xi_0^0| \le R$. As in the proof of Corollary 5.3,

$$\Lambda_0 = \lim_{T\to\infty} \frac{1}{T} V^{0*}(T, x),$$

where $V^{0*} = V^0$ for the above choices g^*, k^*. We then define, as in (5.11),

$$J^{0*}(x; v_\cdot) = \limsup_{T\to\infty} \frac{1}{T} \int_0^T k^*(\xi_t^0, v_t)dt.$$

The proof of Theorem 5.5 is unchanged. Thus

$$\Lambda^0 = \sup_{v_\cdot \in \mathcal{W}_M} J^{0*}(x; v_\cdot).$$

Moreover, (5.13) holds: for every T and $v_\cdot \in \mathcal{W}_M$

$$W^0(\xi_T^*) - W^0(x) \le -\int_0^T [\ell^*(\xi_t^0) - \gamma^2|v_t|^2]dt + \Lambda^0 T.$$

In particular, if $\Lambda^0 = 0$ this is the dissipation inequality (5.15).

Appendix A. *Proof of Lemma* 3.1. Let $\xi_\cdot$ be a solution to (3.6), and let $k \ge |x|$. We define the stopped process $\xi_\cdot^k$ by

$$\tau_k = \inf\{t : |\xi_t| \ge k\},$$

$$\xi_t^k = \begin{cases} \xi_t & \text{if } t \le \tau_k, \\ \xi_{\tau_k} & \text{if } t > \tau_k. \end{cases}$$

By the continuity of ∇F and the boundedness of $\xi_\cdot^k$, we see that

$$\int_0^t \nabla F(\xi_r^k) \cdot dw_r$$

is a square-integrable martingale. Therefore, by Ito's rule

$$E_x F(\xi_t^k) = F(x) + E_x \left[\int_0^{t\wedge\tau_k} \hat{G}F(\xi_r^k)dr\right], \tag{A.1}$$

where

$$\hat{G}\psi = \frac{\epsilon}{4\gamma^2}\Delta\psi + (g + v)\cdot\nabla\psi.$$

Note that, when $|v| \le M$,

$$\begin{aligned}\hat{G}F(\xi_t^k) \le \frac{KF(\xi_t^k)}{\sqrt{1+|\xi_t^k|^2}} \Bigg\{ & 2KM|\xi_t^k| - 2cK|\xi_t^k|^2 \\ & + \frac{\epsilon}{4\gamma^2}(n+K)(1+|\xi_t^k|)\Bigg\} \\ \equiv h(\xi_t^k).\end{aligned}$$

There exists $R_0 < \infty$ such that $h(x) \le 0$ if $|x| \ge R_0$. Thus, letting $C_2 = \max_{|x| \le R_0} h(x)$, we have by (A.1)

$$E_x F(\xi_t^k) \le F(x) + C_2 t \text{ for all } t < \infty. \tag{A.2}$$

Since C_2 is independent of k, applying Fatou's lemma to (A.2) yields

$$E_x F(\xi_t) \le F(x) + C_2 t. \tag{A.3}$$

Applying Ito's rule to the original system, (3.6), one has

$$\begin{aligned} F(\xi_t) = F(x) &+ \int_0^t \hat{G}F(\xi_r)dr \\ &+ \sqrt{\frac{\epsilon}{2\gamma^2}}K \int_0^t \frac{F(\xi_r)}{\sqrt{1+|\xi_r|^2}}\xi_r \cdot dw_r. \end{aligned} \tag{A.4}$$

But since (A.3) holds for arbitrary K, the last term in (A.4) is a square-integrable martingale. Therefore, taking expectations, one obtains

$$E_x F(\xi_t) = F(x) + E_x \int_0^t \hat{G}F(\xi_r)dr. \tag{A.5}$$

Note that (3.1c) implies that there exists $m < \infty$ such that $x \cdot g(x) \le -c|x|^2 + m|x|$ for all x. This yields a $C_3 < \infty$ such that

$$\hat{G}F \le [C_3 - c \log F]F.$$

Employing this inequality, Tonelli's theorem, and (A.5) yields

$$E_x F(\xi_t) \le F(x) + \int_0^t E_x\{[C_3 - c\log G(\xi_r)]G(\xi_r)\}ds.$$

Then, by Jensen's inequality, we have

$$E_x F(\xi_t) \le F(x) + \int_0^t \{C_3 - c \log E_x F(\xi_r)\} E_x F(\xi_r) dr.$$

A simple Lyapunov argument then yields the result. □

Appendix B. In this appendix we review, without proofs, some results about semilinear parabolic PDEs which were used in §§4 and 7. Consider an initial value problem for a PDE of the form

$$\frac{\partial V}{\partial T} + \rho V = D\Delta_x V + H(x, \nabla_x V), \qquad \rho > 0, D > 0 \tag{B.1}$$

with the initial data

$$V(0, x) = 0. \tag{B.2}$$

In (4.3), we have $D = \epsilon(4\gamma^2)^{-1}$ and

$$H(x,p) = \max_{|v| \le M} [(g(x) + v) \cdot p + k(x, v)], \tag{B.3}$$

with $k(x,v) = \ell(x) - \gamma^2|v|^2$, while in (7.14) we have the same D and

$$H(x,p) = \min_{u \in V}[f(x,u) \cdot p + L(x,u)] \tag{B.4}$$
$$+ \max_{|v| \leq M}[v \cdot p - \gamma^2 |v|^2].$$

Assumptions (3.1a) and (3.1b), (7.1)–(7.2) in both cases lead to

$$\text{(B.5)}\qquad \begin{aligned} &\text{(a)} \quad |H(x,p) - H(x',p)| \leq C_1(1+|p|)|x - x'|, \\ &\text{(b)} \quad |H(x,p) - H(x,p')| \leq C_2(1+|x|)|p - p'| \end{aligned}$$

for suitable constant C_1, C_2. If H is as in either (B.3) or (B.4), then there exists a "classical" solution $V(T,x)$ to (B.1)–(B.2) with all partial derivatives $V_T, V_{x_i}, V_{x_i x_j}$, $i,j = 1,\ldots,n$, continuous and $|\nabla_x V|$ bounded. In fact, these partial derivatives are Hölder continuous on every compact set. (In §§3 and 4 this solution is denoted by $V_\rho(T,x)$, and in §7 by $\bar{V}(T,x)\cdot$.) This existence theorem can be obtained from standard existence theorems for parabolic PDEs (see [17], [29], or [31]) by making some approximations. This procedure is described in detail in [30], and hence we merely sketch it here. If in addition to (3.1a) and (3.1b), the function g is bounded, then on the right side of (B.5b) one can replace $C_2(1+|x|)$ by some constant C_3. Existence of a classical solution then follows from [31, Thm. 14] or corresponding results in [17], [29]. One then approximates g by a sequence of bounded functions g_k, with $(g_k)_x$ bounded independent of $k = 1, 2, \ldots$. If we use the control interpretation (4.2) for the correponding solution $V_k(T,x)$, a uniform bound for $|\nabla_x V_k|$ is obtained. (Unlike the bound (4.4b) the bound may depend on T.) Once a uniform bound for $|\nabla_x V_k|$ is obtained, estimates for solutions of linear parabolic PDEs imply Hölder estimates for all partial derivatives $(V_k)_T, (V_k)_{x_i x_j}, i,j = 1,\ldots,n$, which are uniform on every compact set. Finally, the uniform bound $0 \leq V_k \leq \rho^{-1}\|\ell\|$ holds. We obtain V as the limit of V_k as $k \to \infty$ through some sequence.

The Hölder estimates for $V_{x_i}, V_{x_i x_j}$ are also uniform with respect to ρ and T, since the bound $|\nabla_x V| \leq \|\nabla \ell\| c^{-1}$ holds independent of ρ and T according to (4.4b). (Here $V = V_\rho$ in the notation of §4.) Therefore, the limit $W = W_\rho$ as $T \to \infty$ is a classical solution to the semilinear elliptic PDE (4.9).

For §7, with H as in (B.4), one proceeds similarly by making bounded approximations f_k to f, with $(f_k)_x$ uniformly bounded. A uniform bound for $|\nabla_x V_k|$ can be obtained by using the interpretation of $V_k(T,x)$ as the upper value of a stochastic differential game (with f replaced by f_k in (7.10).)

REFERENCES

[1] E. N. Barron and R. Jensen, *Total risk aversion, stochastic optimal control and differential games*, Appl. Math. Optim., 19 (1989), pp. 313–327.

[2] T. Basar and P. Bernhard, *H^∞-Optimal Control and Related Minimax Design Problems*, Birkhäuser, Boston, 1991.

[3] A. Bensoussan, *Perturbation Methods in Optimal Control*, John Wiley, New York, 1988.

[4] M. G. Crandall and P. -L. Lions, *Viscosity solutions of Hamilton–Jacobi equations*, Trans. Amer. Math. Soc., 277 (1983), pp. 1–38.

[5] ———, *Remarks on the existence and uniqueness of unbounded viscosity solutions of Hamilton–Jacobi equations*, Illinois J. Math., 31 (1987), pp. 665–688.

[6] M. D. Donsker and S. R. S. Varadhan, *Asymptotic evaluation of certain Markov process expectations for large time*, I, II, III, Comm. Pure Appl. Math., 28 (1975), pp. 1–45, 279–301; 29 (1976), pp. 389–461.

[7] P. Dupuis and R. S. Ellis, *A Weak Convergence Approach to the Theory of Large Deviations*, John Wiley, New York, 1995, to appear.

[8] R. J. Elliott and N. J. Kalton, *Boundary value problems for nonlinear partial differential operators*, J. Math. Anal. Appl., 46 (1974), pp. 228–241.

[9] L. C. Evans and P. E. Souganidis, *Differential games and representation formulas for solutions of Hamilton–Jacobi–Isaacs equations*, Indiana Univ. Math. J., 33 (1984), pp. 773–797.

[10] W. H. Fleming, *The convergence problem for differential games, Part* I, J. Math. Anal. Appl., 3 (1961), pp. 102–116; part II, in Advances in Game Theory, Annals of Math. Studies 52, Princeton University Press, Princeton, NJ, 1964, pp. 195–210.

[11] W. H. Fleming and W. M. McEneaney, *Risk Sensitive Control and Differential Games*, Springer Lecture Notes in Control and Info. Sci. 184, 1992, Springer-Verlag, New York, pp. 185–197.

[12] W. H. Fleming and W. M. McEneaney, *Risk sensitive control with ergodic cost criteria*, in Proc. 31st IEEE Conf. on Decision and Control, Tucson, AZ, Dec. 1992.

[13] W. H. Fleming and R. W. Rishel, *Deterministic and Stochastic Optimal Control*, Springer-Verlag, New York, 1975.

[14] W. H. Fleming, S.-J. Sheu, and H. M. Soner, *A remark on the large deviations of an ergodic Markov process*, Stochastics 22 (1987), pp. 187–199.

[15] W. H. Fleming and H. M. Soner, *Controlled Markov Processes and Viscosity Solutions*, Springer-Verlag, New York, 1992.

[16] W. H. Fleming and P. E. Souganidis, *On the existence of value functions of two-player, zero-sum stochastic differential games*, Indiana Univ. Math J., 38 (1989), pp. 293–314.

[17] A. Friedman, *Partial Differential Equations of Parabolic Type*, Prentice-Hall, Englewood Cliffs, NJ, 1964.

[18] A. Friedman, *Differential Games*, John Wiley, New York, 1971.

[19] D. Gilbarg and N. Trudinger, *Elliptic Differential Equations of Second Order*, 2nd ed., Springer-Verlag, 1985.

[20] K. Glover and J. C. Doyle, *State-space formulae for all stabilizing controllers that satisfy an H^∞-norm bound and relations to risk sensitivity*, Systems Control Lett., 11 (1988), pp. 167–172.

[21] C. J. Holland, *A new energy characterization of the smallest eigenvalue of the Schrödinger equation*, Comm. Pure Appl. Math., 3 (1977), pp. 755–765.

[22] A. Isidori and A. Astolfi, *Nonlinear H^∞-control via measurement feedback*, J. Math Systems, Estim. Control, 2 (1992), pp. 31–44.

[23] A. Isidori and C. J. Byrnes, *Output regulation of nonlinear systems*, IEEE Trans. Automat. Control, 35 (1990), pp. 131–140.

[24] D. H. Jacobson, *Optimal stochastic linear systems with exponential criteria and their relation to deterministic differential games*, IEEE Trans. Automat. Control, AC-18 (1973), pp. 124–131.

[25] M. R. James, *Asymptotic analysis of nonlinear stochastic risk-sensitive control and differential games*, Math. Control Signals Systems, 5 (1992), pp. 401–417.

[26] M. R. James, R. J. Elliott, and J. S. Baras, *Risk sensitive control and dynamic games with partially observed discrete-time nonlinear systems*, IEEE Trans. Automat. Control, AC-39 (1994), pp. 780–792.

[27] I. Karatzas and S. E. Shreve, *Brownian Motion and Stochastic Calculus*, Springer-Verlag, New York, 1988.

[28] N. V. Krylov, *Controlled Diffusion Processes*, Springer-Verlag, New York, 1980.

[29] O. A. Ladyzhenskaya, V. A. Solonnikov, and N. N. Uralseva, *Linear and Quasilinear Equations of Parabolic Type*, American Math. Soc., Providence, RI, 1968.

[30] W. M. McEneaney, *Connections between Risk-Sensitive Stochastic Control, Differential Games and H-infinity Control: The Nonlinear Case*, Ph.D. Thesis, Brown University, 1993.

[31] O. A. Oleinik and S. N. Kruzhkov, *Quasilinear parabolic equations of second order with many independent variables*, Russian Math. Surveys, 16 (1961), pp. 105–146.

[32] T. Runolfsson, *Stationary risk-sensitive LQG control and its relation to LQG and H^∞-control*, in Proc. 29th IEEE CDC, Honolulu, HI, Dec. 1990, pp. 1018–1023.

[33] ———, *The Equivalence between Infinite Horizon Control of Stochastic Systems with Exponential-of-Integral Performance Index and Stochastic Differential Games*, Technical report JHU/ECE 91-07, Johns Hopkins University, 1991.

[34] S.-J. Sheu, *Stochastic control and exit probabilities of jump Markov processes*, SIAM J. Control Optim., 23 (1985), pp. 306–328.

[35] P. E. SOUGANIDIS, *Approximation schemes for viscosity solutions of Hamilton–Jacobi equations with applications to differential games*, J. Nonlinear Anal., T.M.A., 9 (1985), pp. 217–257.

[36] D. W. STROOCK, *An Introduction to the Theory of Large Deviations*, Springer-Verlag, New York, 1984.

[37] A. J. VAN DER SCHAFT, *Nonlinear state space H^∞ control theory*, in Perspectives in Control, H. L. Trentelman and J. C. Willems, eds., Progressess in Systems and Control, 2nd ECC, Groningen, Birkhäuser, Boston, 1993.

[38] A. YU. VERETENNIKOV, *On strong solutions and explicit formulas for solutions of stochastic integral equations*, Math USSR-Sb., 39 (1981), pp. 387–403.

[39] P. WHITTLE, *Risk Sensitive Optimal Control*, John Wiley, New York, 1990.

[40] ———, *A risk sensitive maximum principle*, Syst. Control Lett., 15 (1990), pp. 183–192.

[41] ———, *A risk sensitive maximum principle: The case of imperfect state information*, IEEE Trans. Automat. Control, 36 (1991), pp. 793–801.

[42] W. H. FLEMING AND M. R. JAMES, *The risk-sensitive index and the H_2 and H_∞ norms for nonlinear systems*, Math. Control Signals Systems, to appear.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 33, No. 6, pp. 1916–1925, November 1995

© 1995 Society for Industrial and Applied Mathematics

015

PARALLEL GRADIENT DISTRIBUTION IN UNCONSTRAINED OPTIMIZATION*

O. L. MANGASARIAN†

Abstract. A parallel version is proposed for a fundamental theorem of serial unconstrained optimization. The parallel theorem allows each of k parallel processors to use simultaneously a different algorithm, such as a descent, Newton, quasi-Newton, or conjugate gradient algorithm. Each processor can perform one or many steps of a serial algorithm on a portion of the gradient of the objective function assigned to it, independently of the other processors. Eventually a synchronization step is performed which, for differentiable convex functions, consists of taking a strong convex combination of the k points found by the k processors. A more general synchronization step, applicable to convex as well as nonconvex functions, consists of taking the best point found by the k processors or any point that is better. The fundamental result that we establish is that any accumulation point of the parallel algorithm is stationary for the nonconvex case and is a global solution for the convex case. Computational testing on the Thinking Machines CM-5 multiprocessor indicates a speedup of the order of the number of processors employed.

Key words. parallel optimization, gradient methods, unconstrained optimization

AMS subject classification. 90C30

1. Introduction.

In this work we are interested in parallel algorithms for solving the unconstrained minimization problem

$$\min_{x \in R^n} f(x), \tag{1}$$

where f is a differentiable function from the n-dimensional real space R^n into R. The basic idea behind our approach is to assign a portion of the gradient ∇f of f to one of k processors, let each processor perform one or more steps of a serial algorithm on its portion of the gradient, and then synchronize the processors eventually. The synchronization consists of taking a strong convex combination of the k points found by the k processors when f is convex. For nonconvex f, the best point found by the k processors can be taken, or any other point with a lower value of f will work.

The fundamental theorem we intend to parallelize is related to some classical forcing function theorems given in [7], [4], [11] that establish convergence for a wide class of algorithms. Such algorithms typically consist of a direction choice followed by a stepsize choice. The combined direction–stepsize choice generates a decrease in the objective function that forces the eventual satisfaction of an optimality condition, namely, the vanishing of the gradient. Direction choices include descent directions, Newton, quasi-Newton, and conjugate directions. Stepsize choices along the chosen direction include minimization, finding the first stationary point, interval stepsize, the Armijo stepsize, and others. Related algorithms, wherein the objective function is sequentially minimized with respect to certain variables, include the serial algorithm proposed by Warga [16] for a strictly convex function in each block of variables and in which the function is sequentially minimized for each block of variables, and the coordinate descent methods of Tseng [15] and Luo and Tseng [8]. Other parallelization schemes are discussed extensively in [2].

* Received by the editors June 15, 1993; accepted for publication (in revised form) July 9, 1994.

† Computer Sciences Department, University of Wisconsin, 1210 West Dayton Street, Madison, Wisconsin 53706 (olvi@cs.wisc.edu). This research was supported by Air Force Office of Scientific Research grants AFOSR-89-0410 and F49620-94-1-0036 and National Science Foundation grants CCR-9101801, CCR-9322479, and CDA-9024618.

We note that our parallelization proofs are direct extensions of those for general serial algorithms. However the resulting parallel algorithms are quite general and have significant theoretical and computational implications. For example, the parallelization proposed here played an important role in establishing the convergence and computational results of the parallel back-propagation algorithm of neural networks [9], the parallel variable distribution algorithm for unconstrained and constrained optimization [6], and the parallel multicategory discrimination problem [1].

We give now an outline of the paper. In §2 we establish two serial convergent algorithm theorems, 2.1 and 2.2 (SCAT1 and SCAT2), which cover many unconstrained direction–stepsize algorithms that are suitable for parallelization. We also give a number of specific instances of well-known algorithms satisfying conditions of these theorems. In §3 we establish a number of parallel convergent algorithm theorems that utilize the serial algorithms. In Theorem 3.1 (convex PCAT1), which covers the convex case, each processor takes one step of any serial algorithm covered by SCAT1 or SCAT2, and then a strong convex combination (positively weighted average) of all the points is taken as the next iterate. Corollary 3.1 (nonconvex PCAT1) differs from convex PCAT1 in that the synchronization step consists of taking the best point found by the k processors or searching for a better point. By better we mean, of course, lower f value. Corollary 3.2 (partially asynchronous nonconvex PCAT1) allows partial asynchronization among the k processors in the sense that each processor is free to perform any number of steps of the serial algorithm that is desirable (say, until further improvement in each processor is very small), followed by a synchronization step that consists of taking the best point or searching for a better one. Theorem 3.2 (partially asynchronous nonconvex PCAT2) combines, in a manner similar to SCAT2 for the serial case, the direction and stepsize choices of Corollary 3.2 into a simpler and more general forcing function condition (20). However, this theorem is not as suggestive of an explicit computational scheme as the partially asynchronous nonconvex PCAT1 of Corollary 3.2. In the concluding section we report briefly on computational experience with parallel gradient distribution algorithms on multicategory discrimination problems [1] and on publicly available test problems [6] from the constrained and unconstrained testing environment CUTE [3]. Computations were carried out on the Thinking Machines CM-5 multiprocessor. Speedup efficiency depended on problem size and number of processors employed (2 to 32) and averaged between 129% and 20%.

We now briefly describe our notation. The sequence $\{x_i\}$, $i = 0, 1, \ldots,$ will represent iterates in the n-dimensional real space R^n generated by some algorithm. For $\ell = 1, \ldots, k$, $x_i^\ell \in R^{n^\ell}$ will represent an n^ℓ-dimensional subset of components of x_i, where $\sum_{\ell=1}^k n^\ell = n$. The complement of ℓ in $\{1, \ldots, k\}$ will be denoted by $\bar{\ell}$, and we write $x_i = (x_i^\ell, x_i^{\bar{\ell}})$, $\ell = 1, \ldots, k$. For a differentiable function $f\colon R^n \to R$, ∇f will denote the n-dimensional vector of partial derivatives with respect to x, and $\nabla_\ell f$ will denote the n^ℓ-dimensional vector of partial derivatives with respect to $x^\ell \in R^{n^\ell}$, $\ell = 1, \ldots, k$. For k points y in R^n, $\sum_{j=1}^k \lambda_j y_j$, such that $\lambda_j \geq \delta > 0$ and $\sum_{j=1}^k \lambda_j = 1$, is said to be a strong convex combination of the points y_j, $j = 1, \ldots, k$. If f has continuous first partial derivatives on R^n, we say $f \in C^1(R^n)$. If f has Lipschitz continuous first partial derivatives on R^n with constant $K > 0$, that is,

$$\|\nabla f(y) - \nabla f(x)\| \leq K\|y - x\| \qquad \forall x,\ y \in R^n,$$

we write $f \in LC_K^1(R^n)$. Here and throughout, $\|\cdot\|$ denotes the two norm, that is, $\|z\| = (z^T z)^{\frac{1}{2}}$ for z in a finite-dimensional real space of unspecified dimension.

2. Serial convergent algorithm theorems. We begin first with a simple serial convergent algorithm theorem (SCAT1) for the solution of the unconstrained minimization problem (1). The theorem is related to some classical forcing function theorems given in [7] and [4] that establish convergence for a wide class of algorithms that consist of a direction choice followed by a stepsize choice. The decrease in the objective function forces the satisfaction of an optimality condition, namely, the vanishing of the gradient. Before stating and proving SCAT1 we adapt the definition of a forcing function [10, p. 479] for our purposes.

DEFINITION 2.1 (forcing function). *A continuous function σ from the nonnegative real line R_+ into itself such that $\sigma(0) = 0$, $\sigma(\zeta) > 0$ for $\zeta > 0$, and for the sequence of nonnegative real numbers $\{\zeta_i\}$*

$$\{\sigma(\zeta_i)\} \rightarrow 0 \text{ implies } \{\zeta_i\} \rightarrow 0$$

is said to be a forcing function on the sequence $\{\zeta_i\}$.

Some simple typical examples of forcing functions are

$$\alpha\zeta,\ \alpha\zeta^2,\ \max\{\sigma_1(\zeta),\ \sigma_2(\zeta)\},\ \min\{\sigma_1(\zeta),\ \sigma_2(\zeta)\} \text{ and } \sigma_2(\sigma_1(\zeta)),$$

where α is a positive number and $\sigma_1(\zeta)$ and $\sigma_2(\zeta)$ are forcing functions. We now state and prove SCAT1.

THEOREM 2.1 (serial convergent algorithm theorem 1 (SCAT1)). *Let $f \in C^1(R^n)$. Start with any $x_0 \in R^n$. Given x_i, stop if $\nabla f(x_i) = 0$, else compute x_{i+1} from a direction d_i and stepsize λ_i as follows. Choose direction d_i such that*

$$-\nabla f(x_i)^T d_i \geq \sigma_1(\|\nabla f(x_i)\|), \tag{2}$$

where σ_1 is a forcing function on $\{\|\nabla f(x_i)\|\}$. Choose stepsize λ_i such that

$$x_{i+1} = x_i + \lambda_i d_i \tag{3}$$

and

$$f(x_i) - f(x_{i+1}) \geq \sigma_2(-\nabla f(x_i)^T d_i) \geq 0, \tag{4}$$

where σ_2 is a forcing function on the sequence of nonnegative real numbers $\{-\nabla f(x_i)^T d_i\}$ for bounded $\{d_i\}$. Then either $\{x_i\}$ terminates at a stationary point $x_{\bar{i}}$, that is, $\nabla f(x_{\bar{i}}) = 0$, or $\nabla f(\bar{x}) = 0$ for each accumulation point $(\bar{x}, \bar{d})$ of the sequence $\{x_i, d_i\}$.

Proof. The algorithm terminates at an $x_{\bar{i}}$ only if $\nabla f(x_{\bar{i}}) = 0$. Suppose now it does not terminate and that $\{(x_{i_j},\ d_{i_j})\} \rightarrow (\bar{x},\ \bar{d})$. Since f is continuous, $\lim_{j\rightarrow\infty} f(x_{i_j}) = f(\bar{x})$. By the stepsize condition (4), the sequence $\{f(x_i)\}$ is nonincreasing and has an accumulation point $f(\bar{x})$, and hence converges to $f(\bar{x})$. By (4) and the continuity of $\sigma_2(\zeta)$

$$0 = \lim_{j\rightarrow\infty} f(x_{i_j}) - f(x_{i_j+1}) \geq \lim_{j\rightarrow\infty} \sigma_2(-\nabla f(x_{i_j})^T d_{i_j}) = \sigma_2(-\nabla f(\bar{x})^T \bar{d}) \geq 0.$$

Hence $\nabla f(\bar{x})^T \bar{d} = 0$. But by the direction condition (2)

$$0 = -\nabla f(\bar{x})^T \bar{d} = -\lim_{j\rightarrow\infty} \nabla f(x_{i_j})^T d_{i_j} \geq \lim_{j\rightarrow\infty} \sigma_1(\|\nabla f(x_{i_j})\|) = \sigma_1(\|\nabla f(\bar{x})\|) \geq 0.$$

Hence $\nabla f(\bar{x}) = 0$. □

We note that the boundedness condition on $\{d_i\}$, which does not restrict Theorem 2.1, was not explicitly used in the proof. However, this condition simplifies the application of the theorem to specific stepsize choices, such as the first stationary point and Armijo stepsize choices given below.

We give now examples of direction and stepsize choices that satisfy the assumptions of Theorem 2.1.

Example 2.1 (serial direction choices). For $f \in C^1(R^n)$ and σ a forcing function, a direction $d_i \in R^n$ satisfying any of the following conditions will satisfy condition (2):

(i) descent direction:

$$-d_i^T \nabla f(x_i) \geq \alpha \|\nabla f(x_i)\|^\beta \text{ for some } \alpha > 0,\ \beta > 0.$$

(ii) quasi-Newton direction:

$$d_i = -H_i \nabla f(x_i),\ H_i \in R^{n \times n},\ z^T H_i z \geq \alpha \|z\|^2 \qquad \forall z \in R^n, \text{ for some } \alpha > 0.$$

(iii) conjugate direction:

$$d_i = -\nabla f(x_i) + \alpha_i d_{i-1},$$

$$\frac{\|\nabla f(x_i)\|^2}{\|\nabla f(x_i)\| + |\alpha_i| \|d_{i-1}\|} \geq \sigma(\|\nabla f(x_i)\|), \tag{5}$$

where σ is a forcing function on $\{\|\nabla f(x_i)\|\}$.

We note that the conjugate direction conditions of (iii) are satisfied by the Polyak–Polak–Ribière [11]–[14] coefficient

$$\alpha_i := \frac{(\nabla f(x_i) - \nabla f(x_{i-1}))^T \nabla f(x_i)}{\|\nabla f(x_{i-1})\|^2} \tag{6}$$

for $f \in C^2(R^n)$ and such that

$$\beta \|z\|^2 \geq z^T \nabla^2 f(x) z \geq \alpha \|z\|^2 \qquad \forall z \in R^n \text{ for some } \beta \geq \alpha > 0. \tag{7}$$

We also note that the Newton direction $d_i = -\nabla^2 f(x_i)^{-1} \nabla f(x_i)$ satisfies (ii) above under the same condition (7).

We give now stepsize choices that satisfy conditions (3) and (4) of Theorem 2.1.

Example 2.2 (serial stepsize choices). For $d_i \in R^n$ and $f \in C^1(R^n)$, a $\lambda_i \geq 0$ satisfying any one of the following conditions will satisfy conditions (3) and (4) of Theorem 2.1:

(i) minimum along d_i:

$$\lambda_i \in \arg\min_{\lambda \geq} f(x_i + \lambda d_i),\ f \in LC_K^1(R^n);$$

(ii) first stationary point:

$$\lambda_i \in \arg\min_{\lambda \geq} \{\lambda | \nabla f(x_i + \lambda d_i)^T d_i = 0\},\ f \in LC_K^1(R^n);$$

(iii) interval stepsize:

$$0 < \varepsilon_1 \leq \lambda_i \leq \frac{2}{\rho K} - \varepsilon_2,\ \|d_i\|^2 \leq -\rho \nabla f(x_i)^T d_i,\ f \in LC^1_K(R^n)$$

for some $\varepsilon_1 > 0$, $\varepsilon_2 > 0$ and $\rho > 0$;

(iv) Armijo [5, pp. 118–119]:

$$\lambda_i = \max\left\{\bar{\lambda}_i,\ \frac{\bar{\lambda}_i}{2}, \ldots\right\} \text{ such that}$$

$$f(x_i) - f(x_i + \lambda_i d_i) \geq -\delta \lambda_i \nabla f(x_i)^T d_i \quad \text{for some } \delta \in (0,1),$$

and $\bar{\lambda}_i \geq (\sigma(-\nabla f(x_i)^T d_i))/(-\nabla f(x_i)^T d_i)$, where σ is a forcing function and $f \in LC^1_K(R^n)$.

It takes a bit of algebra to show that each of the four stepsizes (i) to (iv) above satisfy conditions (3) and (4) of Theorem 2.1. We omit the details here.

We note that Theorem 2.1 can be written in a more general and simpler but algorithmically less suggestive form by combining conditions (2) and (4) into the single condition (8) below. This results in the following theorem, the proof of which either follows from that of Theorem 2.1 or can be given in a few lines as is done below.

THEOREM 2.2 (serial convergent algorithm theorem 2 (SCAT2)). *Let $f \in C^1(R^n)$. Start with any $x_0 \in R^n$. Given x_i, stop if $\nabla f(x_i) = 0$, else determine x_{i+1} such that*

$$f(x_i) - f(x_{i+1}) \geq \sigma(\|\nabla f(x_i\|), \tag{8}$$

where σ is a forcing function on $\{\|\nabla f(x_i)\|\}$. Then either $\{x_i\}$ terminates at a stationary point $x_{\bar{i}}$ or each accumulation point $\bar{x}$ of $\{x_i\}$ is stationary.

Proof. Suppose $\nabla f(x_i) \neq 0$ for all i and that $\{x_{i_j}\}$ converges to $\bar{x}$. Since the nonincreasing sequence $\{f(x_i)\}$ has an accumulation $f(\bar{x})$, it converges to $f(\bar{x})$. By (8) we have that

$$0 = \lim_{j\to\infty} (f(x_{i_j}) - f(x_{i_j+1})) \geq \lim_{j\to\infty} \sigma(\|\nabla f(x_{i_j})\|) \geq 0.$$

Hence $\lim_{j\to\infty} \|\nabla f(x_{i_j})\| = 0$ and $\nabla f(\bar{x}) = 0$. □

We note that the full sequences $\{f(x_i)\}$ and $\{\|\nabla f(x_i)\|\}$ converge if f is bounded below. We state this as the following corollary.

COROLLARY 2.1 (function and gradient convergence). *Let f be bounded below on the level set $S(x_0) = \{x | f(x) \leq f(x_0)\}$. Then the sequence $\{f(x_i)\}$ of Theorems* 2.1 *and* 2.2 *converges, and* $\lim_{i\to\infty} \|\nabla f(x_i)\| = 0$.

Proof. From (8) the sequence $\{f(x_i)\}$ is nonincreasing, and since $\{x_i\}$ remains in $S(x_0)$, $\{f(x_i)\}$ is bounded below and hence converges. From (8), we have that $\lim_{i\to\infty} \sigma(\|\nabla f(x_i)\|) = 0$, and hence $\lim_{i\to\infty} \|\nabla f(x_i)\| = 0$. □

We now proceed to establish parallel versions of Theorems 2.1 and 2.2 and other parallel results.

3. Parallel convergent algorithm theorems.

We shall establish in this section parallel versions of Theorems 2.1 and 2.2. The importance of these theorems, PCAT1 and PCAT2, is that they enable each of k processors to perform, on a portion of the gradient that is assigned to it, one or more iterations of the serial algorithms independently of the other processors. The processor picks a direction and stepsize based on the partial gradient assigned to it. A simple synchronization step follows in which a new point is generated by a strong convex combination of the k points obtained by the k processors for the convex case and by using the best, or better, point obtained by the k processors for the nonconvex case. We first state and prove Theorem 3.1, our parallel theorem for the convex case. Corollary 3.1 extends Theorem 3.1 to the nonconvex case. Corollary 3.2 further extends Corollary 3.1 by allowing partial asynchronization by letting each processor take as many steps as desirable. Finally Theorem 3.2 gives a more general version of Theorem 3.1 for the nonconvex case. We note here that a referee pointed out that the distribution of the gradient can also be made with respect to subspaces induced by other decompositions of R^n. For example, the iterate x_i can be decomposed into $x_i^\ell = P_\ell x_i$, $\ell = 1, \ldots, k$, instead of into subvectors x_i^ℓ of x_i. Here $P_1, \ldots, P_k$ are projection matrices (that is $P_i^2 = P_i$, $P_i^T = P_i$, $i = 1, \ldots, k$) such that $\sum_{i=1}^k P_i = I$.

THEOREM 3.1 (convex parallel convergent algorithm theorem 1 (convex PCAT1)). *Let $f \in C^1(R^n)$ be convex on R^n. Start with any $x_0 \in R^n$. Given x_i, stop if $\nabla f(x_i) = 0$, else compute x_{i+1} from directions $d_i^\ell \in R^{n^\ell}$ and stepsizes $\lambda_i^\ell \in R$, $\ell = 1, \ldots, k$, $\sum_{\ell=1}^k n^\ell = n$. Choose directions d_i^ℓ such that*

$$-\nabla_\ell f(x_i)^T d_i^\ell \geq \tau_\ell(\|\nabla_\ell f(x_i)\|),\ \ell = 1, \ldots k, \tag{9}$$

where τ_ℓ is a forcing function on $\{\|\nabla_\ell f(x_i)\|\}$, $\ell = 1, \ldots, k$. Choose stepsizes λ_i^ℓ, choose λ_i^ℓ, $\ell = 1, \ldots, k$ such that for $\bar{\ell}$, the complement of ℓ in $\{1, \ldots, k\}$,

$$f(x_i) - f(x_i^\ell + \lambda_i^\ell d_i^\ell,\ x_i^{\bar{\ell}}) \geq \mu_\ell(-\nabla_\ell f(x_i)^T d_i^\ell) \geq 0, \qquad \ell = 1, \ldots, k, \tag{10}$$

where μ_ℓ is a forcing function on the sequence of nonnegative real numbers $\{-\nabla_\ell f(x_i)^T d_i^\ell\}$ for bounded $\{d_i^\ell\}$, $\ell = 1, \ldots, k$. For synchronization, let

$$x_{i+1}^\ell = x_i^\ell + \nu_i^\ell \lambda_i^\ell d_i^\ell, \qquad \ell = 1, \ldots, k, \tag{11}$$

$$\sum_{\ell=1}^k \nu_i^\ell = 1, \qquad \nu_i^\ell \geq \delta > 0,\ \ell = 1, \ldots, k. \tag{12}$$

Then either $\{x_i\}$ terminates at a solution $x_{\bar{i}}$ of (1) or, for each accumulation point $(\bar{x}, \bar{d})$ of $\{x_i, d_i\}$, $\bar{x}$ is a solution of (1).

Proof. First we show that the sequence $\{f(x_i)\}$ is nonincreasing.

$$\begin{aligned}
&f(x_i) - f(x_{i+1}) \\
&\quad = f(x_i^1, \ldots, x_i^k) - f(x_i^1 + \nu_i^1 \lambda_i^1 d_i^1, \ldots, x_i^k + \nu_i^k \lambda_i^k d_i^k) \\
&\quad = f(x_i^1, \ldots, x_i^k) \\
&\qquad -f\left(\nu_i^1(x_i^1 + \lambda_i^1 d_i^1) + \left(\sum_{\ell=2}^k \nu_i^\ell\right) x_i^1, \ldots, \nu_i^k(x_i^k + \lambda_i^k d_i^k) + \left(\sum_{\ell=1}^{k-1} \nu_i^\ell\right) x_i^k\right) \\
&\quad \geq \nu_i^1[f(x_i^1, \ldots, x_i^k) - f(x_i^1 + \lambda_i^1 d_i^1,\ x_i^2, \ldots\ x_i^k)] \\
&\qquad + \cdots + \nu_i^k[f(x_i^1, \ldots,\ x_i^k) - f(x_i^1, \ldots,\ x_i^{k-1},\ x_i^k + \lambda_i^k d_i^k)] \quad \text{(by convexity of } f) \\
&\quad \geq \delta \sum_{\ell=1}^k \mu_\ell(-\nabla_\ell f(x_i)^T d_i^\ell) \qquad \text{(by (12) and (10))}. \qquad (13)
\end{aligned}$$

Hence

$$f(x_i) - f(x_{i+1}) \geq \delta \sum_{\ell=1}^{k} \mu_\ell(-\nabla_\ell f(x_i)^T d_i^\ell) \geq 0, \tag{14}$$

and the sequence $\{f(x_i)\}$ is nonincreasing.

Now the sequence $\{x_i\}$ of the PCAT1 algorithm terminates only if $\nabla f(x_{\bar{i}}) = 0$, in which case $x_{\bar{i}}$ solves (1). Now suppose that it does not terminate and that $\{(x_{i_j}, d_{i_j})\} \to (\bar{x}, \bar{d})$. Since f is continuous, $\lim_{j\to\infty} f(x_{i_j}) = f(\bar{x})$. Hence the nonincreasing sequence $\{f(x_i)\}$ has an accumulation point $f(\bar{x})$, and consequently the sequence converges to $f(\bar{x})$. By (14) and the continuity of μ_ℓ, $\ell = 1, \ldots, k$, we have that

$$\begin{aligned} 0 &= \lim_{j\to\infty} (f(x_{i_j}) - f(x_{i_j+1})) \\ &\geq \delta \lim_{j\to\infty} \sum_{\ell=1}^{k} \mu_\ell(-\nabla_\ell f(x_{i_j})^T d_{i_j}^\ell) = \delta \sum_{\ell=1}^{k} \mu_\ell(-\nabla_\ell f(\bar{x})^T \bar{d}^\ell) \geq 0. \end{aligned}$$

Hence $\nabla f(\bar{x})^T \bar{d} = 0$. But by the direction condition (9)

$$\begin{aligned} 0 = \nabla f(\bar{x})^T \bar{d} &= -\lim_{j\to\infty} \nabla f(x_{i_j})^T d_{i_j} \\ &\geq \lim_{j\to\infty} \sum_{\ell=1}^{k} \tau_\ell(\|\nabla f_\ell(x_{i_j})\|) = \sum_{\ell=1}^{k} \tau_\ell(\|\nabla f_\ell(\bar{x})\|) \geq 0. \end{aligned}$$

Hence $\nabla_\ell f(\bar{x}) = 0$, $\ell = 1, \ldots, k$, and consequently $\nabla f(\bar{x}) = 0$ and $\bar{x}$ solves (1). □

We note that the convexity of f was needed in (13) in the proof above, as well as to show that the stationary point generated by PCAT1 is a global solution of $\min_{x\in R^n} f(x)$. However, it is easy to extend Theorem 3.1 to nonconvex f by changing the synchronization procedure (11)–(12) to one that takes the best of the points found by the k processors or a better point. We state this as the following corollary.

COROLLARY 3.1 (nonconvex parallel convergence algorithm theorem 1 (nonconvex PCAT1)). *Theorem* 3.1 *holds for nonconvex f, with a resulting stationary point, if the synchronization procedure* (11)–(12) *is replaced by the following:*

For synchronization, find x_{i+1} such that

$$f(x_{i+1}) \leq \min_{1\leq\ell\leq k} f(x_i^\ell + \lambda_i^\ell d_i^\ell, x_i^{\bar{\ell}}). \tag{15}$$

Proof. The only changes needed in the proof of Theorem 3.1 in order to apply it here are the following. Replace δ with $\frac{1}{k}$ in (14) and the string of inequalities of (13), which establish the monotonicity of $\{f(x_i)\}$ through the convexity of f, with the following:

$$\begin{aligned} f(x_i) - f(x_{i+1}) &\geq \frac{1}{k}\left[f(x_i) - f(x_i^1 + \lambda_i^1 d_i^1, x_1^2, \ldots, x_i^k)\right] \\ &\quad + \cdots + \frac{1}{k}\left[f(x_i) - f(x_i^1, \ldots, x_i^{k-1}, x_i^k + \lambda_i^k d_i^k)\right] \quad \text{(by (15))} \\ &\geq \frac{1}{k} \sum_{\ell=1}^{k} \mu_\ell(-\nabla_\ell f(x_i)^T d_i^\ell). \qquad \square \end{aligned}$$

We note now that partial asynchronization of the k processors for the nonconvex PCAT1 is possible if we allow each of the k processors to take as many steps as desired until, say, they encounter slow convergence, provided we terminate each processor ℓ, $\ell = 1, \dots, k$, at a point $(y_i^\ell, x_i^{\bar{\ell}})$ such that

$$f(y_i^\ell, x_i^{\bar{\ell}}) \leq f(x_i^\ell + \lambda_i^\ell d_i, x_i^{\bar{\ell}}), \qquad \ell = 1, \dots k, \tag{16}$$

where λ_i^ℓ, $\ell = 1, \dots, k$, satisfy (10). Such an inequality is easily satisfied, for example, when each processor takes a desired number of steps in R^{n^ℓ} determined by any of the standard serial algorithms described in §2 on the function $f(x_i^\ell, x_i^{\bar{\ell}})$ starting at $(x_i^\ell + \lambda_i^\ell d_i^\ell, x_i^{\bar{\ell}})$. After these parallel steps are performed by each processor, then an eventual synchronization step is needed that consists of determining x_{i+1} such that

$$f(x_{i+1}) \leq \min_{1 \leq \ell \leq k} f(y_i^\ell, x_i^{\bar{\ell}}). \tag{17}$$

We summarize these procedures as the following partially asynchronous algorithm.

COROLLARY 3.2 (partially asynchronous nonconvex PCAT1). *Theorem* 3.1 *holds for nonconvex f, with a resulting stationary point, if the stepsize choices* (10) *and synchronization procedure* (11)–(12) *are changed to the following.*

Partially asynchronous stepsize: Choose y_i^ℓ, $\ell = 1, \dots, k$, such that for $\bar{\ell}$, the complement of ℓ in $\{1, \dots, k, \}$,

$$f(x_i) - f(y_i^\ell, x_i^{\bar{\ell}}) \geq \mu_\ell(-\nabla_\ell f(x_i)^T d_i^\ell) \geq 0, \qquad \ell = 1, \dots, k, \tag{18}$$

where μ_ℓ is a forcing function on the sequence of nonnegative real numbers $\{-\nabla_\ell f(x_i)^T d_i^\ell\}$ for bounded $\{d_i^\ell\}$, $\ell = 1, \dots, k$.

Comment: Inequality (18) *is easily implemented by satisfying* (16) *and* (10).

Synchronization: Find x_{i+1} such that

$$f(x_{i+1}) \leq \min_{1 \leq \ell \leq k} f(y_i^\ell, x_i^{\bar{\ell}}). \tag{19}$$

Proof. The only changes needed in the proof of Theorem 3.1 in order to apply it here are to replace δ with $\frac{1}{k}$ in (14) and the string of inequalities of (13) that establish the monotonicity of $\{f(x_i)\}$ by using (18)and (19) as follows:

$$\begin{aligned} f(x_i) - f(x_{i+1}) &\geq \frac{1}{k}\left[f(x_i) - f(y_i^1, x_i^2, \dots, x_i^k)\right] \\ &\quad + \cdots + \frac{1}{k}\left[f(x_i) - f(x_i^1, \dots, x_i^{k-1}, y_i^k)\right] \quad \text{(by (19))} \\ &\geq \frac{1}{k} \sum_{\ell=1}^{k} \mu_\ell(-\nabla_\ell f(x_i)^T d_i^\ell). \qquad \square \end{aligned}$$

By combining the direction (9) and stepsize (18) choices of the partially asynchronous nonconvex PCAT1 of Corollary 3.2 into a single forcing function condition (20) below, we obtain Theorem 3.2, which is a simpler and more general theorem than PCAT1 of Corollary 3.2. We omit the proof, which is similar to that of Theorem 2.2.

THEOREM 3.2 (partially asynchronous nonconvex PCAT2). *Let $f \in C^1(R^n)$ on R^n. Start with any $x_0 \in R^n$. Given x_i, stop if $\nabla f(x_i) = 0$, else determine x_{i+1} in the following manner.*

Parallel steps: Determine y^ℓ, $\ell = 1, \ldots, k$, *such that for* $\bar{\ell}$, *the complement of* ℓ *in* $\{i, \ldots, k\}$,

$$f(x_i) - f(y_i^\ell, x_i^{\bar{\ell}}) \geq \sigma_\ell(\|\nabla_\ell f(x_i)\|), \qquad \ell = 1, \ldots, k, \tag{20}$$

where σ_ℓ *is a forcing function on* $\{\|\nabla_\ell f(x_i)\|\}$ *for* $\ell = 1, \ldots, k$.

Synchronization step: Choose x_{i+1} *such that*

$$f(x_{i+1}) \leq \min_{1 \leq \ell \leq k} f(y_i^\ell, x_i^{\bar{\ell}}). \tag{21}$$

Either $\{x_i\}$ *terminates at a stationary point* $x_{\bar{i}}$ *or each accumulation point* $\bar{x}$ *of* $\{x_i\}$ *is stationary.*

We conclude this section with the remark that the synchronization step in all the proposed methods in this section can be further modified if desired. In particular, we can search along the direction $x^i + \lambda(x^{i+1} - x^i)$, $\lambda \in R$, for a better point than x^{i+1} as the next iterate and replace x^{i+1} by this better point. All the convergence results remain valid because of the forcing function arguments used to establish them.

4. Conclusion and numerical results. We have given a number of parallel versions of fundamental convergence theorems for unconstrained minimization. These basic results enable k possible massively large, parallel processors to perform on portions of the gradient what one processor performs on the entire gradient in a serial algorithm. The direction choices in these theorems include many of the popular directions (gradient, quasi-Newton, Newton, conjugate gradient) and stepsizes (minimization, first stationary point, interval, Armijo). Note that each processor can apply direction and stepsize choices different from those of the other processors. A synchronization step is then used to obtain a strongly convex combination of the k points obtained by the k processors for the convex case, or alternatively the best of the k points or a better point can be taken as the next iterate for the convex as well as the nonconvex case.

Numerical implementations of parallel gradient distribution algorithms have been carried out in [1] and [6] on the Thinking Machines CM-5 multiprocessor. In these implementations, inexact quasi-Newton minimization was used in each parallel processor so as to satisfy (16). Each processor was allowed to take a number of steps before synchronization. The synchronization consisted of searching the affine hull of the points generated by the parallel processors as well as the current point. The problems solved in [1] consisted of real-world multicategory discrimination problems, formulated as unconstrained minimization of piecewise convex quadratic functions with Lipschitz continuous gradients. Problem size varied between 70 and 140 variables. For these multicategory discrimination problems, it is most efficient to use as many parallel processors as there are categories. This happened to be 7 for the problems tested. A standard measure of efficiency for parallel algorithms is the speedup efficiency, defined as

$$\text{speedup efficiency} = \frac{\text{time on 1 processor}}{(\text{time on } k \text{ processors}) * k}.$$

Thus, a speedup efficiency of 100% means that the time taken by one processor is cut exactly by a factor of k when k processors are employed. An efficiency of over 100% indicates that some of the parallel processors that are solving smaller subproblems have obtained very good points or that the affine hull generated by these points spans some very good points. For the multicategory discrimination problems, speedup efficiency was between 50% and 91%. For more details see [1].

In [6], 30 unconstrained problems from the publicly available CUTE [3] were tested. Among others, the parallel variable distribution algorithm version PVD0 was tested, which is equivalent to a parallel gradient distribution algorithm. Problems solved were between 100 and 1024 variables in size. These problems were solved on 2, 4, 8, 16, and 32 processors, with respective average speedup efficiencies of 129%, 122%, 77%, 44%, and 20%. These figures indicate that for problems of the size attempted, parallel gradient distribution is capable of producing a speedup equal to or better than 44% of the number of processors used for 16 or less processors. In order to exploit more fully a larger number of processors, larger problems need to be solved. We believe, however, that we have demonstrated that parallel gradient distribution can achieve speedups of the order of the processors employed and hence warrant further study and testing.

REFERENCES

[1] K. P. Bennett and O. L. Mangasarian, *Serial and parallel multicategory dicrimination*, SIAM J. Optim., 4 (1994), pp. 722–734.

[2] D. P. Bertsekas and J. N. Tsitsiklis, *Parallel and Distributed Computation*, Prentice–Hall, Englewood Cliffs, NJ, 1989.

[3] I. Bongartz, A. R. Conn, N. Gould, and Ph. L. Toint, *CUTE: Constrained and unconstrained testing environment*, Report 93/10, Publications du Départment de Mathématique, Facultés Universitaires de Namur, 1993.

[4] J. W. Daniel, *The Approximate Minimization of Functionals*, Prentice-Hall, Englewood Cliffs, NJ, 1971.

[5] J. E. Dennis and R. B. Schnabel, *Numerical Methods for Unconstrained Optimization and Nonlinear Equations*, Prentice-Hall, Englewood Cliffs, NJ, 1983.

[6] M. C. Ferris and O. L. Mangasarian, *Parallel variable distribution*, SIAM J. Optim., 4 (1994), pp. 815–832.

[7] E. S. Levitin and B. T. Polyak, *Constrained minimization methods*, Comput. Math. Math. Phys., 6 (1968), pp. 1–50. (Translated from Russian.)

[8] Z.-Q. Luo and P. Tseng, *On the convergence of the coordinate descent method for convex differentiable minimization*, J. Optim. Theory Appl., 72 (1992), pp. 7–35.

[9] O. L. Mangasarian and M. V. Solodov, *Serial and parallel backpropagation convergence via nonmonotone perturbed minimization*, Optimization Methods and Software, 4 (1994), pp. 103–116.

[10] J. M. Ortega and W. C. Rheinboldt, *Iterative Solution of Nonlinear Equations in Several Variables*, Academic Press, New York, 1970.

[11] E. Polak, *Computational Methods in Optimization*; *A Unified Approach*, Academic Press, New York, 1971.

[12] E. Polak and G. Ribière, *Note sur la convergence de méthodes de directions conjugées*, Revue Francaise Informatique et Recherche Opérationelle, 16-R1 (1969), pp. 35–43.

[13] B. T. Polyak, *The conjugate gradient method in extremal problems*, Comput. Math. Math. Phys., 9 (1969), pp. 94–112, 1969. (Translated from Russian.)

[14] ———, *Introduction to Optimization*, Optimization Software, Inc., Publications Division, New York, 1987.

[15] P. Tseng, *Dual ascent methods with strictly convex costs and linear constraints: A unified approach*, SIAM J. Control Optim., 28 (1990), pp. 214–242.

[16] J. Warga, *Minimizing certain convex functions*, J. SIAM, 11 (1963), pp. 588–593.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 33, No. 6, pp. 1926–1951, November 1995

© 1995 Society for Industrial and Applied Mathematics

016

THE CONTINUUM-ARMED BANDIT PROBLEM*

RAJEEV AGRAWAL[†]

Abstract. In this paper we consider the multiarmed bandit problem where the arms are chosen from a subset of the real line and the mean rewards are assumed to be a continuous function of the arms. The problem with an infinite number of arms is much more difficult than the usual one with a finite number of arms because the built-in learning task is now infinite dimensional. We devise a kernel estimator-based learning scheme for the mean reward as a function of the arms. Using this learning scheme, we construct a class of certainty equivalence control with forcing schemes and derive asymptotic upper bounds on their learning loss. To the best of our knowledge, these bounds are the strongest rates yet available. Moreover, they are stronger than the $o(n)$ required for optimality with respect to the average-cost-per-unit-time criterion.

Key words. bandit problems, controlled i.i.d. process, stochastic adaptive control, certainty equivalence with forcing, learning loss, continuous arms

AMS subject classifications. Primary, 93E35, 62G20, 62L05; Secondary, 60F15, 60F25.

1. Introduction.

In this paper we consider the multiarmed bandit problem where the arms are chosen from a subset of the real line and the mean rewards are assumed to be a continuous function of the arms. Prior work on the multiarmed bandit problem has dealt almost exclusively with only a finite number of independent arms. One of the early papers on this topic was by Robbins [26] who constructs a *consistent* (or optimal average-reward-per-unit-time) policy. More recently, the seminal work of Lai and Robbins [23], [22] addressed this problem with the stronger *learning loss* criterion (defined in (5)). They obtained asymptotic lower bounds on the learning loss and constructed *asymptotically efficient* schemes that achieved those bounds. Various extensions of the basic Lai and Robbins formulation have been obtained by Anantharam, Varaiya, and Walrand [6], [7]; by Agrawal, Hegde, and Teneketzis [2], [3]; and by Agrawal, Teneketzis, and Anantharam [5], [4]. In [5] the arms are allowed to be dependent, and the dependence is explicitly exploited to improve the performance. One of the few papers that does deal with an infinite set of arms is by Yakowitz and Lowe [28]. They consider only the *ϵ-learning loss* criterion (defined in (6)), and for this criterion they get a weaker rate than the one obtained in this paper (Corollary 5.4). However, unlike us, they do not make any continuity assumptions.

Note that the case of an infinite number of arms is much more difficult than the usual one of a finite number of arms because the built-in learning task is now infinite dimensional whereas previously it was only finite dimensional. In this paper we exploit the continuity of the mean reward as a function of the arms to devise a class of learning schemes based on kernel estimators. We obtain an upper bound on the almost sure and L^p uniform consistency rates for these estimators. These bounds strengthen the ones available in the nonparametric regression literature as detailed in §3, and may thus be of independent interest.

Subsequently, using the approach taken in [4], we construct a class of adaptive control schemes based on certainty equivalence control with forcing and derive asymptotic upper bounds on their learning loss. These bounds are not only much stronger

* Received by the editors September 16, 1992; accepted for publication (in revised form) July 11, 1994. This research was supported by National Science Foundation grant ECS-8919818.

[†] Department of Electrical and Computer Engineering, University of Wisconsin–Madison, Madison, Wisconsin 53706-1691 (agrawal@engr.wisc.edu).

than the $o(n)$ required for optimality with respect to the average-cost-per-unit-time criterion, but are also, to the best of our knowledge, the best rates available to date.

The rest of the paper is organized as follows: In §2 we give the precise problem formulation. In §3 we concentrate solely on the learning aspect of the problem. We construct a class of learning schemes and derive an upper bound on its rate of convergence (Corollary 3.4). In §4 we derive various limit laws for "moving averages" that are needed to obtain the rates of convergence in §3. Finally, in §5 we construct a class of adaptive control schemes based on the learning schemes of §3, and obtain upper bounds on their learning loss (Corollaries 5.2 and 5.4).

2. The problem. Consider a (memoryless) discrete-time stochastic system modeled by a controlled i.i.d. process, i.e.,

$$\begin{aligned} P(X_n \in B | U_1, X_1, \ldots, U_{n-1}, X_{n-1}, U_n = u) &= P(X_n \in B | U_n = u) \\ &= P(X_1 \in B | U_1 = u) \end{aligned} \tag{1}$$

where $\{U_n, X_n\}_{n=1}^{\infty}$ is the chronological sequence of controls and states. The states X_n take values in some arbitrary set $\mathcal{X}$, and the controls U_n are chosen from a bounded set $\mathcal{U} \subset \mathbb{R}$. In particular, we will assume that $\mathcal{U} = [\Delta, 1 - \Delta]$ for some $0 < \Delta < 1/2$; any arbitrary bounded subset of $\mathbb{R}$ can be handled easily by a slight modification. There is a one-step reward, $r(U_n, X_n)$, associated with each pair $(U_n, X_n), n \geq 1$, where $r : \mathcal{U} \times \mathcal{X} \to \mathbb{R}$. Let $m : \mathcal{U} \to \mathbb{R}$ be defined by

$$m(u) := E[r(U_n, X_n) | U_n = u] = E[r(u, X_1) | U_1 = u],$$

and let $W_n := r(U_n, X_n) - m(U_n)$. Then we can write

$$r(U_n, X_n) = m(U_n) + W_n,$$

where $E[W_n | U_n = u] = 0$ and

$$\begin{aligned} P(W_n \in B | U_1, W_1, \ldots, U_{n-1}, W_{n-1}, U_n = u) &= P(W_n \in B | U_n = u) \\ &= P(W_1 \in B | U_1 = u). \end{aligned} \tag{2}$$

Throughout the rest of the paper we shall assume that $\{U_n, W_n\}$ satisfy the following condition. There exist $\varsigma, s_0 > 0$ such that

$$E[\exp(sW_n) | U_n = u] \leq \exp(\varsigma^2 s^2 / 2) \quad \forall |s| \leq s_0, u \in \mathcal{U}. \tag{3}$$

In that case define

$$\sigma := \inf\{\varsigma > 0 : \text{ there exists } s_0 > 0 \text{ such that (3) holds}\}. \tag{4}$$

Then $E[|W_n|^2 | U_n = u] \leq \sigma^2$ for all $u \in \mathcal{U}$.

The problem is to design an adaptive control scheme $\gamma = \{\gamma_n\}_{n=1}^{\infty}$, i.e., $U_n = \gamma_n(U_1, X_1, \ldots, U_{n-1}, X_{n-1})$, so as to "maximize" the total reward

$$J_n^{\gamma} := \sum_{i=1}^{n} r(U_i, X_i) = \sum_{i=1}^{n} m(U_i) + \sum_{i=1}^{n} W_i\ ,$$

as $n \to \infty$. First note that

$$J_n^{\gamma} \leq nm^* + \sum_{i=1}^{n} W_i,$$

where $m^* := \sup_{u \in \mathcal{U}} m(u)$. Now if $m : \mathcal{U} \to \mathbb{R}$ were known, then for any constant $M > 0$, we could construct a scheme γ^M such that

$$J_n^{\gamma^M} \geq nm^* + \sum_{i=1}^{n} W_i - M.$$

In the absence of knowledge of m it is desirable to approach this performance as closely as possible. For this purpose define the *learning loss*

$$L_n := \sum_{i=1}^{n} m^* - m(U_i)\ . \tag{5}$$

Also define the *ϵ-learning loss*

$$L_n^{\epsilon} := \sum_{i=1}^{n} I\{m^* - m(U_i) > \epsilon\}\ . \tag{6}$$

Therefore, more precisely, the problem is to design adaptive control schemes for which the learning loss (or ϵ-learning loss) increases slowly regardless of the actual m. In this paper we will investigate almost sure and L^p $(p \geq 1)$ rates for the learning loss and L^1 rate for the ϵ-learning loss.

Throughout the rest of the paper we shall assume that $m : \mathcal{U} \to \mathbb{R}$ is uniformly locally Lipschitz with constant L $(0 \leq L < \infty)$, exponent α $(0 < \alpha \leq 1)$, and restriction δ $(\delta > 0)$, i.e.,

$$u, u' \in \mathcal{U}, |u' - u| \leq \delta \quad \Rightarrow \quad |m(u') - m(u)| \leq L|u' - u|^{\alpha}.$$

Let $ulL(\alpha, L, \delta)$ denote the class of all such functions.

3. The learning scheme. In this section we concentrate solely on the learning aspect of the problem as a first step toward the construction of adaptive control schemes. More precisely, we are interested in choosing the controls $\{U_n\}$ and in constructing the estimates $\{U_n^*\}$ based on the observations $\{m(U_n) + W_n\}$ made at those points, so that $m^* - m(U_n^*)$ converges to 0 as rapidly as possible. Note that the estimates U_n^* need not be the same as the control values U_n. In this section we construct a class of learning schemes and obtain bounds on their rates of convergence. These bounds are important because they determine precisely how the learning schemes can be used to design good adaptive control schemes.

First note that the only assumption we have made on the function m is that it is uniformly locally Lipschitz. Since we do not make any unimodality assumptions we will need a "global search" strategy. Moreover, since the function m may not be differentiable, we cannot use algorithms that rely on the estimation of the gradient. For both these reasons, the Kiefer–Wolfowitz-type (K–W-type) stochastic approximation algorithm is not appropriate for the problem at hand. Even with the unimodality assumption, we would require additional differentiability conditions (existence, boundedness, continuity of the second derivative) in order to get any rate of convergence results for the K–W-type algorithms (see Fabian [12], Nevel'son and Has'minskii [25], Kushner and Clark [21]). Furthermore, the rates of convergence obtained for these algorithms are slower than the ones obtained for the algorithm used in this paper. This is possibly due to the fact that we are interested in constructing the estimates U_n^* so as to minimize the difference from the maximum, $m^* - m(U_n^*)$, rather than to minimize the actual distance $|u^* - U_n^*|$ from the point of maximum u^*. In fact, a point of maximum need not even exist for the problem considered in this paper.

Recently, several researchers have studied a variant of the K–W-type gradient algorithm, which incorporates the global search aspects of simulated annealing algorithms by adding a slowly decreasing noise term. For this algorithm, convergence to the set of global maxima can be established without any unimodality assumption, but with suitable differentiability conditions (see [15], [20], [8], [14]). However, to the best of our knowledge, no results are available on the rate of convergence of these algorithms.

The learning schemes constructed in this section are based on the approach used by Devroye [11]. We first obtain a "uniformly good" estimate $\hat{m}_n$ of the function m, and then use the point of maximum of $\hat{m}_n$ as the estimate U_n^*. We use a nearly equispaced control (*design*) sequence $\{u_n\}$ defined below in (7)–(8). This is a natural choice that corresponds to a progressively finer sampling of $[0, 1]$ along the dyadic rationals. Note that the equispaced design scheme itself is not a *progressive* design scheme, i.e., the set of design points at stage n is not a subset of the design points at stage $n + 1$. The nearly equispaced design scheme described below is the best progressive approximation to the equispaced design scheme in the sense that it visits the equispaced design at stages $n = 2^m, m = 0, 1, \ldots,$ which occur earlier than the successive visits of any other progressive design scheme.[1] Based on the observations obtained at these points, we estimate the function m by means of a kernel estimator defined below in (9)–(11). In Theorem 3.1 we obtain an upper bound (with an explicit constant) on the almost sure and L^p $(p \geq 1)$ uniform consistency rates of this estimation scheme. In Theorem 3.2 we show that the above upper bound is also a lower bound on the in-probability uniform consistency rates of this estimation scheme for the i.i.d. noise case. This shows us that the rate and associated constant obtained in Theorem 3.1 cannot be improved in general.

The problem of estimating the function $m : \mathcal{U} \to \mathbb{R}$ on the basis of "noisy" measurements $\{m(U_n) + W_n\}_{n=1}^{\infty}$ taken at a sequence of points $\{U_n\}_{n=1}^{\infty} \in \mathcal{U}$ has been extensively investigated in the nonparametric regression literature in statistics. While the almost sure rate itself is well known in the nonparametric regression literature, the associated constant is not (see Stone [27]; Mack and Silverman [24]; Härdle and Luckhaus [18]; Härdle, Janssen and Serfling [17]). Also see Härdle [16, Chap. 4, pp. 89–98] for excellent coverage of results available to date on this nonparametric regression problem. The identification of a sharp constant associated with this rate is an important and challenging problem, as evidenced by the work of Fabian [13], who obtains a constant associated with the in-probability rate for a nonprogressive design scheme with a piecewise polynomial estimator. In fact, Fabian comments that specifying the constants in the order of convergence seems difficult for the estimates considered up to now. One of the key contributions of this section is to provide precisely such a constant for a kernel estimator, which is one of the popular estimators considered in the nonparametric regression literature. As explained in §4, we crucially exploit the structure of the progressive nearly equispaced scheme to obtain the sharp constant associated with the almost sure rate. The constant corresponding to the design/estimation scheme considered in this paper is better than the constant obtained

[1] There are other progressive design schemes which visit the equispaced design scheme at stages $n = 2^m, m = 0, 1, \ldots.$ The design scheme described in this paper is the "best" progressive approximation to the equispaced design scheme in possibly a much stronger sense.

by Fabian [13] for his design/estimation scheme (see the second remark following Corollary 3.3). Another contribution of the results obtained in this section is to establish that the same rates also hold both almost surely and in L^p, whereas all of the papers cited above obtain the rates only in probability or almost surely.

Note that the uniform error between the estimate and the true function m can easily be decomposed into the sum of a bias term and a variance term (see (14) below). The bias term is controlled by the continuity of the function m, while the variance term is controlled by the "noise" in the measurements and by the choice of the design scheme. The variance term is essentially a "moving average." Our main result in Theorem 3.1 on the rate of convergence makes use of some fundamental limit theorems on these moving averages that are developed in §4.

We use the nearly equispaced design sequence $\{u_n\}_{n=1}^{\infty} \in \mathcal{U} \not\subset [0,1]$ described below. For $n = 1, 2, \ldots$, consider the binary representation of $n-1$:

$$n - 1 = \ldots b_3 b_2 b_1. \tag{7}$$

First, choose the points $\tilde{u}_n \in [0,1]$ to be dyadic rationals with the binary representation:

$$\tilde{u}_n = 0.b_1 b_2 b_3 \ldots. \tag{8}$$

Note that this is the best progressive approximation to the equispaced design scheme on $[0,1]$. The actual design points, $u_n \in \mathcal{U} = [\Delta, 1-\Delta] \subset [0,1]$, are obtained by projecting $\tilde{u}_n$ onto the control set $\mathcal{U}$. Thus

$$u_n := (\tilde{u}_n \wedge (1-\Delta)) \vee \Delta.$$

The reason for choosing the actual design points in this manner is to ensure that asymptotically we get enough observations close to the two boundaries. If we had chosen $\mathcal{U} = [0,1]$ and $u_n = \tilde{u}_n$, then we would have gotten only half as many observations in a window centered at one of the boundary points as we would in the interior. Note that under any deterministic design scheme (such as the one above), $\{W_n\}$ are independent by (2).

We use a window estimator which is a special case of the more general class of Nadaraya–Watson kernel estimators to estimate m. Thus,

$$\hat{m}_n(u) = \frac{\sum_{i=1}^{n} K_{h_n}(u - \tilde{u}_i)(m(u_i) + W_i)}{\sum_{i=1}^{n} K_{h_n}(u - \tilde{u}_i)}, \tag{9}$$

where

$$K_{h_n}(u) = h_n^{-1} K(u/h_n), \tag{10}$$

$$K(u) = I\{|u| \le 1/2\} \tag{11}$$

is the window kernel, and $\{h_n\}$ is a sequence of bandwidths to be specified later. Finally, we choose the estimate U_n^* by

$$U_n^* := \operatorname*{argmax}_{u \in \mathcal{U}} \hat{m}_n(u).$$

Note that $\hat{m}_n$ is a piecewise constant function with at most a finite number of discontinuities. Hence, there exists at least one such argmax. Ties may be resolved in some fixed but arbitrary measurable manner. This completes the description of our class of learning schemes.

In the remaining part of this section we obtain bounds on the rate of convergence of the above scheme: in Theorem 3.1 and Corollary 3.3 for the estimates $\hat{m}_n$, and in Corollary 3.4 for U_n^*. Let

$$d_\infty(\hat{m}, m) = \sup_{u \in \mathcal{U}} |\hat{m}(u) - m(u)|$$

be the uniform metric. The following theorem obtains (1) almost sure and L^p rates of convergence for $\sup_{m \in ulL(\alpha,L,\delta)} d_\infty(\hat{m}_n, m)$, along with a sharp constant, and (2) a large deviations-type result for $P(\sup_{m \in ulL(\alpha,L,\delta)} d_\infty(\hat{m}_n, m) > \epsilon)$ for any $\epsilon > 0$.

THEOREM 3.1. *For the above window estimator with any bandwidth sequence, h_n, such that* (i) *h_n is nonincreasing,* (ii) *nh_n is nondecreasing, and* (iii) *$A'n^{a'} \le nh_n \le An^a$ for some $0 < a' \le a < 1$, $A', A > 0$, we have*

$$\overline{\lim_{n\to\infty}} \frac{\sup_{m \in ulL(\alpha,L,\delta)} d_\infty(\hat{m}_n, m)}{\frac{L}{\alpha+1}\left(\frac{h_n}{2}\right)^\alpha + \left(\frac{2\log(1/h_n)}{nh_n}\right)^{1/2}\sigma} \le 1 \quad \textit{a.s. and in } L^p, p \ge 1. \tag{12}$$

Also, for the above window estimator with any bandwidth sequence, h_n, such that (i) *$h_n \to 0$ as $n \to \infty$, and* (ii) *$nh_n \ge A'n^{a'}$ for some $0 < a' < 1$, $A' > 0$, we have*

$$\overline{\lim_{\epsilon\searrow 0}} \frac{1}{\epsilon^2} \overline{\lim_{n\to\infty}} \frac{1}{nh_n} \log P\left(\sup_{m \in ulL(\alpha,L,\delta)} d_\infty(\hat{m}_n, m) > \epsilon\right) \le -\frac{1}{8\sigma^2}. \tag{13}$$

Remark. The above theorem can easily be adapted to handle $\mathcal{U} = [a,b]$ for any $-\infty < a < b < \infty$. For each $m : [a,b] \to \mathbb{R}$, $m \in ulL(\alpha, L, \delta)$, simply consider the shifted and rescaled function $\tilde{m} : [\Delta, 1-\Delta] \to \mathbb{R}$ defined by $\tilde{m}(u) = m(a + (u - \Delta)(b-a)(1-2\Delta)^{-1})$. Note that

$$m \in ulL(\alpha, L, \delta) \Leftrightarrow \tilde{m} \in ulL\left(\alpha, L\left(\frac{b-a}{1-2\Delta}\right)^\alpha, \delta\left(\frac{1-2\Delta}{b-a}\right)\right).$$

Next observe that (12) holds for any $\delta > 0$ and $0 < \Delta < 1/2$; however, it does explicitly depend on α and L. Hence, in light of the above discussion, it would be desirable to choose Δ arbitrarily small.

Proof. Since $nh_n \ge A'n^{a'}$ for some $0 < a' < 1$, $A' > 0$, we can pick n_0 such that for all $n \ge n_0$, $\sum_{i=1}^n I\{|u - \tilde{u}_i| \le h_n/2\} > 0 \quad \forall u \in \mathcal{U}$. Then, for all $n \ge n_0$,

$$\begin{aligned}\hat{m}_n(u) &= \frac{\sum_{i=1}^n K_{h_n}(u-\tilde{u}_i)(m(u_i)+W_i)}{\sum_{i=1}^n K_{h_n}(u-\tilde{u}_i)} \\ &= \frac{\sum_{i=1}^n h_n^{-1} I\{|u-\tilde{u}_i| \le h_n/2\}(m(u_i)+W_i)}{\sum_{i=1}^n h_n^{-1} I\{|u-\tilde{u}_i| \le h_n/2\}} \\ &= \frac{\sum_{i=1}^n I\{|u-\tilde{u}_i| \le h_n/2\}(m(u_i)+W_i)}{\sum_{i=1}^n I\{|u-\tilde{u}_i| \le h_n/2\}}.\end{aligned}$$

Thus,

$$\begin{aligned}|\hat{m}_n(u) - m(u)| \le \frac{\sum_{i=1}^n I\{|u-\tilde{u}_i| \le h_n/2\}|m(u_i) - m(u)|}{\sum_{i=1}^n I\{|u-\tilde{u}_i| \le h_n/2\}} \\ + \frac{|\sum_{i=1}^n I\{|u-\tilde{u}_i| \le h_n/2\}W_i|}{\sum_{i=1}^n I\{|u-\tilde{u}_i| \le h_n/2\}}.\end{aligned} \tag{14}$$

Since $h_n \to 0$ as $n \to \infty$, we can pick $n_1 \geq n_0$ so that $h_n/2 \leq \min\{\delta, \Delta\}$ for all $n \geq n_1$. Then, for any $u \in \mathcal{U} = [\Delta, 1-\Delta]$ and $n \geq n_1$, we have

$$b_n := nh_n - \lfloor \log_2 n \rfloor - 1 \leq \sum_{i=1}^{n} I\{|u - \tilde{u}_i| \leq h_n/2\} \leq nh_n + \lfloor \log_2 n \rfloor + 1 =: a_n. \tag{15}$$

Thus, by the condition $nh_n \geq A'n^{a'}$ for some $0 < a' < 1$, $A' > 0$, it follows that $\sum_{i=1}^{n} I\{|u - \tilde{u}_i| \leq h_n/2\} \approx a_n \approx b_n \approx nh_n$, where by $x_n \approx y_n$ we mean that $x_n/y_n \to 1$ as $n \to \infty$.

Also, for any $m \in ulL(\alpha, L, \delta)$ and $n \geq n_1$, we have

$$\begin{aligned} &\sum_{i=1}^{n} I\{|u - \tilde{u}_i| \leq h_n/2\}|m(u_i) - m(u)| \\ &\qquad \leq \sum_{i=1}^{n} I\{|u - \tilde{u}_i| \leq h_n/2\} L|\tilde{u}_i - u|^{\alpha} \\ &\qquad \leq \frac{L}{\alpha+1}\left(\frac{h_n}{2}\right)^{\alpha} [nh_n + 2(\alpha+1)(\lfloor \log_2 n \rfloor + 1)], \end{aligned} \tag{16}$$

where the second inequality is established in the appendix. Combining this with (15), we get for any $m \in ulL(\alpha, L, \delta)$ and $n \geq n_1$,

$$\begin{aligned} \frac{\sum_{i=1}^{n} I\{|u - \tilde{u}_i| \leq h_n/2\}|m(u_i) - m(u)|}{\sum_{i=1}^{n} I\{|u - \tilde{u}_i| \leq h_n/2\}} &\leq \frac{L}{\alpha+1}\left(\frac{h_n}{2}\right)^{\alpha} \frac{1 + 2(\alpha+1)\frac{\lfloor \log_2 n \rfloor + 1}{nh_n}}{1 - \frac{\lfloor \log_2 n \rfloor + 1}{nh_n}} \\ &= \frac{L}{\alpha+1}\left(\frac{h_n}{2}\right)^{\alpha} (1 + \gamma_n) \quad \text{(say)} \\ &\approx \frac{L}{\alpha+1}\left(\frac{h_n}{2}\right)^{\alpha}. \end{aligned} \tag{17}$$

Let $\{\tilde{W}_i^n\}_{i=1}^n$ be a permutation of $\{W_i\}_{i=1}^n$ arranged in order of $\{\tilde{u}_i\}_{i=1}^n$. Define the partial sums of $\{\tilde{W}_i^n\}_{i=1}^n$ as follows:

$$\tilde{S}_{j,k}^n := \sum_{i=j+1}^{k} \tilde{W}_i^n. \tag{18}$$

Also let

$$\tilde{B}_{n,a_n} := \max_{\substack{0 \leq j \leq k \leq n \\ k-j \leq a_n}} |\tilde{S}_{j,k}^n|. \tag{19}$$

Then

$$\left|\sum_{i=1}^{n} I\{|u - \tilde{u}_i| \leq h_n/2\} W_i\right| \leq \tilde{B}_{n,a_n}. \tag{20}$$

Putting together (14)–(20) we get that for all $n \geq n_1$,

$$\begin{aligned} \sup_{m \in ulL(\alpha,L,\delta)} d_{\infty}(\hat{m}_n, m) &= \sup_{m \in ulL(\alpha,L,\delta)} \sup_{u \in \mathcal{U}} |\hat{m}_n(u) - m(u)| \\ &\leq \frac{L}{\alpha+1}\left(\frac{h_n}{2}\right)^{\alpha} (1 + \gamma_n) + \frac{\tilde{B}_{n,a_n}}{b_n}. \end{aligned} \tag{21}$$

Asymptotic bounds on the "moving average" $\tilde{B}_{n,a_n}$ are obtained in §4. In particular, by Theorems 4.5 and 4.7 we know that $\overline{\lim}_n \tilde{B}_{n,a_n}/\beta_n \leq \sigma$ a.s. and in L^p, $p \geq 1$, respectively, where $\beta_n = (2a_n(\log(n/a_n) + \log\log n))^{1/2}$, provided a_n satisfies some conditions. It is straightforward to check that $a_n = nh_n + \lfloor \log_2 n \rfloor + 1$ satisfies the conditions of Theorems 4.5 and 4.7 whenever h_n satisfies the conditions of this theorem. Finally, we have

$$\begin{aligned}\beta_n/b_n &= \frac{(2a_n(\log(n/a_n) + \log\log n))^{1/2}}{nh_n - \lfloor \log_2 n \rfloor - 1}\\ &\approx \frac{(2nh_n(\log(n/nh_n) + \log\log n))^{1/2}}{nh_n}\\ &\approx \left(\frac{2\log(1/h_n)}{nh_n}\right)^{1/2}. \qquad (22)\end{aligned}$$

Combining all of the above, we get (12).

We will now prove the remaining part of the theorem. Let $0 < \eta < \epsilon$ be fixed but arbitrary. Since the bandwidth sequence $h_n \to 0$ and $\gamma_n \to 0$ as $n \to 0$, we can pick $n_2 \geq n_1$ so that $h_n/2 \leq \min\{\delta, \Delta, (\eta(\alpha+1)/2L(1+\gamma_n))^{1/\alpha}\}$, for all $n \geq n_2$. Also, by the condition that $nh_n \geq A'n^{a'}$ for some $0 < a' < 1$, $A' > 0$, it follows that $b_n/a_n \to 1$ as $n \to \infty$. Thus, we can pick $n_3 \geq n_2$ so that we have $(\epsilon - \eta/2)b_n \geq (\epsilon - \eta)a_n$ for all $n \geq n_3$. Then, from (21), we have for all $n \geq n_3$

$$\begin{aligned}P\left(\sup_{m \in ulL(\alpha,L,\delta)} d_\infty(\hat{m}_n, m) > \epsilon\right) &\leq P\left(\frac{L}{\alpha+1}(h_n/2)^\alpha + \tilde{B}_{n,a_n}/b_n > \epsilon\right)\\ &\leq P(\tilde{B}_{n,a_n}/b_n > \epsilon - \eta/2)\\ &\leq P(\tilde{B}_{n,a_n}/a_n > \epsilon - \eta). \qquad (23)\end{aligned}$$

Substituting n, m, p, t, s, a in Lemma 4.1 by $n, a_n, \lfloor nd/a_n \rfloor + 1, (\epsilon - 2\eta), \eta, (\epsilon - 2\eta - 2\rho)/(\epsilon - 2\eta)$, respectively, we have

$$\begin{aligned}P(\tilde{B}_{n,a_n}/a_n > \epsilon - \eta) \leq \frac{1}{1-c}\left(\frac{nd}{a_n} + 1\right)\Big[&\max_{0 \leq j \leq n} P(|\tilde{S}^n_{j,j+a_n}|/a_n > (\epsilon - 2\eta - 2\rho))\\ &+ \max_{0 \leq j \leq n} P(|\tilde{S}^n_{j,j+\lfloor a_n/d \rfloor + 1}|/a_n > \rho)\Big] \qquad (24)\end{aligned}$$

where c is such that

$$\max_{0 \leq j \leq n} \max_{1 \leq k \leq a_n + 1} P(|\tilde{S}^n_{j,j+k}|/a_n > \eta/4) \leq c < 1. \qquad (25)$$

By Chebyshev's inequality we have

$$\max_{0 \leq j \leq n} \max_{1 \leq k \leq a_n + 1} P(|\tilde{S}^n_{j,j+k}|/a_n > \eta/4) \leq \frac{\sigma^2(a_n+1)}{(a_n\eta/4)^2}.$$

By the assumptions on h_n, $a_n = nh_n + \lfloor \log_2 n \rfloor + 1 \to \infty$ as $n \to \infty$. Therefore, given any $0 < c < 1$, there exists $n_4 \geq n_3$ such that (25) holds for all $n \geq n_4$. We now apply Lemma 4.2 to upper bound the right-hand side (RHS) of (24). Given any $\varsigma > \sigma$, let s_0 be such that (30) holds. Let $\epsilon \leq \varsigma^2 s_0$. Then by Lemma 4.2 we have

$$P(|\tilde{S}^n_{j,j+a_n}|/a_n > (\epsilon - 2\eta - 2\rho)) \leq 2\exp\left(-\frac{(\epsilon - 2\eta - 2\rho)^2 a_n^2}{2a_n\varsigma^2}\right).$$

Similarly, if $d\rho \le \varsigma^2 s_0$, then by Lemma 4.2 we have

$$P(|\tilde{S}^n_{j,j+\lfloor a_n/d\rfloor+1}|/a_n > \rho) \le 2\exp\left(-\frac{d\rho^2 a_n^2}{2a_n\varsigma^2}\right).$$

Thus, by choosing $d = 4$ and $\rho = \epsilon/4$, we get that for all $\epsilon \le \varsigma^2 s_0$, $\eta < \epsilon/4$, and $n \ge n_4$,

$$P(\tilde{B}_{n,a_n}/a_n > \epsilon - \eta) \le \frac{1}{1-c}\left(\frac{nd}{a_n}+1\right)4\exp\left(-\frac{((\epsilon/2)-2\eta)^2 a_n}{2\varsigma^2}\right).$$

Thus, by (23) and the above, it follows that

$$\begin{aligned}\overline{\lim_{n\to\infty}}\frac{1}{a_n}\log P\left(\sup_{m\in ulL(\alpha,L,\delta)} d_\infty(\hat{m}_n,m) > \epsilon\right) &\le \overline{\lim_{n\to\infty}}\frac{\log(1/h_n)}{nh_n+\lfloor\log_2 n\rfloor+1} - \frac{((\epsilon/2)-2\eta)^2}{2\varsigma^2}\\ &= -\frac{((\epsilon/2)-2\eta)^2}{2\varsigma^2}.\end{aligned}$$

Note that the last equality follows from the assumptions on h_n. The left-hand side (LHS) above does not depend on η. By letting $\eta \to 0$, we get for all $\epsilon \le \varsigma^2 s_0$,

$$\overline{\lim_{n\to\infty}}\frac{1}{a_n}\log P\left(\sup_{m\in ulL(\alpha,L,\delta)} d_\infty(\hat{m}_n,m) > \epsilon\right) \le -\frac{\epsilon^2}{8\varsigma^2}.$$

Now dividing by ϵ^2, taking limits as $\epsilon \to 0$, and finally letting $\varsigma \to \sigma$, we obtain (13). □

Below, we show that the rate and the associated constant identified in (12) in the previous theorem are the best possible ones.

THEOREM 3.2. *If, in addition to the conditions already imposed on* $\{W_n\}$, *we assume that*

$$P(W_n \in B|U_n = u') = P(W_1 \in B|U_1 = u), \quad \forall B,\ u', u,\ n,$$

then we can also show that (12) *holds with equality in probability, i.e.,*

$$\lim_{n\to\infty} P\left(\frac{\sup_{m\in ulL(\alpha,L,\delta)} d_\infty(\hat{m}_n,m)}{\frac{L}{\alpha+1}(\frac{h_n}{2})^\alpha + (\frac{2\log(1/h_n)}{nh_n})^{1/2}\sigma} < 1-\epsilon\right) = 0 \quad \forall\epsilon > 0.$$

Proof. This follows because by choosing $m \in ulL(\alpha,L,\delta)$ to be given by $m(u) = \pm|u-c|^\alpha$ for $c \in [\Delta', 1-\Delta']$ with $\Delta < \Delta' < 1/2$, one can show that there exists an n_0' such that for all $n \ge n_0'$,

$$\begin{aligned}\sup_{m\in ulL(\alpha,L,\delta)} d_\infty(\hat{m}_n,m) \ge \sup_{c\in[\Delta',1-\Delta']}&\left[\frac{\sum_{i=1}^n I\{|c-\tilde{u}_i| \le h_n/2\}L|\tilde{u}_i - c|^\alpha}{\sum_{i=1}^n I\{|c-\tilde{u}_i| \le h_n/2\}}\right.\\ &\left.+\frac{|\sum_{i=1}^n I\{|c-\tilde{u}_i| \le h_n/2\}W_i|}{\sum_{i=1}^n I\{|c-\tilde{u}_i| \le h_n/2\}}\right].\end{aligned}\tag{26}$$

Corresponding to the upper bound on the first term on the RHS used earlier in (16), we have the following lower bound which is also established in the appendix. For all $n \ge 1$ and $u \in \mathcal{U}$,

$$\sum_{i=1}^n I\{|u-\tilde{u}_i| \le h_n/2\}L|\tilde{u}_i - u|^\alpha \ge \frac{L}{\alpha+1}\left(\frac{h_n}{2}\right)^\alpha [nh_n - 2(\alpha+1)(\lfloor\log_2 n\rfloor+1)].\tag{27}$$

Using (27) and (15) in (26) we get for all $n \geq n_0'$,

$$\begin{aligned}\sup_{m\in ulL(\alpha,L,\delta)} d_\infty(\hat{m}_n, m) \geq \frac{L}{\alpha+1}\left(\frac{h_n}{2}\right)^\alpha \frac{1-2(\alpha+1)\frac{\lfloor \log_2 n\rfloor+1}{nh_n}}{1+\frac{\lfloor \log_2 n\rfloor+1}{nh_n}} \\ + \max_{\Delta' n\leq j\leq(1-\Delta')n} \frac{|\tilde{S}^n_{j,j+a_{n,j}}|}{\beta_n}\frac{\beta_n}{a_n} \\ =: r_n + s_n t_n,\end{aligned}$$

where $b_n \leq a_{n,j} \leq a_n$ for all $1 \leq j \leq n$, $n = 1, 2, \ldots$. As shown earlier in (17) and (22), for any $\epsilon > 0$ there exists an $n_1' \geq n_0'$ such that for all $n \geq n_1'$,

$$r_n \geq (1-\epsilon)\frac{L}{\alpha+1}\left(\frac{h_n}{2}\right)^\alpha$$

and

$$t_n \geq (1-\epsilon)^{1/2}\left(\frac{2\log(1/h_n)}{nh_n}\right)^{1/2}.$$

Thus, for all $n \geq n_1'$ we get

$$\begin{aligned}&P\left(\frac{\sup_{m\in ulL(\alpha,L,\delta)} d_\infty(\hat{m}_n, m)}{\frac{L}{\alpha+1}(\frac{h_n}{2})^\alpha + (\frac{2\log(1/h_n)}{nh_n})^{1/2}\sigma} < 1-\epsilon\right) \\ &\leq P\left(r_n + s_n t_n < (1-\epsilon)\left(\frac{L}{\alpha+1}\left(\frac{h_n}{2}\right)^\alpha + \left(\frac{2\log(1/h_n)}{nh_n}\right)^{1/2}\sigma\right)\right) \\ &\leq P(s_n < (1-\epsilon)^{1/2}\sigma)\end{aligned}$$

which goes to 0 as $n \to \infty$ by Theorem 4.3. Note that the collection of random variables $\{V_i^n\}$ obtained by setting $\{V_i^n\}_{i=1}^n = \{\tilde{W}_i^n\}_{i=1}^n$ for $n = 1, 2, \ldots$ satisfies conditions (C1)–(C4), but not (C5) (see §4). However, since the RHS above does not depend on (C5), we can still apply Theorem 4.3, whereas we would need condition (C5) to obtain the corresponding almost sure result from Theorem 4.3. □

COROLLARY 3.3. *For the above window estimator with the bandwidth sequence* $h_n = h(\log n/n)^{1/(2\alpha+1)}$, $h > 0$, *we have*

$$\overline{\lim_{n\to\infty}} \frac{\sup_{m\in ulL(\alpha,L,\delta)} d_\infty(\hat{m}_n, m)}{(\frac{\log n}{n})^{\alpha/(2\alpha+1)}} \leq c(\alpha, L, \sigma, h) \quad \textit{a.s. and in } L^p, p \geq 1, \tag{28}$$

where

$$c(\alpha, L, \sigma, h) := \frac{L}{\alpha+1}\left(\frac{h}{2}\right)^\alpha + \left(\frac{2}{h(2\alpha+1)}\right)^{1/2}\sigma. \tag{29}$$

Proof. First note that the above choice of h_n satisfies the conditions of Theorem 3.1. Moreover,

$$\frac{L}{\alpha+1}\left(\frac{h_n}{2}\right)^\alpha + \left(\frac{2\log(1/h_n)}{nh_n}\right)^{1/2}\sigma$$

$$
\begin{aligned}
&= \frac{L}{\alpha+1}\left(\frac{h}{2}\right)^{\alpha}\left(\frac{\log n}{n}\right)^{\alpha/(2\alpha+1)} \\
&\quad + \left[\frac{-2\log h + \frac{2}{(2\alpha+1)}(\log n - \log\log n)}{hn^{2\alpha/(2\alpha+1)}(\log n)^{1/(2\alpha+1)}}\right]^{1/2}\sigma \\
&\approx \left[\frac{L}{\alpha+1}\left(\frac{h}{2}\right)^{\alpha} + \sigma\left(\frac{2}{h(2\alpha+1)}\right)^{1/2}\right]\left(\frac{\log n}{n}\right)^{\alpha/(2\alpha+1)}. \qquad \square
\end{aligned}
$$

Remark. Note that the bandwidth sequence $\{h_n\}$ chosen above minimizes the rate of convergence given by (12). This is so because if $h_n/(\log n/n)^{1/(2\alpha+1)} \to 0$ or ∞, then

$$
\left[\frac{L}{\alpha+1}\left(\frac{h_n}{2}\right)^{\alpha} + \left(\frac{2\log(1/h_n)}{nh_n}\right)^{1/2}\sigma\right]\left(\frac{\log n}{n}\right)^{\alpha/(2\alpha+1)} \to \infty.
$$

Moreover, the constant $c(\alpha, L, \sigma, h)$ associated with the above rate is minimized by choosing

$$
h^* = 2(\sigma(\alpha+1)/2\alpha L(2\alpha+1)^{1/2})^{2/(2\alpha+1)}.
$$

The minimum value of the above constant, say $c^*(\alpha, L, \sigma)$, is given by

$$
c^*(\alpha, L, \sigma) = \left[\frac{L}{\alpha+1}\sigma^{2\alpha}(2\alpha+1)^{-\alpha}\right]^{1/(2\alpha+1)} [(2\alpha)^{-2\alpha/(2\alpha+1)} + (2\alpha)^{1/(2\alpha+1)}].
$$

Remark. The above constant can be compared with the one obtained by Fabian [13]. Note that while we consider uniformly locally Lipschitz functions with exponent $0 < \alpha \leq 1$, [13] considers functions with bounded rth derivative, $r \geq 1$. Thus $\alpha = r = 1$ is the only case in common between the two papers. In particular, for $\alpha = 1$ and $\sigma = 1$, our constant reduces to

$$
c^*(1, L, 1) = L^{1/3}\frac{(1/4)^{1/3} + 2^{1/3}}{6^{1/3}}.
$$

In comparison, the optimal constant in [13, Thm. 5.4, p. 1359] for the above case ($r = 1, s = 2$) is greater by a factor of $\pi^{1/3}$. Also note that while our results hold both almost surely and in $L^p, p \geq 1$, those of [13] hold only in probability.

The following corollary establishes the same rates for the sequence U_n^*.

COROLLARY 3.4. *For the learning scheme constructed above with the bandwidth sequence $h_n = h(\log n/n)^{1/(2\alpha+1)}$, $h > 0$, we have*

$$
\overline{\lim_{n\to\infty}} \frac{\sup_{m\in ulL(\alpha,L,\delta)}(m^* - m(U_n^*))}{(\frac{\log n}{n})^{\alpha/(2\alpha+1)}} \leq 2c(\alpha, L, \sigma, h) \quad \textit{a.s. and in } L^p, p \geq 1,
$$

where $c(\alpha, L, \sigma, h)$ is given by (29). *Also, for the the same learning scheme with any bandwidth sequence, h_n, such that* (i) *$h_n \to 0$ as $n \to \infty$, and* (ii) *$nh_n \geq A'n^{a'}$ for some $0 < a' < 1$, $A' > 0$, we have*

$$
\overline{\lim_{n\to\infty}} \frac{1}{nh_n}\log P\left(\sup_{m\in ulL(\alpha,L,\delta)}(m^* - m(U_n^*)) > \epsilon\right) \leq -\Gamma(\epsilon/2)
$$

for all $\epsilon > 0$, for some $\Gamma(\epsilon) > 0$.

Proof. Both parts of the corollary follow immediately from Theorem 3.1, Corollary 3.3, and the following observation. Since, $\hat{m}_n(U_n^*) = \max_{u \in \mathcal{U}} \hat{m}_n(u)$,

$$m(u) - m(U_n^*) \leq \hat{m}_n(u) + d_\infty(\hat{m}_n, m) - \hat{m}_n(U_n^*) + d_\infty(\hat{m}_n, m) \leq 2d_\infty(\hat{m}_n, m).$$

Taking $\sup_{u \in \mathcal{U}}$ on both sides we get $m^* - m(U_n^*) \leq 2d_\infty(\hat{m}_n, m)$. □

4. Limit laws for moving averages. In this section we obtain almost sure and L^p limit theorems for the moving averages that comprise the variance term in the analysis of the kernel estimator considered in §3. If the noise were i.i.d. and the design scheme were equispaced (see conditions (C4) and (C5) below), then exact almost sure limit theorems are available in the literature for these moving averages (see Theorem 4.3). These results can be viewed as generalizations of the law of iterated logarithm. In light of the paragraph following Theorem 4.3, the weaker in-probability result also holds for the the i.i.d. noise case with the progressive nearly equispaced design scheme employed in this paper. This identifies the desired rate and constant. However, in order to get the stronger almost sure results for the progressive nearly equispaced design scheme, we have to work considerably harder. We first obtain a coarser constant in Theorem 4.4 and then recover the desired constant in Theorem 4.5, by applying that result on an appropriately chosen subsequence. This subsequence argument relies heavily on the exact structure of the nearly equispaced design scheme. Thus, the specific design scheme plays a crucial role in the analysis, if we desire to get a sharp constant.

It is also worth noting that the techniques used here to obtain the abovementioned limit theorems parallel those of de Acosta and Kuelbs [10], by relying on classical probability inequalities rather than on the strong approximations approach of Csörgő and Révész [9]. This allows us to obtain results for $\{W_i\}$ which are independent but not necessarily identically distributed. Also, it gives us a way of extending these results to higher dimensions (obtained in a forthcoming paper [1]), for which the strong approximation results are not yet available, to the best of our knowledge (see Mack and Silverman [24]).

In Theorem 4.7 we establish the same rates (with the same constant) in L^p. This is done by means of a device due to Hoffman-Jørgensen [19]. The proof of Theorem 4.7 can be used more generally for establishing L^p counterparts of in-probability rates.

Throughout this section we will consider a collection of random variables $\{V_i^n : n = 1, 2, \ldots; i = 1, 2, \ldots, n\}$ satisfying some of the following conditions.

(C1) $E[V_i^n] = 0$ for all $n = 1, 2, \ldots; i = 1, 2, \ldots, n$.

(C2) $\{V_i^n\}_{i=1}^n$ is an independent collection for each $n = 1, 2, \ldots$.

(C3) There exist $\varsigma, s_0 > 0$ such that

$$E[\exp(sV_i^n)] \leq \exp(\varsigma^2 s^2/2) \quad \forall |s| \leq s_0,\ n = 1, 2, \ldots; i = 1, 2, \ldots, n. \tag{30}$$

In that case define

$$\sigma := \inf\{\varsigma > 0 : \text{ there exists } s_0 > 0 \text{ such that (30) holds}\}. \tag{31}$$

(C4) $\{V_i^n : n = 1, 2, \ldots; i = 1, 2, \ldots, n\}$ are identically distributed.

(C5) $V_i^n = V_i^i$ for all $n = 1, 2, \ldots; i = 1, 2, \ldots, n$.

Note that under (C1) and (C3), it follows that $E[|V_i^n|^2] \leq \sigma^2$, and under (C1), (C3), and (C4), $E[|V_i^n|^2] = \sigma^2$. Also, note that the collection of random variables $\{\tilde{W}_i^n : n = 1, 2, \ldots; i = 1, 2, \ldots, n\}$ obtained in the previous section with $\{\tilde{W}_i^n\}_{i=1}^n$

being a permutation of $\{W_i\}_{i=1}^n$ arranged in order of $\{\tilde{u}_i\}_{i=1}^n$, $n = 1, 2, \ldots$, satisfies conditions (C1), (C2), and (C3).

For each $n = 1, 2, \ldots$, define the partial sums of the random variables $\{V_i^n\}_{i=1}^n$ by

$$S_{j,k}^n := \sum_{i=j+1}^{k} V_i^n. \tag{32}$$

For $m \leq n$, define

$$B_{n,m} := \max_{0 \leq j \leq n} \max_{1 \leq k \leq (m \wedge n-j)} |S_{j,j+k}^n|. \tag{33}$$

Note that by setting $V_i^n = \tilde{W}_i^n$ of the previous section, we get $S_{j,k}^n = \tilde{S}_{j,k}^n$ and $B_{n,m} = \tilde{B}_{n,m}$ defined in (18) and (19), respectively.

Let $\{a_n\}$ be a sequence of positive integers satisfying some of the following conditions:

(A1) $1 \leq a_n \leq n$,
(A2) a_n is nondecreasing,
(A3) n/a_n is nondecreasing,
(A4) $a_n/\log n \to \infty$ as $n \to \infty$.
(A5) For some $a > 0$, $n^{a-1}a_n = o(n^{\varepsilon})$
(A6) For some $a', A' > 0$, $a_n \geq A'n^{a'}$. for every $\varepsilon > 0$.

Finally, let

$$\beta_n = (2a_n(\log(n/a_n) + \log\log n))^{1/2}.$$

In this section we determine the limiting behavior of $\{B_{n,a_n}/\beta_n\}$.

Below we give two lemmas that will be used to establish the limit theorems. The first is a minor modification of a maximal inequality due to de Acosta and Kuelbs [10, Lem. 3.1].

LEMMA 4.1. *Let $\{V_i\}$ be a sequence of independent random variables and let $S_{j,k} := \sum_{i=j+1}^{k} V_i$. Then for every integer $n \geq 0$, $m \geq 0$, $p \geq 0$, $p < n$, $m \leq n$, and $t > 0$, $s > 0$, $0 < a < 1$,*

$$P(B_{n,m} > t+s) \leq \frac{p}{1-c}\left[\max_{0 \leq j \leq n} P(|S_{j,j+m}| > at) + \max_{0 \leq j \leq n} P\left(|S_{j,j+\lfloor \frac{n}{p} \rfloor+1}| > \frac{(1-a)}{2}t\right)\right],$$

provided

$$\max_{0 \leq j \leq n} \max_{1 \leq k \leq m \vee \lfloor \frac{n}{p} \rfloor+1} P(|S_{j,j+k}| \geq s/4) \leq c < 1.$$

LEMMA 4.2. *Let $\{V_i^n\}$ be a collection of random variables satisfying conditions (C2) and (C3) with σ as defined in (31). Given any $\varsigma > \sigma$, let $s_0 > 0$ be such that (30) holds. Then for all $j \geq 0$, $k \geq 1$, and $0 < a \leq k\varsigma^2 s_0$,*

$$P(|S_{j,j+k}^n| \geq a) \leq 2\exp\left(-\frac{a^2}{2k\varsigma^2}\right).$$

Proof. For any $0 < s \leq s_0$,

$$\begin{aligned} P(S_{j,j+k}^n \geq a) &\leq \exp(-sa)E[\exp(sS_{j,j+k}^n)] \\ &= \exp(-sa)E[\exp(sV_{j+1}^n)] \cdots E[\exp(sV_{j+k}^n)] \\ &\leq \exp\left(\frac{k\varsigma^2 s^2}{2} - sa\right). \end{aligned}$$

The same upper bound also holds for $P(-S^n_{j,j+k} \geq a)$. Optimizing the upper bound over $0 \leq s \leq s_0$, we get the desired result. □

The first theorem from [9] gives almost sure rates for the process $\{B_{n,a_n}\}_{n=1}^{\infty}$ under conditions (C1)–(C5).

THEOREM 4.3 (Csörgő and Révész [9, Thm. 3.1.1]). *Let $\{V_i^n\}$ be a be a collection of random variables satisfying conditions* (C1)–(C5) *with variance σ^2. Let $\{a_n\}$ satisfy conditions* (A1)–(A4). *Then*

$$\overline{\lim_{n\to\infty}} \frac{B_{n,a_n}}{\beta_n} = \sigma \ a.s.$$

If in addition $\{a_n\}$ satisfies (A5), *then*

$$\lim_{n\to\infty} \frac{\max_{\Delta n \leq j \leq (1-\Delta)n} |S^n_{j,j+a_{n,j}}|}{\beta_n} = \sigma \ a.s.$$

for all $0 \leq \Delta < 1/2$, $0 < \rho \leq 1$, and $a_{n,j}$ satisfying $\rho a_n \leq a_{n,j} \leq a_n, \forall 1 \leq j \leq n$, $n = 1, 2, \ldots$.

From the above theorem we immediately have the same rate in probability for any collection $\{V_i^n\}$ satisfying conditions (C1)–(C4) (but not necessarily (C5)), i.e.,

$$\lim_n P(B_{n,a_n}/\beta_n > (\sigma + \epsilon)) = 0, \quad \forall \epsilon > 0, \tag{34}$$

and σ is the smallest constant for which (34) holds. In the next theorem we obtain upper bounds on the in-probability rate (with the above constant) and on the almost sure rate (with a larger constant) for $\{V_i^n\}$ that do not necessarily satisfy conditions (C4) or (C5).

THEOREM 4.4. *Let $\{V_i^n\}$ be a collection of random variables satisfying conditions* (C1)–(C3) *with σ as defined in* (31). *Let $\{a_n\}$ satisfy conditions* (A1)–(A4). *Then* (34) *holds. If in addition $\{a_n\}$ satisfies* (A5) *for some $a > 0$, then*

$$\overline{\lim_{n\to\infty}} \frac{B_{n,a_n}}{\beta_n} \leq \left(1 + \frac{1}{a}\right)^{1/2} \sigma \ a.s.$$

Proof. The theorem will follow from the Borel–Cantelli lemma if we can show that for all $\varsigma > \sigma$ and $\epsilon > 0$,

$$\sum_n P(B_{n,a_n}/\beta_n > (\rho\varsigma + 4\epsilon)) < \infty, \tag{35}$$

where $\rho = (1 + a^{-1})^{1/2}$. Substituting n, m, p, t, s in Lemma 4.1 by $n, a_n, \lfloor nd/a_n \rfloor + 1, (\rho\varsigma + 3\epsilon)\beta_n, \epsilon\beta_n$, respectively, we have

$$\begin{aligned} P(B_{n,a_n}/\beta_n > (\rho\varsigma + 4\epsilon)) &\leq 2\left(\frac{nd}{a_n} + 1\right) \max_{0\leq j\leq n} P(|S^n_{j,j+a_n}|/\beta_n > (\rho\varsigma + \epsilon)) \\ &\quad + 2\left(\frac{nd}{a_n} + 1\right) \max_{0\leq j\leq n} P(|S^n_{j,j+\lfloor a_n/d \rfloor + 1}|/\beta_n > \epsilon) \end{aligned} \tag{36}$$

provided

$$\max_{0\leq j\leq n} \max_{1\leq k\leq a_n+1} P(|S^n_{j,j+k}|/\beta_n \geq \epsilon/4) \leq 1/2. \tag{37}$$

By Chebyshev's inequality we have

$$\max_{0\le j\le n}\max_{1\le k\le a_n+1} P(|S^n_{j,j+k}|/\beta_n \ge \epsilon/4) \le \frac{\sigma^2(a_n+1)}{(\beta_n\epsilon/4)^2}.$$

By the definition of β_n and the assumptions on a_n, $a_n/\beta_n^2 \to 0$ as $n\to\infty$. So there exists an n_0 such that (37) holds for all $n \ge n_0$. We now apply Lemma 4.2 to upper bound the RHS of (36). Let s_0 be such that (30) holds. Note that by condition (A4) on $\{a_n\}$ it follows that $\beta_n/a_n \to 0$ as $n\to\infty$. Hence, there exists an $n_1 \ge n_0$ such that for all $n \ge n_1$, $(\rho\varsigma+\epsilon)\beta_n \le a_n\varsigma^2 s_0$, and consequently for all $n \ge n_1$,

$$\begin{aligned} P(|S^n_{j,j+a_n}|/\beta_n > (\rho\varsigma+\epsilon)) &\le 2\exp\left(-\frac{(\rho\varsigma+\epsilon)^2\beta_n^2}{2\varsigma^2 a_n}\right)\\ &= 2\exp\left(-\frac{(\rho\varsigma+\epsilon)^2 2a_n(\log(n/a_n)+\log\log n)}{2\varsigma^2 a_n}\right)\\ &= 2\exp(-(\rho')^2(\log(n/a_n)+\log\log n))\\ &= 2\left(\frac{a_n}{n\log n}\right)^{(\rho')^2}, \end{aligned}$$

where $\rho' = \rho + \epsilon/\varsigma$. Thus, for all $n \ge n_1$, the first term on the RHS of (36) is bounded above by

$$(38)\ 2\left(\frac{nd}{a_n}+1\right)\max_{0\le j\le n} P(|S^n_{j,j+a_n}|/\beta_n > (\rho\varsigma+\epsilon)) \le 8d\left(\frac{a_n}{n}\right)^{(\rho')^2-1}(\log n)^{-(\rho')^2}.$$

In view of the additional condition on a_n it is easy to check that

$$\left(\frac{a_n}{n}\right)^{(\rho')^2-1}(\log n)^{-(\rho')^2} \le \left(\frac{a_n}{n}\right)^{(\rho')^2-1} = o(1/n^\gamma),$$

where $\gamma = a((\rho+\epsilon/2\varsigma)^2-1) > 1$. By choosing the constant $d > ((\rho\varsigma+\epsilon)/\epsilon)^2$, the same asymptotic upper bound can be obtained for the second term on the RHS of (36). This establishes (35).

Finally, by setting $\rho = 1$ in (36) through (38), it can easily be seen that (34) holds. □

In the next theorem we obtain the same upper bound as in Theorem 4.3 for the sequence $\{\tilde B_{n,a_n}\}$ that was defined in the previous section.

THEOREM 4.5. *Let $\{\tilde W_i^n\}$ be the collection of random variables defined in the previous section with σ as defined in* (31). *Let $\{\tilde B_{n,a_n}\}$ be as defined in* (19). *Let $\{a_n\}$ satisfy conditions* (A1)–(A5) *with $a > 0$ as in* (A5). *Then*

$$\overline{\lim_{n\to\infty}}\frac{\tilde B_{n,a_n}}{\beta_n} \le \sigma \ a.s.$$

Proof. Let $\{\tilde W_i^{m,n}\}_{i=m}^n$ be a permutation of $\{W_i\}_{i=m}^n$ arranged in order of $\{\tilde u_i\}_{i=m}^n$. Define their partial sums

$$(39)\qquad \tilde S^{m,n}_{j,k} := \sum_{i=j+1}^{k}\tilde W_i^{m,n},\qquad m-1\le j\le k\le n.$$

Define

$$\tilde{B}^m_{n,a_n} := \max_{m-1\le j\le n}\ \max_{1\le k\le(a_n\wedge n-j)} |\tilde{S}^{m,n}_{j,j+k}|. \tag{40}$$

Then, in terms of our prior notation, $\tilde{B}_{n,a_n} = \tilde{B}^1_{n,a_n}$. The proof of this theorem depends crucially on the following simple observation:

$$\tilde{B}_{n,a_n} \le \tilde{B}_{m,a_n} + \tilde{B}^m_{n,2\frac{n-m}{n}a_n+1} \qquad \forall m \le n.$$

This fact is a consequence of the nearly uniformly spaced design sequence $\{u_n\}$. Now let $n_k = \lfloor \theta^k \rfloor$, $k = 1, 2, \dots$ for some $\theta > 1$. Given any n, let k be such that $n_k \le n < n_{k+1}$. Then by substituting $m = n_k$ above and using the fact that a_n is increasing, we get

$$\tilde{B}_{n,a_n} \le \tilde{B}_{n_k,a_{n_{k+1}}} + \tilde{B}^{n_k}_{n,2\frac{n-n_k}{n}a_n+1}.$$

Now, since $\beta_{n_k} \le \beta_n \le \beta_{n_{k+1}}$, we get

$$\frac{\tilde{B}_{n,a_n}}{\beta_n} \le \frac{\beta_{n_{k+1}}}{\beta_{n_k}} \frac{\tilde{B}_{n_k,a_{n_{k+1}}}}{\beta_{n_{k+1}}} + \frac{\tilde{B}^{n_k}_{n,2\frac{n-n_k}{n}a_n+1}}{\beta_n}.$$

Let $\rho = (1 + a^{-1})^{1/2}$. The theorem will follow if for all $\varsigma > \sigma$ and $0 < \eta < 2\varsigma\rho^2$, we can find a $\theta > 1$ such that

$$\overline{\lim_{k\to\infty}} \frac{\beta_{n_{k+1}}}{\beta_{n_k}} \frac{\tilde{B}_{n_k,a_{n_{k+1}}}}{\beta_{n_{k+1}}} \le \varsigma + \eta \text{ a.s.}, \tag{41}$$

and

$$\overline{\lim_{n\to\infty}} \frac{\tilde{B}^{n_k}_{n,2\frac{n-n_k}{n}a_n+1}}{\beta_n} \le \eta \text{ a.s.} \tag{42}$$

Choose $\theta = 1 + (\eta/\sqrt{2}\rho\varsigma)^2$. It is easy to check that

$$\overline{\lim_{k\to\infty}} \frac{\beta_{n_{k+1}}}{\beta_{n_k}} \le \theta.$$

Hence, (41) will follow by Borel–Cantelli if we establish that for all $\epsilon > 0$,

$$\sum_k P\left(\frac{\tilde{B}_{n_{k-1},a_{n_k}}}{\beta_{n_k}} > \varsigma + 4\epsilon\right) < \infty. \tag{43}$$

Using Lemmas 4.1 and 4.2, as was done in the proof of Theorem 4.4 (cf. (36), (38)), we get

$$P\left(\frac{\tilde{B}_{n_{k-1},a_{n_k}}}{\beta_{n_k}} > \varsigma + 4\epsilon\right) \le O((\log n_k)^{-(1+\epsilon/\varsigma)}) \le O(k^{-(1+\epsilon/\varsigma)})$$

and (43) holds. Similarly, (42) will follow by Borel–Cantelli if we establish that for all $\epsilon > 0$,

$$\sum_n P\left(\frac{\tilde{B}^{n_k}_{n,2\frac{n-n_k}{n}a_n+1}}{\beta_n} > \eta + 4\epsilon\right) < \infty.$$

Since $\tilde{B}^{n_k}_{n,2\frac{n-n_k}{n}a_n+1} \leq \tilde{B}^{n_k}_{n,2\frac{\theta-1}{\theta}a_n+1}$, the above will follow from

$$\sum_n P\left(\frac{\tilde{B}^{n_k}_{n,2\frac{\theta-1}{\theta}a_n+1}}{\beta_n} > \eta + 4\epsilon\right) < \infty. \tag{44}$$

Again, using Lemmas 4.1 and 4.2, as was done in the proof of Theorem 4.4 (cf. (36), (38)), we get

$$P\left(\frac{\tilde{B}^{n_k}_{n,2\frac{\theta-1}{\theta}a_n+1}}{\beta_n} > \eta + 4\epsilon\right) \leq O\left(\left(\frac{a_n}{n}\right)^{(1+a^{-1})(1+\epsilon/\eta)-1}\right) \leq o(n^{-\gamma})$$

for some $\gamma > 1$. Hence (44) holds. □

We now proceed to obtain the same upper bound as in Theorem 4.3 but in L^p instead of almost surely. We first establish the following lemma which is needed to obtain the L^p upper bound.

LEMMA 4.6. *Let $\{V_i^n\}$ be a collection of random variables satisfying condition* (C2). *If $\{B_{n,a_n}/\beta_n\}_{n=1}^{\infty}$ is stochastically bounded and* $\sup_n E[\sup_{1\leq i\leq n}(|V_i^n|/\beta_n)^{p'}] < \infty$, *then* $\sup_n E[(B_{n,a_n}/\beta_n)^{p'}] < \infty$ *and consequently, $\{(B_{n,a_n}/\beta_n)^p\}_{n=1}^{\infty}$ is uniformly integrable for all $1 \leq p < p'$.*

Proof. The proof of this lemma is based on a technique borrowed from [19]. We first prove the following claim:

$$P(B_{n,a_n} \geq 2t + s) \leq P(B_{n,a_n} \geq t)^2 + P(N_n \geq s),$$

where $N_n := \sup_{1\leq i\leq n} |V_i^n|$. Let

$$B^{l,m}_{n,a_n} := \max_{\substack{l-1\leq j\leq k\leq m\\ k-j\leq a_n}} |S^n_{j,k}|, \qquad 1 \leq l \leq m \leq n.$$

Thus,

$$B_{n,a_n} = B^{1,n}_{n,a_n} \leq B^{1,m-1}_{n,a_n} + |V^n_m| + B^{m+1,n}_{n,a_n} \quad \text{for any } 1 \leq m \leq n. \tag{45}$$

Let T be the stopping time defined by

$$T := \inf\{1 \leq m \leq n : B^{1,m}_{n,a_n} \geq t\},$$

where $\inf\emptyset = \infty$. Then $B^{1,n}_{n,a_n} \geq 2t + s$ implies that $T \leq n$, and so we have

$$\begin{aligned} P(B_{n,a_n} \geq 2t + s) &= \sum_{m=1}^{n} P(B_{n,a_n} \geq 2t + s, T = m) \\ &\leq \sum_{m=1}^{n} P(B^{1,m-1}_{n,a_n} + |V^n_m| + B^{m+1,n}_{n,a_n} \geq 2t + s, T = m) \\ &\leq \sum_{m=1}^{n} P(B^{m+1,n}_{n,a_n} \geq t + s - N_n, T = m) \\ &\leq \sum_{m=1}^{n} P(B^{m+1,n}_{n,a_n} \geq t, T = m) + P(N_n \geq s) \end{aligned}$$

$$
\begin{aligned}
&= \sum_{m=1}^{n} P(B_{n,a_n}^{m+1,n} \geq t) P(T = m) + P(N_n \geq s) \\
&\leq P(B_{n,a_n}^{1,n} \geq t) \sum_{m=1}^{n} P(T = m) + P(N_n \geq s) \\
&= P(B_{n,a_n} \geq t)^2 + P(N_n \geq s).
\end{aligned}
$$

The first inequality follows from (45), the second from the fact that $T = m$ implies that $B_{n,a_n}^{1,m-1} < t$, and the third by the definition of N_n. The next equality follows from the fact that $\{B_{n,a_n}^{m+1,n} \geq t\}$ and $\{T = m\}$ are independent events.

Pick A so that $P(B_{n,a_n}/\beta_n \geq A) \leq 1/(2.3^{p'})$. Note that this is possible by the assumption that $\{B_{n,a_n}/\beta_n\}_{n=1}^{\infty}$ is stochastically bounded. Now we have

$$
\begin{aligned}
E[(B_{n,a_n}/\beta_n)^{p'}] &= \int_0^\infty p' x^{p'-1} P(B_{n,a_n}/\beta_n \geq x)\, dx \\
&\leq (3A)^{p'} + \int_{3A}^\infty p' x^{p'-1} P(B_{n,a_n}/\beta_n \geq x)\, dx \\
&\leq (3A)^{p'} + \int_{3A}^\infty p' x^{p'-1} P(N_n/\beta_n \geq x/3)\, dx \\
&\qquad + \int_{3A}^\infty p' x^{p'-1} P(B_{n,a_n}/\beta_n \geq x/3)^2\, dx \\
&= (3A)^{p'} + 3^{p'} \int_A^\infty p' x^{p'-1} P(N_n/\beta_n \geq x)\, dx \\
&\qquad + 3^{p'} \int_A^\infty p' x^{p'-1} P(B_{n,a_n}/\beta_n \geq x)^2\, dx \\
&\leq (3A)^{p'} + 3^{p'} \int_A^\infty p' x^{p'-1} P(N_n/\beta_n \geq x)\, dx \\
&\qquad + \frac{1}{2} \int_A^\infty p' x^{p'-1} P(B_{n,a_n}/\beta_n \geq x)\, dx \\
&\leq (3A)^{p'} + 3^{p'} E[(N_n/\beta_n)^{p'}] + \frac{1}{2} E[(B_{n,a_n}/\beta_n)^{p'}].
\end{aligned}
$$

Thus

$$
E[(B_{n,a_n}/\beta_n)^{p'}] \leq 2(3A)^{p'} + 2.3^{p'} E[(N_n/\beta_n)^{p'}] < \infty.
$$

The uniform integrability of $\{(B_{n,a_n}/\beta_n)^p\}_{n=1}^{\infty}$ for all $1 \leq p < p'$ is an immediate consequence. This completes the proof of Lemma 4.6. □

The L^p upper bound is now established in the following theorem.

THEOREM 4.7. *Let* $\{V_i^n\}$ *be a collection of random variables satisfying conditions* (C1)–(C3) *with* σ *as defined in* (31). *Let* $\{a_n\}$ *satisfy conditions* (A1)–(A4) *and* (A6). *Then*

$$
\varlimsup_{n \to \infty} E[(B_{n,a_n}/\beta_n)^p]^{1/p} \leq \sigma \quad \forall p \geq 1.
$$

Proof. The theorem will follow from the in-probability bound (34) obtained in Theorem 4.4, if we establish uniform integrability of $\{(B_{n,a_n}/\beta_n)^p\}_{n=1}^{\infty}$. In view of Lemma 4.6, it suffices to establish that $\{B_{n,a_n}/\beta_n\}_{n=1}^{\infty}$ is stochastically bounded and that $\sup_n E[\sup_{1\le i\le n}(|V_i^n|/\beta_n)^{p'}] < \infty$ for some $p' > p$. That $\{B_{n,a_n}/\beta_n\}_{n=1}^{\infty}$ is stochastically bounded also follows from the in-probability bound (34) obtained in Theorem 4.4. Finally, using the fact that $E[|X|] \le E[|X|^r]^{1/r}$ for any $r \ge 1$, we get

$$\begin{aligned}\sup_n E\left[\sup_{1\le i\le n}(|V_i^n|/\beta_n)^{p'}\right] &= \sup_n (1/\beta_n)^{p'} E\left[\sup_{1\le i\le n}|V_i^n|^{p'}\right]\\ &\le \sup_n (1/\beta_n)^{p'} E\left[\sup_{1\le i\le n}|V_i^n|^{p'r}\right]^{1/r}\\ &\le \sup_n (1/\beta_n)^{p'} E\left[\sum_{i=1}^n |V_i^n|^{p'r}\right]^{1/r}\\ &\le \sup_n (1/\beta_n)^{p'}\left(\sum_{i=1}^n E[|V_i^n|^{p'r}]\right)^{1/r}\\ &\le \sup_n (n/(\beta_n)^{p'r})^{1/r}\sup_{\substack{1\le i\le n\\ 1\le n<\infty}} E[|V_i^n|^{p'r}]^{1/r}\\ &\le \sup_n (n/((2A')^{1/2}n^{a'/2})^{p'r})^{1/r}\sup_{\substack{1\le i\le n\\ 1\le n<\infty}} E[|V_i^n|^{p'r}]^{1/r}\\ &\le (2A')^{-p'/2}(1\vee n^{1/r-a'p'/2})\sup_{\substack{1\le i\le n\\ 1\le n<\infty}} E[|V_i^n|^{p'r}]^{1/r}\\ &= (2A')^{-p'/2}\sup_{\substack{1\le i\le n\\ 1\le n<\infty}} E[|V_i^n|^{p'r}]^{1/r}\\ &< \infty\end{aligned}$$

if we choose $r = 1 \vee 2/a'p'$. Note that we have used the condition that $a_n \ge A'n^{a'}$ to obtain a corresponding bound on β_n in one of the above steps. Also, we have used condition (C3) on the moment-generating function of $\{V_i^n\}$ to deduce the last inequality. This completes the proof of the theorem. □

5. The adaptive control scheme. In this section we construct a class of certainty equivalence control with forcing-type adaptive control schemes based on the learning schemes constructed in §3. Let $\{\tau_i\}_{i=1}^{\infty}$ be a positive integer-valued sequence to be specified later. Define the related sequence $\{t_i\}_{i=1}^{\infty}$ as follows:

$$t_i := 1 + \sum_{k=1}^{i-1}(\tau_k+1) = \sum_{k=1}^{i-1}\tau_k + i, \quad i \ge 1.$$

At times t_i, $i \ge 1$ use (force) the ith control u_i from the design sequence of the learning scheme such as the one described in the previous section. Let U_i^* be the estimate based on the corresponding observations at times t_k, $1 \le k \le i$, with the design sequence u_k, $1 \le k \le i$. Use the control U_i^* from time t_i+1 to time $t_{i+1}-1$, i.e., τ_i times. Thus,

$$U_{t_i} = u_i, \quad U_n = U_i^* \text{ for } t_i+1 \le n \le t_{i+1}-1, \quad i \ge 1.$$

This completes the description of the adaptive control scheme.

Let

$$\kappa(n) := \min\{i : t_i > n\} - 1 = \max\{i : t_i \le n\} = \max\left\{i : \sum_{k=1}^{i-1}\tau_k + i \le n\right\}.$$

The following theorems and their corollaries provide upper bounds on the learning loss associated with the class of schemes constructed above in terms of $\kappa(n)$.

THEOREM 5.1. *Assume that for a certain learning scheme $m^* - m(U_i^*) = O(r_i)$ a.s. (resp., in L^p with $p \geq 1$) for some known sequence r_i. Then for the certainty equivalence control with forcing-type scheme constructed above with the sequence $\tau_i = \lfloor br_i^{-1} \rfloor$ for some $b > 0$, we have $L_n = O(\kappa(n))$ a.s. (resp., in L^p with $p \geq 1$).*

Proof. Since, $m : \mathcal{U} = [\Delta, 1 - \Delta] \to \mathbb{R}$ is uniformly locally Lipschitz, it follows that $K := \sup_{u \in \mathcal{U}} m(u) - \inf_{u \in \mathcal{U}} m(u) < \infty$. Thus,

$$\begin{aligned} L_n &= \sum_{l=1}^{n} m^* - m(U_l) \\ &\leq K\kappa(n) + \sum_{i=1}^{\kappa(n)} \sum_{l=t_i+1}^{t_{i+1}-1} m^* - m(U_l) \\ &= K\kappa(n) + \sum_{i=1}^{\kappa(n)} (m^* - m(U_i^*))\tau_i \\ &\leq K\kappa(n) + \sum_{i=1}^{\kappa(n)} (m^* - m(U_i^*))br_i^{-1}. \end{aligned}$$

Now, for the first part we are given that $m^* - m(U_i^*) = O(r_i)$ a.s. Thus, $\overline{\lim}_i (m^* - m(U_i^*))r_i^{-1} \leq C$ a.s. for some $C \geq 0$ that could depend on ω. That is, for all $\epsilon > 0$, there exists an $i_0 \geq 1$, such that for all $i \geq i_0$, $m^* - m(U_i^*) \leq (C + \epsilon)r_i$ a.s. Then, clearly,

$$\begin{aligned} L_n &\leq K\kappa(n) + \sum_{i=1}^{i_0 - 1} (m^* - m(U_i^*))br_i^{-1} + \sum_{i=i_0}^{\kappa(n)} (m^* - m(U_i^*))br_i^{-1} \\ &\leq K\kappa(n) + M(\epsilon) + (C + \epsilon)b\kappa(n) \qquad \text{a.s.}, \end{aligned}$$

where $M(\epsilon) = \sum_{i=1}^{i_0 - 1} (m^* - m(U_i^*))br_i^{-1}$ depends on ϵ but not on n. Dividing by $\kappa(n)$ and taking the limit as $n \to \infty$ we get

$$\overline{\lim_n} \frac{L_n}{\kappa(n)} \leq K + (C + \epsilon)b \qquad \text{a.s.}$$

Now, by letting $\epsilon \to 0$ we get

$$\overline{\lim_n} \frac{L_n}{\kappa(n)} \leq K + Cb \qquad \text{a.s.}$$

For the second part we are given that $m^* - m(U_i^*) = O(r_i)$ in L^p with $p \geq 1$. Thus, $\overline{\lim}_i E[|m^* - m(U_i^*)|^p]r_i^{-p} \leq C^p$ for some $C \geq 0$. That is, for all $\epsilon > 0$ there exists an $i_0 \geq 1$, such that for all $i \geq i_0$, $E[|m^* - m(U_i^*)|^p]r_i^{-p} \leq C^p + \epsilon$. Also by the well-known inequality $|x_1 + \cdots + x_k|^p \leq k^{p-1}(|x_1|^p + \cdots + |x_k|^p)$ we get

$$|L_n|^p \leq 2^{p-1} \left(K^p(\kappa(n))^p + (\kappa(n))^{p-1} \sum_{i=1}^{\kappa(n)} |m^* - m(U_i^*)|^p b^p r_i^{-p} \right).$$

Therefore,

$$\begin{aligned} E[|L_n|^p] &\le 2^{p-1}\left(K^p(\kappa(n))^p + (\kappa(n))^{p-1}\sum_{i=1}^{\kappa(n)} E[|m^* - m(U_i^*)|^p]b^p r_i^{-p}\right) \\ &\le 2^{p-1}\Bigg(K^p(\kappa(n))^p + (\kappa(n))^{p-1}\sum_{i=1}^{i_0-1} E[|m^* - m(U_i^*)|^p]b^p r_i^{-p} \\ &\qquad + (\kappa(n))^{p-1}\sum_{i=i_0}^{\kappa(n)} E[|m^* - m(U_i^*)|^p]b^p r_i^{-p}\Bigg) \\ &\le 2^{p-1}(K^p(\kappa(n))^p + (\kappa(n))^{p-1}M(\epsilon) + (C^p+\epsilon)b^p(\kappa(n))^p), \end{aligned}$$

where $M(\epsilon) = \sum_{i=1}^{i_0-1} E[|m^* - m(U_i^*)|^p]b^p r_i^{-p}$ depends on ϵ but not on n. Dividing by $(\kappa(n))^p$ and taking the limit as $n \to \infty$ we get

$$\varlimsup_n \frac{E[|L_n|^p]}{(\kappa(n))^p} \le 2^{p-1}(K^p + (C^p+\epsilon)b^p).$$

Now, by letting $\epsilon \to 0$ we get

$$\varlimsup_n \frac{E[|L_n|^p]}{(\kappa(n))^p} \le 2^{p-1}(K^p + C^p b^p). \qquad \square$$

Note that we can also obtain a constant (in terms of K, C, b) associated with the rate of increase of the learning loss, $\kappa(n)$. Moreover, $\kappa(n)$ itself depends on b. In fact, if $\tau_i \to \infty$ as $i \to \infty$, then asymptotically $\kappa_b(n) \approx \kappa_1(n/b)$, where the subscript on $\kappa(n)$ denotes the dependence on b. We may therefore want to choose b to minimize the rate along with the constant.

COROLLARY 5.2. *For the certainty equivalence control with forcing-type scheme constructed above, with the learning scheme of* §3 *with the bandwidth sequence* $h_i = h(\log i/i)^{1/(2\alpha+1)}$, *and with the sequence* $\tau_i = \lfloor b(i/\log i)^{\alpha/(2\alpha+1)}\rfloor$ *for some* $b > 0$, *we have* $L_n = O(\tilde{B}^{-1}(n))$ *a.s. and in* $L^p, p \ge 1$, *where* $\tilde{B}^{-1} : [0,\infty) \to [e,\infty)$ *is the inverse of the function* $\tilde{B} : [e,\infty) \to [0,\infty)$ *defined by*

$$\tilde{B}(t) := \int_e^t b\left(\frac{s}{\log s}\right)^{\alpha/(2\alpha+1)} ds.$$

Moreover, $\tilde{B}^{-1}(n) = o(n^{\frac{2\alpha+1}{3\alpha+1}+\eta})$ *for all* $\eta > 0$.

Proof. This corollary will follow immediately from Theorems 5.1 and 3.1, if we can show that $\kappa(n) = O(\tilde{B}^{-1}(n))$, and that $\tilde{B}^{-1}(n) = o(n^{\frac{2\alpha+1}{3\alpha+1}+\eta})$ for all $\eta > 0$, when $\tau_i = \lfloor b(i/\log i)^{\alpha/(2\alpha+1)}\rfloor$. To this end, define the functions $v, B : \mathbb{R}^+ \to \mathbb{R}^+$ based on the sequence $\{\tau_i\}$ as follows:

$$\begin{aligned} v(s) &:= \tau_{\lceil s\rceil} + 1, \\ B(t) &:= \int_0^t v(s)\, ds. \end{aligned}$$

Then

$$B(i) = \sum_{k=1}^{i} \tau_k + i, \quad i = 0, 1, \ldots.$$

Also, observe that (1) $B(t)$ is continuous in $t > 0$, (2) $B(0) = 0$, (3) $B(t)$ is strictly increasing in $t > 0$ (since $v(s) > 0$), and (4) $B(t) \geq t$ (since $v(s) \geq 1$). Hence, given any n, there exists a unique $0 \leq t \leq n$ such that $B(t) = n$. Denote this solution by $B^{-1}(n)$. Then

$$\begin{aligned} \sum_{k=1}^{\lceil B^{-1}(n)\rceil+1-1} \tau_k + \lceil B^{-1}(n)\rceil + 1 &= B(\lceil B^{-1}(n)\rceil) + 1 \\ &\geq B(B^{-1}(n)) + 1 \\ &= n + 1 > n. \end{aligned}$$

Therefore, by the definition of $\kappa(n)$ it follows that $\kappa(n) \leq \lceil B^{-1}(n)\rceil \leq B^{-1}(n)+1$. It is also easy to see that if we have a function $\tilde{v} : \mathbb{R}^+ \to \mathbb{R}^+$ such that $v(s) \geq \tilde{v}(s)$, $s > 0$, then $B(t) \geq \tilde{B}(t)$, $t > 0$, and hence $B^{-1}(n) \leq \tilde{B}^{-1}(n)$, $n \geq 1$. For the sequence $\tau_i = \lfloor b(i/\log i)^{\alpha/(2\alpha+1)} \rfloor$ under consideration, note that $v(s) = \lfloor b(\lceil s\rceil/\log\lceil s\rceil)^{\alpha/(2\alpha+1)} \rfloor + 1 \geq b(s/\log s)^{\alpha/(2\alpha+1)} I\{s \geq e\} =: \tilde{v}(s)$. Thus $\kappa(n) = O(\tilde{B}^{-1}(n))$. It is easy to verify that $\tilde{B}^{-1}(n) = o(n^{\frac{2\alpha+1}{3\alpha+1}+\eta})$ for any $\eta > 0$. □

THEOREM 5.3. *Assume that for a certain learning scheme $P(m^* - m(U_i^*) > \epsilon) = O(r_i)$ for some known sequence r_i. Then for the certainty equivalence control with forcing-type scheme constructed above with the sequence $\tau_i = \lfloor br_i^{-1} \rfloor$ for some $b > 0$, we have $E[L_n^\epsilon] = O(\kappa(n))$ for all $\epsilon > 0$.*

Proof.

$$\begin{aligned} E[L_n^\epsilon] &= \sum_{l=1}^{n} P(m^* - m(U_l) > \epsilon) \\ &\leq \kappa(n) + \sum_{i=1}^{\kappa(n)} \sum_{l=t_i+1}^{t_{i+1}-1} P(m^* - m(U_l) > \epsilon) \\ &= \kappa(n) + \sum_{i=1}^{\kappa(n)} P(m^* - m(U_i^*) > \epsilon)\tau_i \\ &\leq \kappa(n) + \sum_{i=1}^{\kappa(n)} P(m^* - m(U_i^*) > \epsilon) b r_i^{-1}. \end{aligned}$$

Now, we are given that $P(m^* - m(U_i^*) > \epsilon) = O(r_i)$. Thus, $\overline{\lim}_i P(m^* - m(U_i^*) > \epsilon) r_i^{-1} \leq C$ for some $C \geq 0$. That is, for all $\eta > 0$ there exists an $i_0 \geq 1$, such that for all $i \geq i_0$, $P(m^* - m(U_i^*) > \epsilon) \leq (C + \eta) r_i$. Then, clearly

$$\begin{aligned} E[L_n^\epsilon] &\leq \kappa(n) + \sum_{i=1}^{i_0-1} P(m^* - m(U_i^*) > \epsilon) b r_i^{-1} + \sum_{i=i_0}^{\kappa(n)} P(m^* - m(U_i^*) > \epsilon) b r_i^{-1} \\ &\leq \kappa(n) + M(\eta) + (C + \eta) b \kappa(n), \end{aligned}$$

where $M(\eta) = \sum_{i=1}^{i_0-1} P(m^* - m(U_i^*) > \epsilon) b r_i^{-1}$ depends on η but not on n. Dividing by $\kappa(n)$ and taking the limit as $n \to \infty$ we get

$$\overline{\lim_{n}} \frac{E[L_n^\epsilon]}{\kappa(n)} \leq 1 + (C + \eta) b.$$

Now, by letting $\eta \to 0$ we get

$$\overline{\lim_n} \frac{E[L_n^\epsilon]}{\kappa(n)} \leq 1 + Cb. \qquad \square$$

COROLLARY 5.4. *For the certainty equivalence control with forcing-type scheme constructed above, with the learning scheme of* §3 *with any bandwidth sequence* h_i *such that* (i) $h_i \to 0$ *as* $i \to \infty$ *and* (ii) $ih_i \geq A'i^{a'}$ *for some* $0 < a' < 1$, $A' > 0$, *and with the sequence* $\tau_i = \lfloor be^{i(h_i)^2} \rfloor$ *for some* $b > 0$, *we have*

$$E[L_n^\epsilon] = O((h_n)^{-2} \log n), \tag{46}$$

for all $\epsilon > 0$.

Proof. This corollary will follow immediately from Theorems 5.3 and 3.1, if we can show that

$$\kappa(n) = O((h_n)^{-2} \log n) \tag{47}$$

when $\tau_i = \lfloor be^{i(h_i)^2} \rfloor$. Now by the definition of $\kappa(n)$, it follows that

$$\sum_{k=1}^{\kappa(n)-1} \tau_k + \kappa(n) \leq n.$$

Taking only the last term in the above sum we get

$$(\kappa(n) - 1) \leq h_{\kappa(n)-1}^{-2} (\log n - \log b) \leq h_n^{-2} (\log n - \log b).$$

This establishes (47). $\square$

Remark. In light of (46) above, we can choose $h_n \to 0$ arbitrarily slowly, thereby making $E[L_n^\epsilon]$ arbitrarily close to $O(\log n)$.

6. Concluding remarks. The ϵ-learning loss of the class of adaptive control schemes constructed in this paper for the case of an infinite number of arms is of the same order as those obtained previously for the finite case. For the finite case it is easy to see that the learning loss is within a constant factor of the ϵ-learning loss. The infinite case is fundamentally different in this respect. Thus, the learning loss that we obtain for the infinite case is considerably worse than those available for the finite case. Since we do not have any tighter lower bounds on the learning loss other than those available for the finite case, there may be room for improvement. However, to the best of our knowledge, these are the best rates available to date. Moreover, the rates obtained by us are still stronger than the $o(n)$ required for optimality with respect to the average-cost-per-unit-time criterion.

In a forthcoming paper [1] we extend the results of this paper to the multiarmed bandit problem as well as to the adaptive control of Markov chains, both with a control set $\mathcal{U}$ which is a bounded subset of $\mathbb{R}^d$, $d > 1$. The principal difficulty with this extension is that strong limit laws for moving averages in higher dimensions are not available in the literature, and are also much more difficult to obtain for the kind of sampling/design scheme that we employ.

Appendix: Proof of (15), (16), (27). First let $l(n) := \lfloor \log_2 n \rfloor + 1$, and consider the representation of n in base 2:

$$n = n_{l(n)} \dots n_2 n_1.$$

Note that we can partition $\{1 \dots n\}$ as

$$\{1 \dots n\} = \bigcup_{\substack{i=1,\dots,l(n):\\ n_i=1}} E_i^n,$$

where for $i = 1, \dots, l(n)$ such that $n_i = 1$, E_i^n are defined by

$$\begin{aligned} E_i^n &= \sum_{l=i+1}^{l(n)} n_l 2^{l-1} + \{1, \dots, 2^{i-1}\} \\ &= \{(n_{l(n)} \dots n_{i+1} 00 \dots 01), \dots, \\ &\qquad (n_{l(n)} \dots n_{i+1} 01 \dots 11), (n_{l(n)} \dots n_{i+1} 10 \dots 00)\} \\ &=: \{\underline{e}_i^n, \dots, \overline{e}_i^n\}. \end{aligned}$$

We can now partition $\{\tilde{u}_i\}_{i=1}^n$ correspondingly as

$$\{\tilde{u}_i\}_{i=1}^n = \bigcup_{\substack{i=1,\dots,l(n):\\ n_i=1}} \{\tilde{u}_k, k \in E_i^n\}.$$

It is easy to see that $\{\tilde{u}_k, k \in E_i^n\}$ is a uniform lattice of 2^{i-1} points in $[0, 1]$. More precisely,

$$\tilde{u}_k = \tilde{u}_{\underline{e}_i^n} + \tilde{u}_{k+1-\underline{e}_i^n}$$

for $k \in E_i^n$. As k ranges over the set E_i^n, the first term on the RHS stays fixed, and the second term gives us precisely the set of all dyadic rationals in $[0, 1]$ with denominators dividing 2^{i-1}. There are a total of 2^{i-1} such points. Thus we can think of $\{\tilde{u}_i\}_{i=1}^n$ as an overlay of these lattices of various levels of coarseness.

Then, we get for all $n \geq 1$ and $u \in \mathcal{U}$,

$$\begin{aligned} \sum_{i=1}^n I\{|u - \tilde{u}_i| \leq h_n/2\} &= \sum_{\substack{i=1,\dots,l(n):\\ n_i=1}} \sum_{k \in E_i^n} I\{|u - \tilde{u}_k| \leq h_n/2\} \\ &\leq \sum_{\substack{i=1,\dots,l(n):\\ n_i=1}} (2^{i-1} h_n + 1) \\ &= n h_n + l(n). \end{aligned}$$

Similarly, for all $n \geq n_1$ and $u \in \mathcal{U}$,

$$\begin{aligned} \sum_{i=1}^n I\{|u - \tilde{u}_i| \leq h_n/2\} &= \sum_{\substack{i=1,\dots,l(n):\\ n_i=1}} \sum_{k \in E_i^n} I\{|u - \tilde{u}_k| \leq h_n/2\} \\ &\geq \sum_{\substack{i=1,\dots,l(n):\\ n_i=1}} (2^{i-1} h_n - 1) \\ &= n h_n - l(n). \end{aligned}$$

This establishes (15).

Also, all $n \geq 1$ and $u \in \mathcal{U}$,

$$
\begin{aligned}
&\sum_{i=1}^{n} I\{|u-\tilde{u}_i| \leq h_n/2\}|\tilde{u}_i - u|^\alpha \\
&= \sum_{\substack{i=1,\ldots,l(n):\\ n_i=1}} \sum_{k\in E_i^n} I\{|u-\tilde{u}_k| \leq h_n/2\}|\tilde{u}_k - u|^\alpha \\
&= \sum_{\substack{i=1,\ldots,l(n):\\ n_i=1}} \sum_{k\geq 0} [(s(u,n,i)+k2^{1-i})^\alpha I\{(s(u,n,i)+k2^{1-i}) \leq h_n/2\} \\
&\qquad\qquad +(2^{1-i}-s(u,n,i)+k2^{1-i})^\alpha I\{(2^{1-i}-s(u,n,i)+k2^{1-i}) \leq h_n/2\}],
\end{aligned}
$$

where $0 \leq s(u,n,i) < 2^{1-i}$ is the distance from u to the closest point in $E_i^n \cap [u,1]$. Next, note that

$$
\begin{aligned}
&\sum_{\substack{i=1,\ldots,l(n):\\ n_i=1}} \sum_{k\geq 0} [(s(u,n,i)+k2^{1-i})^\alpha I\{(s(u,n,i)+k2^{1-i}) \leq h_n/2\} \\
&\qquad\qquad +(2^{1-i}-s(u,n,i)+k2^{1-i})^\alpha I\{(2^{1-i}-s(u,n,i)+k2^{1-i}) \leq h_n/2\}] \\
&\leq \sum_{\substack{i=1,\ldots,l(n):\\ n_i=1}} \sum_{k\geq 0} 2\left[2^{i-1}\int_0^{h_n/2} x^\alpha dx + \left(\frac{h_n}{2}\right)^\alpha\right] \\
&= 2 \sum_{\substack{i=1,\ldots,l(n):\\ n_i=1}} \sum_{k\geq 0} \left[2^{i-1}\frac{1}{\alpha+1}\left(\frac{h_n}{2}\right)^{\alpha+1} + \left(\frac{h_n}{2}\right)^\alpha\right] \\
&= 2\left[\frac{1}{\alpha+1}\left(\frac{h_n}{2}\right)^{\alpha+1} n + \left(\frac{h_n}{2}\right)^\alpha l(n)\right] \\
&= \frac{1}{\alpha+1}\left(\frac{h_n}{2}\right)^\alpha [nh_n + 2(\alpha+1)l(n)],
\end{aligned}
$$

which establishes (16). Similarly,

$$
\begin{aligned}
&\sum_{\substack{i=1,\ldots,l(n):\\ n_i=1}} \sum_{k\geq 0} [(s(u,n,i)+k2^{1-i})^\alpha I\{(s(u,n,i)+k2^{1-i}) \leq h_n/2\} \\
&\qquad\qquad +(2^{1-i}-s(u,n,i)+k2^{1-i})^\alpha I\{(2^{1-i}-s(u,n,i)+k2^{1-i}) \leq h_n/2\}] \\
&\geq \sum_{\substack{i=1,\ldots,l(n):\\ n_i=1}} \sum_{k\geq 0} 2\left[2^{i-1}\int_0^{h_n/2} x^\alpha dx - \left(\frac{h_n}{2}\right)^\alpha\right] \\
&= \frac{1}{\alpha+1}\left(\frac{h_n}{2}\right)^\alpha [nh_n - 2(\alpha+1)l(n)],
\end{aligned}
$$

which establishes (27).

Acknowledgments. The author would like to thank Professor J. Kuelbs for some very helpful discussions, and also two anonymous referees for comments that helped improve the exposition.

REFERENCES

[1] R. AGRAWAL, *Adaptive control of i.i.d. processes and Markov chains with a multidimensional control set*, working paper, Dept. of Electrical and Computer Engineering, Univ. of Wisconsin–Madison, 1992.

[2] R. AGRAWAL, M. HEGDE, AND D. TENEKETZIS, *Asymptotically efficient adaptive allocation rules for the multi-armed bandit problem with switching cost*, IEEE Trans. Automat. Control, AC-33 (1988), pp. 899–906.

[3] ———, *Multi-armed bandit problems with multiple plays and switching cost*, Stochastics Stochastic Rep., 29 (1990), pp. 437–459.

[4] R. AGRAWAL AND D. TENEKETZIS, *Certainty equivalence control with forcing: Revisited*, Systems Control Lett., 13 (1989), pp. 405–412.

[5] R. AGRAWAL, D. TENEKETZIS, AND V. ANANTHARAM, *Asymptotically efficient adaptive allocation schemes for controlled i.i.d. processes: Finite parameter space*, IEEE Trans. Automat. Control, AC-34 (1989), pp. 258–267.

[6] V. ANANTHARAM, P. VARAIYA, AND J. WALRAND, *Asymptotically efficient allocation rules for the multiarmed bandit problem with multiple plays: Part* I: *IID rewards*, IEEE Trans. Automat. Control, 32 (1987), pp. 968–975.

[7] ———, *Asymptotically efficient allocation rules for the multiarmed bandit problem with multiple plays: Part* II: *Markovian rewards*, IEEE Trans. Automat. Control, 32 (1987), pp. 975–982.

[8] T. S. CHIANG, C. R. HWANG, AND S. J. SHEU, *Diffusions for global optimization in* $\mathbb{R}^n$, SIAM J. Control Optim., 25 (1987), pp. 737–752.

[9] M. CSÖRGŐ AND P. RÉVÉSZ, *Strong Approximations in Probability and Statistics*, Academic Press, New York, 1981.

[10] A. DE ACOSTA AND J. KUELBS, *Limit theorems for moving averages of independent random vectors*, Z. Wahrscheinlichkeitstheorie verw. Gebiete, 64 (1983), pp. 67–123.

[11] L. P. DEVROYE, *The uniform convergence of nearest neighbor regression function estimators and their application in optimization*, IEEE Trans. Inform. Theory, IT-24 (1978), pp. 142–151.

[12] V. FABIAN, *Stochastic approximation*, in Optimizing Methods in Statistics, J. S. Rustagi, ed., Academic Press, New York, 1971, pp. 439–470.

[13] ———, *Polynomial estimation of regression functions with the supremum norm error*, Ann. Statist., 16 (1988), pp. 1345–1368.

[14] S. B. GELFAND AND S. MITTER, *Recursive stochastic algorithms for global optimization in* $\mathbb{R}^d$, SIAM J. Control Optim., 29 (1991), pp. 999–1018.

[15] S. GEMAN AND C. R. HWANG, *Diffusions for global optimization*, SIAM J. Control Optim., 24 (1986), pp. 1031–1043.

[16] W. HÄRDLE, *Applied Nonparametric Regression*, Cambridge University Press, Cambridge, 1990.

[17] W. HÄRDLE, P. JANSSEN, AND R. SERFLING, *Strong uniform consistency rates for estimators of conditional functionals*, Ann. Statist., 16 (1988), pp. 1428–1449.

[18] W. HÄRDLE AND S. LUCKHAUS, *Uniform consistency of a class of regression function estimators*, Ann. Statist., 12 (1984), pp. 612–623.

[19] J. HOFFMAN-JØRGENSEN, *Sums of independent Banach space valued random variables*, Studia Math., 52 (1974), pp. 159–186.

[20] H. J. KUSHNER, *Asymptotic global behavior for stochastic approximation and diffusions with slowly decreasing noise effects*, SIAM J. Appl. Math., 47 (1987), pp. 169–185.

[21] H. J. KUSHNER AND D. S. CLARK, *Stochastic Approximation Methods for Constrained and Unconstrained Systems*, Springer-Verlag, Berlin, New York, 1978.

[22] T. L. LAI AND H. ROBBINS, *Asymptotically optimal allocation of treatments in sequential experiments*, in Design of Experiments, T. J. Santer and A. J. Tamhane, eds., Marcel Dekker, New York, 1984, pp. 127–142.

[23] ———, *Asymptotically efficient adaptive allocation rules*, Adv. in Appl. Math., 6 (1985), pp. 4–22.

[24] Y. P. MACK AND B. W. SILVERMAN, *Weak and strong uniform consistency of kernel regression estimates*, Z. Wahrscheinlichkeitstheorie verw. Gebiete, 61 (1982), pp. 405–415.

[25] M. B. NEVEL'SON AND R. Z. HAS'MINSKII, *Stochastic Approximation and Recursive Estimation*, American Mathematical Society, Providence, RI, 1973.

[26] H. ROBBINS, *Some aspects of sequential design of experiments*, Bull. Amer. Math. Soc., 55 (1952), pp. 527–535.

[27] C. J. STONE, *Optimal global rates of convergence for nonparametric regression*, Ann. Statist., 10 (1982), pp. 1040–1053.

[28] S. YAKOWITZ AND W. LOWE, *Nonparametric bandit methods*, Ann. Oper. Res., 28 (1991), pp. 297–312.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 33, No. 6, pp. 1952–1977, November 1995

© 1995 Society for Industrial and Applied Mathematics

017

FORMS OF OPTIMAL SOLUTIONS FOR SEPARATED CONTINUOUS LINEAR PROGRAMS*

MALCOLM C. PULLAN†

Abstract. This paper discusses the nature of optimal solutions for a class of continuous linear programs called separated continuous linear programs. It is shown that under various different assumptions on the problem data there exist optimal solutions that are piecewise constant, piecewise polynomial, or, more generally, piecewise analytic. These results are reminiscent of bang-bang results in linear optimal control.

Key words. continuous linear programming, linear optimal control, bang-bang solutions

AMS subject classifications. 49J30, 49N05, 90C45

1. Introduction. In 1953, Bellman [6] introduced a class of optimization problems which he called *bottleneck problems.* These problems have now become known as *continuous linear programs* because they can be formulated as linear programs having variables that are functions of time as follows:

$$\begin{aligned} \text{CLP:}\quad & \text{maximize} && \int_0^T c(t)^T x(t)\,dt \\ & \text{subject to} && B(t)x(t) + \int_0^t K(s,t)x(s)\,ds \le b(t), \\ & && x(t) \ge 0, \qquad t \in [0,T], \end{aligned}$$

with $x(t)$, $c(t)$, and the elements of $B(t)$ and $K(s,t)$ bounded measurable functions.

The usual way to solve these problems is by discretization (see, for example, Buie and Abrham [7]); however, a number of authors have tried to solve this problem by extending the simplex method for finite linear programming. This was first attempted by Lehman [15] and extended by Drews [8], Hartberger [9], and Segers [22]. The most comprehensive, but still incomplete, solution method based on the simplex method is that by Perold [17], later followed up by Anstreicher [4].

In this paper we will be considering the following subclass of CLP called *separated continuous linear programs*, first introduced by Anderson [1] in an attempt to model job-shop scheduling problems:

$$\begin{aligned} \text{SCLP:}\quad & \text{minimize} && \int_0^T c(t)^T x(t)\,dt \\ (1)\qquad & \text{subject to} && \int_0^t Gx(s)\,ds + y(t) = a(t), \\ (2)\qquad & && Hx(t) + z(t) = b(t), \\ & && x(t), y(t), z(t) \ge 0, \qquad t \in [0,T]. \end{aligned}$$

Here $x(t)$, $z(t)$, $b(t)$, and $c(t)$ are bounded measurable functions and $y(t)$ and $a(t)$ are absolutely continuous functions. The dimensions of $x(t)$, $y(t)$, and $z(t)$ are n_1, n_2, and n_3, respectively. We let $\omega(t)$ denote a complete set of variables for SCLP, i.e.,

* Received by the editors April 23, 1993; accepted for publication (in revised form) July 14, 1994.

† Judge Institute of Management Studies, Mill Lane, Cambridge CB2 1RX, United Kingdom.

$\omega(t)^T = (x(t)^T, y(t)^T, z(t)^T)$. SCLP has an alternative formulation as a linear optimal control problem with state positivity constraints but without feedback as follows:

$$\begin{array}{lll} & \text{minimize} & \int_0^T [p(t)^T x(t) + q(t)^T y(t)]\, dt \\ (3) & \text{subject to} & \dot{y}(t) = Bx(t) + v(t), \\ & & H(t)x(t) \leq b(t), \\ & & x(t), y(t) \geq 0, \qquad t \in [0,T], \end{array}$$

where $v(t) = -\dot{a}(t)$ and $B = -G$. The objective functions can be seen to be equivalent by integrating by parts and substituting in (3).

The main existing theoretical results relating to SCLP may be found in Anderson et al. [2]. These consist of three key results, which we now summarise.

THEOREM 1.1. *If the feasible region for SCLP is nonempty and bounded, then there exists an optimal solution for SCLP at an extreme point of the set of feasible solutions (or simply, at an* extreme point solution).

THEOREM 1.2. *A feasible solution, $\omega(t)$, for SCLP is an extreme point solution if and only if the columns of the matrix*

$$K = \begin{bmatrix} G & I & 0 \\ H & 0 & I \end{bmatrix}$$

corresponding to the support of $\omega(t)$ (that is, i such that $\omega_i(t) > 0$) are linearly independent for almost all $t \in [0,T]$.

THEOREM 1.3. *Suppose that $a(t)$ and $c(t)$ are piecewise linear (but with $a(t)$ continuous) and that $b(t)$ is piecewise constant on $[0,T]$. Suppose also that the feasible region for SCLP is nonempty and bounded, then there exists an optimal solution for SCLP with $x(t)$ piecewise constant on $[0,T]$.*

The statements of these results make use of the following standard definitions which will also be used throughout this paper.

DEFINITION 1.

1. *A set $P = \{t_0, \ldots, t_m\}$ is said to be a* partition *of $[a,b]$ if*

$$a = t_0 < t_1 < \ldots < t_m = b.$$

2. *We say that $f : [a,b] \to \mathbb{R}^n$ is* analytic on a neighbourhood of $[a,b]$ *(or $[a,b)$) if there exists $\varepsilon > 0$ and an analytic function $g : (a-\varepsilon, b+\varepsilon) \to \mathbb{R}^n$ such that $f(t) = g(t)$ for all $t \in [a,b]$ (respectively, $[a,b)$).*
3. *We say that $f : [a,b] \to \mathbb{R}^n$ is* piecewise analytic *on $[a,b]$ if there exists a partition $P = \{t_0, \ldots, t_m\}$ of $[a,b]$ such that $f(t)$ is analytic on a neighbourhood of $[t_{i-1}, t_i)$ for $i = 1, \ldots, m$. The elements of the smallest such partition P (excluding a and, if f is continuous at b, b) are called the* breakpoints.

We use similar definitions for piecewise constant, linear, polynomial, and continuous.

In a recent paper by Anderson and Philpott [3] the following result was also proved.

THEOREM 1.4. *Suppose that $a(t)$ and $b(t)$ are piecewise analytic (but with $a(t)$ continuous) and that $c(t)$ is piecewise constant on $[0,T]$. Suppose also that the set $\{\, \xi : H\xi \leq b(t),\ \xi \geq 0 \,\}$ is bounded for each $t \in [0,T]$ and that the feasible region for SCLP is nonempty; then there exists an optimal solution for SCLP with $x(t)$ piecewise analytic on $[0,T]$.*

Results similar to Theorems 1.3 and 1.4 are well known in the context of linear optimal control and are often termed *bang-bang* results. Bang-bang results are results that give conditions under which an optimal solution can be found that is "bang-bang," that is, where the solution is always on the boundary of the feasible region and, moreover, the number of breakpoints is finite. For instance, many bang-bang results exist for the linear optimal control version of SCLP with feedback, i.e., with (3) replaced by

$$\dot{y}(t) = A(t)y(t) + B(t)x(t) + v(t)$$

but without state-positivity constraints, i.e., without $y(t) \geq 0$ (see, for example, Lee and Markus [14] for a sample of such results).

When state positivity is imposed, the results seem much harder to establish (and, in fact, may not even be true), although many have conjectured that such results are true for various instances of CLP (e.g., Perold [17] and Tyndall [23]). This is now known not to be true for all instances of CLP (see, for example, Ilyotovich [10]). In fact we present simple counterexamples in §5 to show that this is not true for SCLP without sufficiently well behaved problem data (Examples 5.1 and 5.2). However, some authors have attempted to give conditions on CLP that guarantee an optimal solution with a finite number of breakpoints, with limited success, and these include Ilyotovich [10], Jasiulek [11], and Jóhannesson and Hanson [12] (and of course Anderson et al. [2] and Anderson and Philpott [3] for SCLP).

There are many reasons why results that guarantee optimal solutions with a finite number of breakpoints for either CLP, SCLP, or linear optimal control problems are important. First, from the point of view of practical problems, it is useful to have such results, as a practical problem with an optimal solution having an infinite number of breakpoints would not be very worthwhile. Second, such results may enable useful duality results to be established. For instance, Theorem 1.3 was used in Pullan [19] to establish a strong duality result for SCLP under the conditions of Theorem 1.3. Also in the context of linear optimal control, some authors, such as Köhler [13] and Maurer [16], have developed maximum principles given that an optimal solution exists that is piecewise continuous. However, these authors did not establish conditions for the existence of such optimal solutions and thus it is not known under what conditions their results apply.

For the problem SCLP there is a third reason why results guaranteeing an optimal solution with a finite number of breakpoints are important. This reason is from an algorithmic point of view. If it is known that an optimal solution exists with a finite number of breakpoints and, moreover, that this solution is an extreme point solution and also that the maximum number of breakpoints is bounded, then the problem of solving SCLP becomes a finite one. The reason for this is as follows. By Theorem 1.2, the value of an extreme point solution for SCLP at any time (or, at least, at almost all times) is uniquely determined by the choice of nonzero variables at that time (see the proof of Lemma 2.1). Therefore, there are two main unknowns one wishes to determine in finding an optimal extreme point solution for SCLP. The first of these is the set of variables that are positive at any particular time, and the second, the set of breakpoints. Thus, if one can show that there is an upper bound on the number of breakpoints of an optimal extreme point solution for a given SCLP problem, then the problem of finding the optimal solution, given that the feasible region of SCLP is nonempty and bounded, becomes a finite one. Hence, in theory, one may construct an algorithm that terminates in a finite number of steps. This was part of the motivation in Pullan [19] for developing an algorithm for solving SCLP under the assumption of piecewise linear and continuous $a(t)$, piecewise constant $b(t)$, and piecewise linear $c(t)$.

In this paper we consider the form of optimal solutions under different conditions to either Theorems 1.3 or 1.4. In particular, in §3 we consider the case of SCLP with piecewise linear and continuous $a(t)$, piecewise constant $b(t)$, and piecewise analytic $c(t)$. In this case, given that the feasible region for SCLP is also nonempty and bounded, we prove that an optimal extreme point solution for SCLP exists with $x(t)$ piecewise constant (Theorem 3.3). This result also gives an upper bound on the number of breakpoints for an optimal solution for a given SCLP problem. In §4 we then consider SCLP with piecewise analytic $a(t)$, $b(t)$, and $c(t)$. The result in this case is that SCLP has an optimal extreme point solution that is piecewise analytic, given again that the feasible region for SCLP is both nonempty and bounded (Theorem 4.3). The proof uses Theorem 3.3 as a starting point and considers a problem with analytic right-hand sides as a limit of a sequence of problems where $a(t)$ is piecewise linear and continuous and $b(t)$ is piecewise constant. The proof also uses ideas from the proof of Theorem 1.4 in Anderson and Philpott [3]. Using a simple observation, however, we are able to weaken the assumption used in [3] that the set $\{\, \xi : H\xi \leq b(t),\ \xi \geq 0 \,\}$ is bounded for each $t \in [0, T]$ to the assumption that just the feasible region for SCLP itself is bounded. As a conclusion to the section we give a simple breakpoint condition, namely that $\dot{c}(t)^T x(t)$ must not increase at a breakpoint (Theorem 4.4). As with Theorem 3.3, Theorem 4.3 also gives an upper bound on the number of breakpoints in an optimal extreme point solution for a given SCLP problem. However, as this bound is slightly cumbersome to state, it is only given implicitly in the proof. Although the results of §3 are just special cases of those in §4, it is convenient to consider these as separate cases in order to bring greater clarity and understanding to the proofs.

It is worth noting that neither Theorem 1.3 nor Theorem 1.4 guarantees an optimal solution of the appropriate form that is also an extreme point. However, the results in this paper do guarantee such a solution.

As a conclusion to this paper we then postulate more general results in §5. First, however, we present simple counterexamples to show that the assumption of analyticity of the problem data cannot be weakened, e.g., to continuous or n-times differentiable for some n, to ensure that an optimal solution exists with only a finite number of breakpoints (Examples 5.1 and 5.2). These examples, however, suggest possible extensions to the results in previous sections. Finally in §5, we comment on what happens if $a(t)$ and $b(t)$ are chosen from a more restrictive class of functions than that of piecewise analytic functions, e.g., piecewise polynomial (Theorem 5.1).

Before beginning the discussion we introduce some more standard definitions and notations that will be used throughout this paper.

DEFINITION 2. 1. *For any set S, we use the notation $|S|$ to denote its cardinality.*

2. *We use the notation a.e. to mean almost everywhere with respect to the standard Lebesgue measure on $\mathbb{R}$.*
3. *Let f be any real-valued function. We use the notations $f(t-)$ to denote $\lim_{s\uparrow t} f(s)$ and $f(t+)$ to denote $\lim_{s\downarrow t} f(s)$ when these limits exist.*
4. *For any $\zeta \in \mathbb{R}^n$ we define $\operatorname{sgn}(\zeta) \in \mathbb{R}^n$ by*

$$(\operatorname{sgn}(\zeta))_i = \begin{cases} -1, & \zeta_i < 0, \\ 0, & \zeta_i = 0, \\ 1, & \zeta_i > 0. \end{cases}$$

5. *For any $\zeta \in \mathbb{R}^n$, $\zeta \geq 0$, we use the notation*

$$\operatorname{supp}(\zeta) = \{\, i : \zeta_i > 0 \,\}$$

to denote the support of ζ.

We now begin the discussion by establishing some preliminary results that will be used throughout the main body of this paper.

2. Preliminary results.

Before we begin the development of the main results in this paper, it is useful to establish some preliminary notation and results that can be frequently referred to. We begin by the introduction of some notation relating to the matrix K used to describe extreme point solutions in Theorem 1.2.

DEFINITION 3. *We define*

$$K = \begin{bmatrix} G & I & 0 \\ H & 0 & I \end{bmatrix}.$$

Let B be a basis matrix for K.

1. *Let $\omega(t)^T = (x(t)^T, y(t)^T, z(t)^T)$ be a set of variables for SCLP. We let $x_B(t)$ denote the elements of $x(t)$ corresponding to those of the first n_1 columns of K that are also in B, arranged in the same order as the columns of B.*
2. *Let $\rho(t)$ be a solution to $B\rho(t) = d(t)$ for some $d(t)$ (that is, $\rho(t) = B^{-1}d(t)$). We use the notation $\rho_x(t)$ to denote the elements of $\rho(t)$ corresponding to those of the first n_1 columns of K that are also in B, arranged in the same order as the columns of B.*

Using this notation we now present a simple lemma that is essentially contained in Anderson et al. [2]. The result shows that if $\omega(t)$ is an extreme point solution for SCLP, then $x(t)$ can only take the values of a finite number of functions, these functions being linear combinations of $\dot{a}(t)$ and $b(t)$. It is this finite nature of any extreme point solution which allows us to show that an optimal extreme point exists with only a finite number of breakpoints, given that the problem data are piecewise analytic.

LEMMA 2.1. *Let $a(t)$ be any absolutely continuous function and $b(t)$ be any bounded measurable function on $[0, T]$. Let $B^{(1)}, \ldots, B^{(L)}$ be the basis matrices for K, and let*

$$\rho^{(i)}(t) = B^{(i)^{-1}} \begin{bmatrix} \dot{a}(t) \\ b(t) \end{bmatrix}.$$

Define $x^{(i)}(t)$ by $x^{(i)}_{B^{(i)}}(t) = \rho^{(i)}_x(t)$ with the other components of $x^{(i)}(t)$ set to zero. Let $\omega(t)$ be any extreme point solution for SCLP; then for almost all $t \in [0, T]$, $x(t) = x^{(i_t)}(t)$ for some i_t.

Proof. The constraints of SCLP, (1) and (2), are equivalent to

$$K \begin{bmatrix} x(t) \\ \dot{y}(t) \\ z(t) \end{bmatrix} = \begin{bmatrix} \dot{a}(t) \\ b(t) \end{bmatrix},$$
$$y(0) = a(0),$$

a.e. on $[0, T]$. However, by Theorem 1.2, we know that the columns of K corresponding to the nonzero variables of $\omega(t)$ are linearly independent for almost all $t \in [0, T]$. Thus for almost all $t \in [0, T]$, the nonzero variables (and possibly some zero variables) of $(x(t)^T, \dot{y}(t)^T, z(t)^T)^T$ are given by $\rho^{(i_t)}$ for some i_t. This establishes the result. □

Throughout the course of the proofs in the following sections, we will frequently derive our results for SCLP with problem data of a particular form and then establish the result in the general case by considering SCLP with the general problem data as a limit of a sequence of SCLP problems with the less general problem data. In such cases it is useful to know that the limit of optimal extreme point solutions for the simpler problems will be an optimal extreme point solution for the general problem. This is the content of the next result. We note that the result we present is by no means the most general result available. We only give the result for the circumstances required in this paper.

LEMMA 2.2. *Let $a^{(n)}$ be absolutely continuous and $b^{(n)}$ and $c^{(n)}$ be piecewise continuous on $[0,T]$ with $a^{(n)}(t) \geq a(t)$ and $b^{(n)}(t) \geq b(t)$ on $[0,T]$ for each $n \in \mathbb{N}$ and $a^{(n)} \to a$, $b^{(n)} \to b$ and $c^{(n)} \to c$ uniformly on $[0,T]$. Let SCLP$_n$ be the problem SCLP with a, b, and c replaced by $a^{(n)}$, $b^{(n)}$, and $c^{(n)}$, respectively. Suppose that $\omega_n(t)$ is an optimal extreme point solution for SCLP$_n$ for $n \in \mathbb{N}$ and that $x_n(t) \to x(t)$ a.e. on $[0,T]$. Suppose also that $x(t)$ and $x_n(t)$ for each n are piecewise continuous. Define $\omega(t)$ from the constraints of SCLP, i.e., from* (1) *and* (2). *Then $\omega(t)$ is an optimal extreme point solution for SCLP.*

Proof. By Lebesgue's dominated convergence theorem (see, for example, Rudin [21]) we have $y_n(t) \to y(t)$ for all $t \in [0,T]$. Let s be such that $x_n(s) \to x(s)$ and such that s is not a point of discontinuity for x or x_n for any n. Then $\omega_n(s) \to \omega(s)$ and thus $\omega(s) \geq 0$. Hence $\omega(t) \geq 0$ a.e. on $[0,T]$ and so $\omega(t)$ is feasible for SCLP.

Now suppose that $\omega(t)$ is not an extreme point solution for SCLP. Then by Theorem 1.2 there exists a set S of nonzero measure such that the columns of K corresponding to $\operatorname{supp}(\omega(t))$ for $t \in S$ are linearly dependent. Choose $s \in S$ such that $x_n(s) \to x(s)$ and such that s is not a point of discontinuity of x or x_n for any n. Now, as noted above, we have $\omega_n(s) \to \omega(s)$. Hence there exists n such that

$$\operatorname{supp}(\omega(s)) \subseteq \operatorname{supp}(\omega_n(s)).$$

Now ω_n is continuous at s and so there exists (α, β) such that $s \in (\alpha, \beta)$ and

$$\operatorname{supp}(\omega_n(s)) \subseteq \operatorname{supp}(\omega_n(t)),$$

for all $t \in (\alpha, \beta)$. Thus $\operatorname{supp}(\omega(s)) \subseteq \operatorname{supp}(\omega_n(t))$ for all $t \in (\alpha, \beta)$, and so by Theorem 1.2, $\omega_n(t)$ is not an extreme point solution for SCLP$_n$. This contradiction establishes that $\omega(t)$ is an extreme point solution for SCLP.

Finally, suppose that $\omega(t)$ is not optimal for SCLP. Then there exists $\bar{\omega}(t)$ feasible for SCLP satisfying

$$\int_0^T c(t)^T \bar{x}(t)\, dt < \int_0^T c(t)^T x(t)\, dt.$$

Now $a^{(n)}(t) \geq a(t)$ and $b^{(n)}(t) \geq b(t)$ on $[0,T]$, and so $\bar{\omega}(t)^T + (0^T, (a^{(n)}(t) - a(t))^T, (b^{(n)}(t) - b(t))^T)$ forms a feasible solution for SCLP$_n$ for each n. But by Lebesgue's dominated convergence theorem again we then have

$$\int_0^T c(t)^T \bar{x}(t)\, dt < \int_0^T c(t)^T x^{(n)}(t)\, dt,$$

for some n, contradicting the optimality of $\omega^{(n)}(t)$ in SCLP$_n$. Thus $\omega(t)$ must be optimal for SCLP. □

Our final result in this section is quite unrelated to the previous ones and concerns the zeros of a linear combination of a finite set of analytic functions on a compact interval $[a,b]$. While it is well known that such a linear combination will only have a finite number of zeros (see, for example, Apostol [5, Thm. 16.25]), we show that the number of zeros in this linear combination is dependent only on the functions involved and not the particular scalars chosen. This result will be useful in establishing that a sequence of SCLP problems with problem data all of some similar general form will have optimal solutions with a uniform upper bound on the number of breakpoints.

LEMMA 2.3. *Let $f : [a,b] \to \mathbb{R}^n$ be a function analytic on a neighbourhood of $[a,b]$. Then there exists $M(f)$ $(< \infty)$ such that for all $\lambda \in \mathbb{R}^n$, if*

$$S(\lambda, f) = \{\, t \in [a,b] : \lambda^T f(t) = 0 \,\},$$

then either $S(\lambda,f) = [a,b]$ or $|S(\lambda,f)| \le M(f)$.

Proof. We prove the result by induction. For $n = 1$ the result is well known (see, for example, Apostol [5, Thm. 16.25]).

Let n be given. Suppose that for all $m < n$ and for all $f : [a,b] \to \mathbb{R}^m$ analytic on a neighbourhood of $[a,b]$ there exists $M(f) < \infty$ such that for all $\lambda \in \mathbb{R}^m$ we have either $S(\lambda,f) = [a,b]$ or $|S(\lambda,f)| \le M(f)$, where $S(\lambda,f)$ is as given in the statement of the theorem. Let $f : [a,b] \to \mathbb{R}^n$ be a given function analytic on a neighbourhood of $[a,b]$. Define $g, h : [a,b] \to \mathbb{R}^{n-1}$ by $g_i = \dot{f}_n f_i - f_n \dot{f}_i$ and $h_i = f_i$ for $i = 1, \ldots, n-1$. We now define

$$M = \max\{(M(h)+1)(M(g)+2)+1, M(f_n)\},$$

where $M(g)$ and $M(h)$ are the required bounds for g and h given by the inductive assumption. We claim that $M(f) = M$ satisfies the requirements of the lemma.

Let $\lambda \in \mathbb{R}^n$. Suppose that $\lambda^T f \not\equiv 0$ (i.e., $\lambda^T f(t) \neq 0$ for some $t \in [a,b]$). Let μ denote the first $n-1$ components of λ. Suppose that $\mu^T h \equiv 0$. Then $\lambda^T f(t) = 0$ if and only if $\lambda_n f_n(t) = 0$. Hence either $S(\lambda,f) = [a,b]$ or $|S(\lambda,f)| \le M(f_n) \le M$.

Similarly, if $\lambda_n f_n \equiv 0$, then either $S(\lambda,f) = [a,b]$ or $|S(\lambda,f)| \le M(h) \le M$.

Now suppose that $\mu^T h \not\equiv 0$ and $\lambda_n f_n \not\equiv 0$. Let $P = \{t_0, \ldots, t_m\}$ be the partition of $[a,b]$ consisting of all the zeros of the analytic function $\mu^T h$ and the points a and b. Then $m \le M(h)+1$ (a and b may not be zeros of $\mu^T h$). Consider the interval (t_{i-1}, t_i). Let $t \in (t_{i-1}, t_i)$, then

$$\lambda^T f(t) = 0 \iff \frac{f_n(t)}{\mu^T h(t)} = -\frac{1}{\lambda_n}. \tag{4}$$

Now either (4) has at most one solution in (t_{i-1}, t_i) or it has more than one solution in (t_{i-1}, t_i). Assume the latter. As $\lambda^T f \not\equiv 0$ there can only be a finite number of solutions to (4). Let s_1 and s_2 be two consecutive solutions to (4) in (t_{i-1}, t_i). By Rolle's theorem there exists $t \in (s_1, s_2)$ such that

$$\mu^T g(t) = \dot{f}_n(t)\mu^T h(t) - f_n(t)\mu^T \dot{h}(t) = 0. \tag{5}$$

Clearly (5) cannot be true for all $t \in [a,b]$; otherwise we would have $\lambda^T f \equiv 0$ on (t_{i-1}, t_i). Hence (5) only has a finite number of solutions. Thus by the inductive assumption,

$$|\{\, t \in (t_{i-1}, t_i) : (5) \text{ holds} \,\}| \le M(g).$$

Thus (and this includes the case where (4) has at most one solution in (t_{i-1}, t_i)),

$$|\{\, t \in (t_{i-1}, t_i) : (4) \text{ holds} \,\}| \leq M(g) + 1.$$

Now $m \leq M(h) + 1$, and so $|S(\lambda, f) - P| \leq (M(h) + 1)(M(g) + 1)$. As P contains $m+1 \leq M(h)+2$ points, we have $|S(\lambda, f)| \leq M$. This proves the result by induction. □

We may now begin the development of the main results of this paper.

3. SCLP with analytic costs. In this section we consider the case of SCLP with $a(t)$ piecewise linear (and continuous), $b(t)$ piecewise constant, and $c(t)$ piecewise analytic on $[0, T]$. Under the assumption that the feasible region for SCLP is nonempty and bounded, we show that there exists an optimal extreme point solution with $x(t)$ piecewise constant (Theorem 3.3). Our proof is based on the proof of Theorem 1.3 contained in Anderson et al. [2]. As with [2], we begin by concentrating on the case where the problem data contain no discontinuities and present the proof in stages.

We outline the stages of the proof as follows. We begin with the assumption that the components of the costs and their derivatives are concave, i.e., that the components of $c(t)$ have nonpositive second and third derivatives. Under this assumption we derive a certain convexity property of all optimal solutions (Lemma 3.1). This is a simple extension of a similar result in [2] (namely, Theorem 3). We now recall Lemma 2.1, which states that in the case of constant $\dot{a}(t)$ and $b(t)$, there are only a finite number of possible values, $x^{(1)}, \ldots, x^{(L)}$, for an extreme point. Now as $c(t)$ is analytic, we either have for some k and j that $\dot{c}(t)^T x^{(k)} = \dot{c}(t)^T x^{(j)}$ for all $t \in [0, T]$, or there is a partition $P = \{t_0, \ldots, t_n\}$ of $[0, T]$ such that for all $k \neq j$ we have $\dot{c}(t)^T x^{(k)} \neq \dot{c}(t)^T x^{(j)}$ for all $t \in [0, T] - P$. In the latter case we consider a typical subinterval (t_{i-1}, t_i) of the partition P. Using the convexity property established in Lemma 3.1 it is then shown in Lemma 3.2 that $\dot{c}(t)^T x(t)$ is decreasing at a breakpoint. Hence there can be at most L breakpoints in $[t_{i-1}, t_i)$ and hence nL overall.

In the case where for some $k \neq j$ we have $\dot{c}(t)^T x^{(k)} = \dot{c}(t)^T x^{(j)}$ for all $t \in [0, T]$, we then construct a sequence of piecewise constant solutions for SCLP whose costs approach the optimal value of SCLP. This then completes the proof of Lemma 3.2, which guarantees an optimal solution with $x(t)$ piecewise constant in the case of costs whose second and third derivatives are nonpositive.

To complete the proof of the result for general analytic costs it is required only to transform a general problem into an equivalent one with costs that have nonpositive second and third derivatives. This simple transformation is contained in Theorem 3.3.

We now begin by establishing the convexity property of all optimal solutions under the restricted assumption of costs with nonpositive second and third derivatives.

LEMMA 3.1. *Suppose that $a(t)$ is linear and $b(t)$ is constant on $[0, T]$. Suppose also that $c(t)$ is analytic on a neighbourhood of $[0, T]$ and has nonpositive second and third derivatives. Let $\omega(t)$ be optimal for SCLP; then*

$$q(t) = -\dot{c}(t)^T \int_0^t x(s)\, ds$$

is convex on $[0, T]$.

Proof. Suppose that $\omega(t)$ is optimal for SCLP. Suppose the result is false; then as q is continuous there is some interval $[t_1, t_2] \subseteq [0, T]$ such that

$$q(t) > r(t), \quad t \in (t_1, t_2),$$

where

$$r(t) = q(t_1) + \left(\frac{t-t_1}{t_2-t_1}\right)(q(t_2) - q(t_1)), \quad t \in [t_1, t_2].$$

Define

$$\bar{q}(t) = \int_0^t x(s)\,ds, \quad t \in [t_1, t_2],$$

and

$$\bar{r}(t) = \bar{q}(t_1) + \left(\frac{t-t_1}{t_2-t_1}\right)(\bar{q}(t_2) - \bar{q}(t_1)), \quad t \in [t_1, t_2];$$

then $q(t) = -\dot{c}(t)^T\bar{q}(t)$ for $t \in [t_1, t_2]$. We now claim that

$$r(t) \geq -\dot{c}(t)^T\bar{r}(t), \quad t \in [t_1, t_2]. \tag{6}$$

Let $t \in (t_1, t_2)$; then

$$\frac{d^2}{dt^2}(-\dot{c}(t)^T\bar{r}(t)) = -\dddot{c}(t)^T\bar{r}(t) - 2\ddot{c}(t)^T\dot{\bar{r}}(t).$$

But

$$\dot{\bar{r}}(t) = \frac{1}{t_2-t_1}\int_{t_1}^{t_2} x(s)\,ds \geq 0,$$

and so $\ddot{c}(t)^T\dot{\bar{r}}(t) \leq 0$ as $\ddot{c}(t)$ is nonpositive on $[0,T]$. Similarly, $\dddot{c}(t)^T\bar{r}(t) \leq 0$ and hence $-\dot{c}(t)^T\bar{r}(t)$ is convex on (t_1, t_2). Thus $r(t) \geq -\dot{c}(t)^T\bar{r}(t)$ on $[t_1, t_2]$, as claimed, as $r(t)$ is linear on $[t_1, t_2]$ and $r(t) = -\dot{c}(t)^T\bar{r}(t)$ for $t = t_1$ and $t = t_2$.

We now define $\bar{x}(t)$ by

$$\bar{x}(t) = \begin{cases} \dfrac{1}{t_2-t_1}\displaystyle\int_{t_1}^{t_2} x(s)\,ds, & t \in [t_1, t_2), \\ x(t), & \text{otherwise.} \end{cases}$$

Define $\bar{\omega}(t)$ from $\bar{x}(t)$ by the constraints of SCLP, i.e., so that $\bar{x}(t)$, $\bar{y}(t)$, and $\bar{z}(t)$ satisfy (1) and (2) in place of $x(t)$, $y(t)$, and $z(t)$, respectively. Then $\bar{\omega}(t)$ is feasible for SCLP. Comparing objective function values for $\omega(t)$ and $\bar{\omega}(t)$ gives

$$\begin{aligned}
&\int_0^T c(t)^T\bar{x}(t)\,dt - \int_0^T c(t)^T x(t)\,dt \\
&= \int_{t_1}^{t_2} c(t)^T(\bar{x}(t) - x(t))\,dt \\
&= \int_{t_1}^{t_2} c(t)^T\frac{d}{dt}(\bar{r}(t) - \bar{q}(t))\,dt \\
&= -\int_{t_1}^{t_2} \dot{c}(t)^T(\bar{r}(t) - \bar{q}(t))\,dt + c(t_2)^T(\bar{r}(t_2) - \bar{q}(t_2)) - c(t_1)^T(\bar{r}(t_1) - \bar{q}(t_1))
\end{aligned}$$

by integrating by parts. But $\bar{r}(t_1) = \bar{q}(t_1)$ and $\bar{r}(t_2) = \bar{q}(t_2)$ and so

$$\begin{aligned}\int_0^T c(t)^T \bar{x}(t)\, dt - \int_0^T c(t)^T x(t)\, dt &= -\int_{t_1}^{t_2} \dot{c}(t)^T (\bar{r}(t) - \bar{q}(t))\, dt \\ &\leq \int_{t_1}^{t_2} (r(t) - q(t))\, dt\end{aligned}$$

by (6). Hence

$$\int_0^T c(t)^T \bar{x}(t)\, dt - \int_0^T c(t)^T x(t)\, dt < 0,$$

since $r(t) < q(t)$ on (t_1, t_2). This contradicts the assumed optimality of $\omega(t)$. Hence $q(t)$ is convex, and this establishes the result. □

We now use this result, as outlined at the beginning of this section, to establish that SCLP has a piecewise constant optimal solution in the case of costs with nonpositive second and third derivatives. We also explicitly state the upper bound on the number of breakpoints derived for such an optimal solution as this bound will be useful in the next section.

LEMMA 3.2. *Suppose that $a(t)$ is linear and $b(t)$ is constant on $[0, T]$. Suppose also that $c(t)$ is analytic on a neighbourhood of $[0, T]$ and has nonpositive second and third derivatives. If, furthermore, the feasible region for SCLP is nonempty and bounded, then there exists an optimal extreme point solution for SCLP with $x(t)$ piecewise constant on $[0, T]$. Moreover, let $x^{(1)}, \dots, x^{(L)}$ be given by Lemma 2.1, where L is the number of basis matrices for K, and $P = \{t_0, t_1, \dots, t_n\}$ be any partition of $[0, T]$ with $\dot{c}(t)^T x^{(i)} = \dot{c}(t)^T x^{(j)}$ for all $t \in (t_{m-1}, t_m)$ or $\dot{c}(t)^T x^{(i)} \neq \dot{c}(t)^T x^{(j)}$ for all $t \in (t_{m-1}, t_m)$, for each $i \neq j$ and each m. Then such $x(t)$ may be chosen so that for all m, the maximum number of breakpoints of $x(t)$ in $[t_{m-1}, t_m)$ is L.*

Proof. Let $\omega(t)$ be an optimal extreme point solution, which exists by Theorem 1.1 as the feasible region is nonempty and bounded. By Lemma 2.1 there exist constants $x^{(1)}, \dots, x^{(L)}$ such that for almost all $t \in [0, T]$, $x(t) = x^{(i_t)}$ for some i_t. Assume, without loss of generality, that this is true for all $t \in [0, T]$.

Now by Lemma 3.1

$$q(t) = -\dot{c}(t)^T \int_0^t x(s)\, ds$$

is convex on $[0, T]$. Hence (see, for example, Rockafellar [20]), $\dot{q}(t)$ is monotonic increasing a.e. on $[0, T]$; i.e., there exists a set $S_1 \subseteq [0, T]$ of measure T such that $\dot{q}(t)$ is monotonic increasing on S_1. But $\dot{q}(t) = -f(t)$ a.e., where

$$f(t) = \dot{c}(t)^T x(t) + \ddot{c}(t)^T \int_0^t x(s)\, ds.$$

Let

$$S = S_1 \cap \{\, t \in [0, T] : \dot{q}(t) = -f(t) \,\};$$

then f is monotonic decreasing on S. We now show that there is no loss of generality by assuming that f is monotonic decreasing on $[0, T]$ by redefining $x(t)$ on $[0, T] - S$, which has measure zero.

Let $s \in [0,T] - S$. Define

$$\gamma = \sup\{ f(t) : t > s,\ t \in S \}.$$

As f is decreasing on S and S is dense in $[0,T]$, we have a sequence $\{t_k\}_{k=1}^{\infty}$ such that $t_k \downarrow t$, $t_k \in S$ and $\lim_{k\to\infty} f(t_k) = \gamma$. As $x(t)$ takes only a finite number of values on S, there is an l and a subsequence $\{t_{i_k}\}_{k=1}^{\infty}$ such that $x(t_{i_k}) = x^{(l)}$ and $\lim_{k\to\infty} f(t_{i_k}) = \gamma$. Redefine $x(s) = x^{(l)}$; then $f(t) \le f(s)$ for all $t > s$, $t \in S$. If this is done for each $s \in [0,T] - S$, then it is clear that f is monotonic decreasing on $[0,T]$.

We now assume that f is monotonic decreasing on $[0,T]$ and that $x(t)$ takes only the values $x^{(1)}, \ldots, x^{(L)}$. We have two cases to consider. Either for each $k \neq j$ we have

$$\dot{c}(t)^T x^{(k)} \neq \dot{c}(t)^T x^{(j)} \quad \text{for some } t \in [0,T], \tag{7}$$

or there exists $k \neq j$ such that $\dot{c}(t)^T x^{(k)} = \dot{c}(t)^T x^{(j)}$ for all $t \in [0,T]$. We consider the former case first; i.e., (7) holds for all $k \neq j$. Now as $\dot{c}(t)^T x^{(i)}$ is analytic on a neighbourhood of $[0,T]$ for each i, there are only a finite number of $t \in [0,T]$ such that $\dot{c}(t)^T x^{(k)} = \dot{c}(t)^T x^{(j)}$ for some $k \neq j$. Label these t as $t_1, \ldots, t_{n-1}$, and set $t_0 = 0$ and $t_n = T$. Then $\dot{c}(t)^T x^{(k)} \neq \dot{c}(t)^T x^{(j)}$ for all $t \in (t_{i-1}, t_i)$ and for all $k \neq j$ and $i = 1, \ldots, n$.

Consider the interval (t_{i-1}, t_i). Assume without loss of generality that

$$\dot{c}(t)^T x^{(1)} > \dot{c}(t)^T x^{(2)} > \cdots > \dot{c}(t)^T x^{(L)}$$

for $t \in (t_{i-1}, t_i)$. We will show that $x(t)$ takes the values $x^{(j)}$ in this interval with j increasing; i.e., that if $x(s_1) = x^{(k)}$ and $x(s_2) = x^{(j)}$ with $s_1 < s_2$, then $k \le j$. Suppose otherwise, then there is some $s_1 < s_2$ with $x(s_1) = x^{(k)}$ and $x(s_2) = x^{(j)}$ but $k > j$. Let

$$\sigma = \sup\left\{ s : x(t) = x^{(l)} \text{ for some } l > j, \text{ for all } t \in [s_1, s] \right\}.$$

Then $s_1 \le \sigma < s_2$. Define

$$\varepsilon = \frac{1}{2} \min\left\{ \left| \dot{c}(\sigma)^T x^{(l)} - \dot{c}(\sigma)^T x^{(j)} \right| : l \neq j \right\};$$

then since $\dot{c}(t)$ is continuous, we can choose $\delta_1 > 0$ with

$$\left| \dot{c}(\tau_2)^T x^{(l)} - \dot{c}(\tau_1)^T x^{(j)} \right| > \varepsilon$$

for all $l \neq j$ and τ_1, $\tau_2 \in (\sigma - \delta_1, \sigma + \delta_1)$. If we now define

$$d(t) = \ddot{c}(t)^T \int_0^t x(s)\, ds,$$

then $d(t)$ is continuous on $[0,T]$ and $f(t) = \dot{c}(t)^T x(t) + d(t)$. We now choose δ_2 so that

$$|d(\tau_2) - d(\tau_1)| < \frac{\varepsilon}{2},$$

for all τ_1, $\tau_2 \in (\sigma - \delta_2, \sigma + \delta_2)$. Set $\delta = \min\{\delta_1, \delta_2\}$. Now from the definition of σ we can choose $\sigma_1 \in (\sigma - \delta, \sigma]$ with $x(\sigma_1) = x^{(l)}$ for some $l > j$ and $\sigma_2 \in [\sigma, \sigma + \delta)$ with $x(\sigma_2) = x^{(m)}$ for some $m \leq j$. Then

$$f(\sigma_2) - f(\sigma_1) > \varepsilon - |d(\sigma_1) - d(\sigma_2)| > 0,$$

which contradicts the assumption that f is monotonic decreasing.

Hence if $x(s_2) = x^{(j)}$ and $x(s_1) = x^{(k)}$ with $s_2 > s_1$ then $k \leq j$. So if we define

$$U_k = \{\, t \in (t_{i-1}, t_i) : x(t) = x^{(k)} \,\}$$

for any k, and if $U_k \neq \emptyset$ we also define

$$\begin{aligned} u_k &= \inf\{\, t : t \in U_k \,\}, \\ v_k &= \sup\{\, t : t \in U_k \,\}, \end{aligned}$$

then $x(t) = x^{(k)}$ for all $t \in (u_k, v_k)$. Hence, by redefining $x(t)$ at u_k and v_k if necessary, we can see that $x(t)$ is piecewise constant on $[t_{i-1}, t_i)$ and hence, as i was arbitrary, on $[0, T]$. Moreover, $x(t)$ has the maximum number of breakpoints as specified in the statement of the lemma.

Now suppose that (7) does not hold for some $k \neq j$, i.e., that

$$\dot{c}(t)^T x^{(k)} = \dot{c}(t)^T x^{(j)} \quad \text{for all } t \in [0, T].$$

Suppose there exists i and l such that

$$\dot{c}(t)^T x^{(i)} \neq \dot{c}(t)^T x^{(l)} \tag{8}$$

for some $t \in [0, T]$. In this case, let $P = \{t_0, \ldots, t_n\}$ be any partition of $[0, T]$ so that (8) holds for all $t \in [0, T] - P$ and for all such i and l. If, on the other hand, no such i and l exist then let $P = \{0, T\}$.

Let $\{\varepsilon_m(t)\}_{m=1}^{\infty}$ be any sequence of vector-valued functions analytic on a neighbourhood of $[0, T]$ satisfying the following properties:

1. $\varepsilon_m(t) \to 0$ uniformly on $[0, T]$ as $m \to \infty$.
2. $\varepsilon_m(t_i) = 0$ for $i = 0, \ldots, n$.
3. $[\dot{c}(t) + \varepsilon_m(t)]^T x^{(\eta)} \neq [\dot{c}(t) + \varepsilon_m(t)]^T x^{(\zeta)}$ for $t \in [0, T] - P$ and η, $\zeta = 1, \ldots, L$ $(\eta \neq \zeta)$.
4. $[\dot{c}(t) + \varepsilon_m(t)]^T x^{(\eta)} < [\dot{c}(t) + \varepsilon_m(t)]^T x^{(\zeta)}$ for some $t \in (t_{i-1}, t_i)$ if and only if the same is true for all $t \in (t_{i-1}, t_i)$.

It is clear that such functions exist as a suitable sequence of polynomials will do.

Let $\omega_m(t)$ be an optimal extreme point solution for the SCLP problem with $\dot{c}(t)$ replaced by $\dot{c}(t) + \varepsilon_m(t)$. From the above we can choose $\omega_m(t)$ so that $x_m(t)$ is piecewise constant on $[0, T]$. Assume now without loss of generality that

$$[\dot{c}(t) + \varepsilon_m(t)]^T x^{(1)} > [\dot{c}(t) + \varepsilon_m(t)]^T x^{(2)} > \cdots > [\dot{c}(t) + \varepsilon_m(t)]^T x^{(L)}$$

for $t \in (t_0, t_1)$ and $m \in \mathbb{N}$. Then by the above, we can find $p_{m,j} \geq 0$ with $\sum_{j=0}^{L} p_{m,j} = t_1 - t_0$ and

$$x_m(t) = x^{(i)} \quad \text{a.e. on} \quad \left[t_0 + \sum_{j=0}^{i-1} p_{m,j}, t_0 + \sum_{j=0}^{i} p_{m,j} \right).$$

As $p_{m,j}$ is bounded for all m and j (by $t_1 - t_0$), we may choose a subsequence, $\{p_{m_k,j}\}_{k=1}^{\infty}$, of $\{p_{m,j}\}_{m=1}^{\infty}$ for each j, converging to p_j, say. Define $\bar{x}(t)$ on $[t_0, t_1)$ by

$$\bar{x}(t) = x^{(i)}, \quad t \in \left[t_0 + \sum_{j=0}^{i-1} p_j, t_0 + \sum_{j=0}^{i} p_j \right).$$

We now repeat the above process for the subsequence $\{x_{m_k}\}_{k=1}^{\infty}$ of $\{x_m\}_{m=1}^{\infty}$ for the next interval $[t_1, t_2)$. Continuing in this manner will produce $\bar{x}(t)$ that is piecewise constant on $[0, T]$ and also the limit of some subsequence of $\{x_m\}_{m=1}^{\infty}$. Hence by Lemma 2.2, $\bar{\omega}(t)$, defined from $\bar{x}(t)$ by the constraints of SCLP, is an optimal extreme point solution for SCLP. Moreover, $x(t)$ has the maximum number of breakpoints as specified in the statement of the lemma. This establishes the result. $\square$

By transforming a general problem into one where the costs have nonpositive second and third derivatives, we may now establish the main result of this section.

THEOREM 3.3. *Suppose that $a(t)$ is piecewise linear and continuous, $b(t)$ is piecewise constant, and $c(t)$ is piecewise analytic on $[0, T]$. Suppose also that the feasible region for SCLP is nonempty and bounded; then there exists an optimal extreme point solution for SCLP with $x(t)$ piecewise constant on $[0, T]$. Moreover, let $x^{(1)}, \ldots, x^{(L)}$ be given by Lemma 2.1, where L is the number of basis matrices for K, and $P = \{t_0, t_1, \ldots, t_n\}$ be any partition of $[0, T]$ containing the breakpoints of $a(t)$, $b(t)$, and $c(t)$ and with $\dot{c}(t)^T x^{(i)} = \dot{c}(t)^T x^{(j)}$ for all $t \in (t_{m-1}, t_m)$ or $\dot{c}(t)^T x^{(i)} \neq \dot{c}(t)^T x^{(j)}$ for all $t \in (t_{m-1}, t_m)$ for each $i \neq j$ and each m. Then such $x(t)$ may be chosen so that for all m, the maximum number of breakpoints of $x(t)$ in $[t_{m-1}, t_m)$ is L.*

Proof. We begin by considering the case where the problem data contain no breakpoints, i.e., that $a(t)$ is linear, $b(t)$ is constant, and $c(t)$ is analytic on a neighbourhood of $[0, T]$. We transform the problem into one where the costs have nonpositive second and third derivatives. Let

$$N = \max\{\|\ddot{c}\|_{\infty}, \|\dddot{c}\|_{\infty}\},$$

and $d(t) = -Nt^2(t + 1)$. Let e denote the n_1-dimensional vector of all ones. We now define

$$\tilde{c}(t) = c(t) + d(t)e,$$

for $t \in [0, T]$. Now by the definition of N, $\tilde{c}$ has nonpositive second and third derivatives. Finally, let M be such that $\|e^T x\|_{\infty} \leq M$ for all feasible solutions $\omega^T = (x^T, y^T, z^T)$ for SCLP. Consider now the following separated continuous linear program:

$$\begin{aligned}
&\text{SCLP}(M, N)\text{:} && \text{minimize} && \int_0^T \left[\tilde{c}(t)^T x(t) + d(t) x_{n_1+1}(t)\right] dt \\
& && \text{subject to} && \int_0^t G x(s)\, ds + y(t) = a(t), \\
& && && H x(t) + z(t) = b(t), \\
&(9) && && e^T x(t) + x_{n_1+1}(t) = M, \\
& && && x(t), y(t), z(t), w(t) \geq 0, \qquad t \in [0, T].
\end{aligned}$$

Strictly speaking, this is not in the form of a separated continuous linear program as (9) is an equality. However, this problem can be removed by replacing (9) by two inequalities.

Now by the definition of M, $\omega(t)^T = (x(t)^T, y(t)^T, z(t)^T)$ is feasible for SCLP if and only if $\tilde{\omega}(t)^T = (x(t)^T, M - e^T x(t), y(t)^T, z(t)^T)$ is feasible for SCLP(M,N). But

$$\int_0^T \left[\tilde{c}(t)^T x(t) + d(t)(M - e^T x(t))\right] dt = M \int_0^T d(t)\, dt + \int_0^T c(t)^T x(t)\, dt,$$

and so $\omega(t)$ is optimal for SCLP if and only if $\tilde{\omega}(t)$ is optimal for SCLP(M,N). But by Lemma 3.2, SCLP(M,N) has a piecewise constant optimal extreme point solution and hence so does SCLP. It is also clear that Lemma 3.2 gives the required bound on the number of breakpoints of $x(t)$ on $[0,T]$.

If we repeat the above argument a finite number of times, it is now not difficult to extend the result to the case of problem data with breakpoints. We omit the details as there is little to be gained from their inclusion. □

4. SCLP with analytic right-hand sides.

In this section we consider the case of SCLP with $a(t)$, $b(t)$, and $c(t)$ piecewise analytic (with of course $a(t)$ continuous). Under the assumption that the feasible region for SCLP is nonempty and bounded, we show that there exists an optimal extreme point solution with $x(t)$ piecewise analytic (Theorem 4.3). Although the result in the previous section, Theorem 3.3, is just a special case of Theorem 4.3, it is convenient to separate the two and use Theorem 3.3 as a starting point. In particular, we approximate $a(t)$ and $b(t)$ by sequences of piecewise linear and piecewise constant functions, respectively, and use Theorem 3.3 on these approximating problems to construct an optimal solution for the original problem. The difficulty is in constructing appropriate approximations to $a(t)$ and $b(t)$.

To perform this construction we need to introduce a definition that allows us to distinguish between two types of breakpoints in an extreme point solution for SCLP when the problem data have breakpoints. We recall that an extreme point solution for SCLP can only take the values of a finite number of functions $x^{(1)}(t), \dots, x^{(L)}(t)$ (Lemma 2.1). Moreover each $x^{(1)}(t), \dots, x^{(L)}(t)$ is given by a linear combination of the problem data and their derivatives. Thus $x^{(1)}(t), \dots, x^{(L)}(t)$ will have breakpoints when the problem data have breakpoints. Hence an extreme point solution will, in general, have breakpoints that arise from these breakpoints and also from switching from $x^{(i)}(t)$ to $x^{(j)}(t)$ for some $i \neq j$. It is this second type of breakpoint which is of importance and that we now define formally.

DEFINITION 4. *Let $\omega(t)$ be a piecewise analytic extreme point solution for* SCLP *and $x^{(1)}(t), \dots, x^{(L)}(t)$ be given by Lemma* 2.1. *By a* change of basis *we mean a time $s \in (0,T)$ such that for some $\varepsilon > 0$, $x(t) = x^{(i)}(t)$ for $t \in (s-\varepsilon, s)$ and $x(t) = x^{(j)}(t)$ for $t \in (s, s+\varepsilon)$ for some i and j such that $x^{(i)}(t) \not\equiv x^{(j)}(t)$ on $(s-\varepsilon, s+\varepsilon)$.*

We note that if a and b are analytic on a neighbourhood of $[0,T]$, then a change of basis is identical to a breakpoint. However, if a and b are piecewise analytic on $[0,T]$, then, in general, the set of changes of basis for an extreme point solution will be a subset of the set of breakpoints for that solution.

Using the above definition we may now give more details about the development of Theorem 4.3. We begin by taking sequences $\{a^{(n)}\}_{n=1}^{\infty}$ and $\{b^{(n)}\}_{n=1}^{\infty}$ of piecewise linear (and continuous) and piecewise constant functions, respectively, so that $a^{(n)}(t) \geq a(t)$ and $b^{(n)}(t) \geq b(t)$ on $[0,T]$ and $a^{(n)} \to a$ and $b^{(n)} \to b$ uniformly on $[0,T]$. This ensures that if SCLP$_n$ is the problem SCLP with a and b replaced by $a^{(n)}$ and $b^{(n)}$, respectively, then feasibility of SCLP implies feasibility of SCLP$_n$. Moreover, we also choose $a^{(n)}$ and $b^{(n)}$ so that SCLP$_n$ has an optimal extreme point solution with a bounded number of changes of basis independent of n. This is the difficult part of the result. It is then relatively easy to construct a piecewise analytic optimal solution for the original SCLP, in a similar way to the proof of the second half of Lemma 3.2.

To establish that the problem SCLP$_n$ mentioned above will have an optimal extreme point solution with a uniform bound on the number of changes of basis, we will need to use ideas from Anderson and Philpott [3]. To use these ideas it is necessary to have the set $\{\, \xi : H\xi \leq b(t),\ \xi \geq 0 \,\}$ bounded for each $t \in [0,T]$. To guarantee this, we begin by considering the case where there are upper bounds on $x(t)$. This extra assumption is removed later. As with the previous section, it is also convenient to begin by considering the case where the problem data contain no breakpoints. We therefore introduce the following assumption that we will use throughout this section.

Assumption 4.1. The costs, $c(t)$, the right-hand sides, $a(t)$ and $b(t)$, are analytic on a neighbourhood of $[0,T]$; H is of the form

$$H = \begin{bmatrix} \bar{H} \\ I \end{bmatrix};$$

and the feasible region for SCLP is nonempty.

We now develop some ideas and notation similar to those Anderson and Philpott [3]. Let

$$\bar{K} = \begin{bmatrix} G & I & -I & 0 \\ H & 0 & 0 & I \end{bmatrix}.$$

We note that $\bar{K}$ has full rank, and so any $n_2 + n_3$ linearly independent columns of $\bar{K}$ form a basis matrix.

Now suppose that $a(t)$ is piecewise analytic and continuous and that $b(t)$ is piecewise analytic on $[0,T]$. Let $S = [\alpha,\beta] \subseteq [0,T]$ satisfy the following: if B is any basis matrix for $\bar{K}$ and

$$\rho(t) = B^{-1} \begin{bmatrix} \dot{a}(t) \\ b(t) \end{bmatrix},$$

then for each i, either $\rho_i(t) > 0$ for all $t \in S$, $\rho_i(t) < 0$ for all $t \in S$, or $\rho_i(t) = 0$ for all $t \in S$. We now define

$$\hat{a}_S = \frac{1}{\beta - \alpha}(a(\beta) - a(\alpha)),$$
$$\hat{b}_S = \frac{1}{\beta - \alpha}\int_\alpha^\beta b(t)\,dt.$$

Now given any $\pi \in \{-1,0,1\}^{n_2}$ we define

$$\hat{E}_S = \{\, \hat{\psi} : H\hat{\psi} \leq \hat{b}_S,\ \hat{\psi} \geq 0,\ \operatorname{sgn}(\hat{a}_S - G\hat{\psi}) = \pi \,\},$$
$$E_S(t) = \{\, \psi : H\psi \leq b(t),\ \psi \geq 0,\ \operatorname{sgn}(\dot{a}(t) - G\psi) = \pi \,\}, \quad t \in S.$$

Strictly speaking, $\hat{E}_S$ and $E_S(t)$ depend on π; however, we choose to omit this dependence in the notation for simplicity. The following straightforward lemma is a simple extension of results in Anderson and Philpott [3, Lems. 2.3–2.5]. As K is a submatrix of $\bar{K}$, we extend the notations Definition 3 (i.e., of x_B and ρ_x) in a natural way to include basis matrices of $\bar{K}$.

LEMMA 4.1. *Let S be as above and $\hat{\psi}$ be any extreme point of $\hat{E}_S$. Then there exists a basis matrix B of $\bar{K}$ such that the nonzero components of $\hat{\psi}$ are contained in $\hat{\psi}_B$ and $\hat{\psi}_B = \hat{\rho}_x$, where*

$$\hat{\rho} = B^{-1} \left[\begin{array}{c} \hat{a}_S \\ \hat{b}_S \end{array} \right].$$

Moreover, let

$$\rho(t) = B^{-1} \left[\begin{array}{c} \dot{a}(t) \\ b(t) \end{array} \right]$$

and define $\psi(t)$ by $\psi_B(t) = \rho_x(t)$ with the other components of $\psi(t)$ set to zero for $t \in S$. Then

$$\int_\alpha^\beta \psi(t)\, dt = \int_\alpha^\beta \hat{\psi}\, dt$$

and $\psi(t)$ is an extreme point for $E_S(t)$ for $t \in S$.

For any piecewise analytic and continuous $a(t)$ and piecewise analytic $b(t)$ it is easy to construct a partition P of $[0,T]$ so that each subinterval satisfies the requirements for an interval S in the above lemma. Indeed, let

$$Q = \left\{ B^{-1} \left[\begin{array}{c} \dot{a}(t) \\ b(t) \end{array} \right] : B \text{ is a basis matrix of } \bar{K} \right\},$$
$$R = \{ \rho_j(t) : \rho \in Q,\ 1 \le j \le n_2 + n_3 \}.$$

Now since each component of $\dot{a}(t)$ and $b(t)$ is piecewise analytic on $[0,T]$, R consists of a finite set of piecewise analytic functions. Thus if $a(t)$ and $b(t)$ are analytic on a neighbourhood of some interval I, then each function in R is either identically zero or contains a finite number of zeros on I. We can now let P be the (finite) partition of $[0,T]$ that contains all the zeros of each function in R that is not identically zero on some interval, along with the breakpoints of $a(t)$ and $b(t)$. We call this the *canonical partition* of $[0,T]$. It was for the subintervals of this partition that the above lemma was proved in [3].

We may now begin the development of Theorem 4.3 by constructing appropriate functions $a^{(n)}$ and $b^{(n)}$ to approximate a and b. This construction is rather complicated although it is based on a relatively simple idea.

LEMMA 4.2. *Suppose that Assumption* 4.1 *holds. Then there exists $M \in \mathbb{N}$, a sequence $\{a^{(n)}\}_{n=1}^{\infty}$ of piecewise linear and continuous functions and a sequence $\{b^{(n)}\}_{n=1}^{\infty}$ of piecewise constant functions with $a^{(n)}(t) \ge a(t)$ and $b^{(n)}(t) \ge b(t)$ on $[0,T]$ for each n, $a^{(n)} \to a$ and $b^{(n)} \to b$ uniformly on $[0,T]$, and such that SCLP$_n$ has a piecewise constant optimal extreme point solution with at most M changes of basis, where SCLP$_n$ is the problem SCLP with right-hand sides $a^{(n)}$ and $b^{(n)}$.*

Proof. Let $B^{(1)}, \ldots, B^{(L)}$ be the possible basis matrices for $\bar{K}$. The construction of the $a^{(n)}$ and $b^{(n)}$ is done in two stages. Let

$$d(t) = \left[\begin{array}{c} \dot{a}(t) \\ b(t) \end{array} \right].$$

First we consider approximating $d(t)$ by $d(t) + 1/n\,e$, where e is the vector of all ones. We now define

$$\rho^{(n_j)}(t) = B^{(j)^{-1}} d(t) + \frac{1}{n} B^{(j)^{-1}} e.$$

Note that $\rho^{(n_j)}$ is analytic on a neighbourhood of $[0, T]$. This completes the first stage of the construction.

Second, we approximate $d(t) + 1/n\,e$ by a piecewise constant function

$$d^{(n)}(t) = \begin{bmatrix} \dot{a}^{(n)}(t) \\ b^{(n)}(t) \end{bmatrix},$$

with the approximation being done so that $a^{(n)}(t) \geq a(t)$ and $b^{(n)}(t) \geq b(t)$ on $[0, T]$ for each n and $a^{(n)} \to a$ and $b^{(n)} \to b$ uniformly on $[0, T]$, where $a^{(n)}$ is the integral of $\dot{a}^{(n)}$ taking the value $a(0) + 1/n\,e$ at zero. This second approximation is achieved by taking a sufficiently fine partition, $P_n = \{t_0^{(n)}, t_1^{(n)}, \ldots, t_{m_n}^{(n)}\}$, and setting $d^{(n)}(t) = d(t_{i-1}^{(n)}) + 1/n\,e$ for $t \in [t_{i-1}^{(n)}, t_i^{(n)})$, $i = 1, \ldots, m_n$.

Now by Lemma 2.3 there exists M_1, independent of n, so that if

$$\mathcal{F}_n = \left\{ t \in [0, T] : \rho_i^{(n_j)}(t) = 0 \text{ for some } i \text{ and } j \text{ but } \rho_i^{(n_j)} \not\equiv 0 \right\},$$

then $|\mathcal{F}_n| \leq M_1$. Again by Lemma 2.3 we may choose M_2 so that if

$$\mathcal{G}_n = \Big\{ t \in [0, T] : \dot{c}_{B^{(j)}}(t)^T \rho_x^{(n_j)}(t) = \dot{c}_{B^{(k)}}(t)^T \rho_x^{(n_k)}(t) \text{ for some } k \text{ and } j \text{ but}$$
$$\dot{c}_{B^{(j)}}^T \rho_x^{(n_j)} \not\equiv \dot{c}_{B^{(k)}}^T \rho_x^{(n_k)} \Big\},$$

then $|\mathcal{G}_n| \leq M_2$. Define

$$N = M_1 + M_2 + 1.$$

Choose n. Let $R_n = \{r_0^{(n)}, r_1^{(n)}, \ldots, r_{N_n}^{(n)}\}$ be the partition of $[0, T]$ that contains all the points in $\mathcal{F}_n$ and $\mathcal{G}_n$. By the definition of N, we have $N_n \leq N$. Such a partition R_n has been chosen for two reasons. First any subinterval $[r_{h-1}^{(n)}, r_h^{(n)}]$ of R_n satisfies the conditions required by an interval $[\alpha, \beta]$ in Lemma 4.1. Second any subinterval $[r_{h-1}^{(n)}, r_h^{(n)}]$ also satisfies the conditions of the form required by an interval $[\alpha, \beta]$ in Theorem 3.3 to bound the number of breakpoints in an optimal solution for SCLP if $\dot{a}$ and b are piecewise constant with no breakpoints in $(r_{h-1}^{(n)}, r_h^{(n)})$.

Having constructed a partition fine enough so that Theorem 3.3 and Lemma 4.1 can be applied when necessary on the subintervals, we now proceed to subdivide it further so that approximations can be made to $d + 1/n\,e$ so that $a^{(n)}(t) \geq a(t)$ and $b^{(n)}(t) \geq b(t)$ on $[0, T]$ for each n. Now by the uniform continuity of $d(t)$ on $[0, T]$, there exists δ_n so that $0 < 2\delta_n < \|R_n\|$ (where $\|R_n\|$, the norm of R_n, is the maximum distance between consecutive points of R_n) and such that

$$(10) \qquad |t - s| \leq \delta_n \Rightarrow \|d(t) - d(s)\|_\infty < \min\left\{ \frac{1}{nT}, \frac{1}{n} \right\}.$$

This δ_n gives the required upper bound on the norm of the partition P_n needed to form $a^{(n)}(t) \geq a(t)$ and $b^{(n)}(t) \geq b(t)$ on $[0, T]$.

We now require one last bound on the size of the partition P_n. This bound will ensure that an optimal solution for the problem SCLP$_n$ has a uniform bound on the number of changes of basis. Define

$$\varepsilon_n = \min\left\{ \left|\dot{c}_{B(j)}(t)^T \rho_x^{(n_j)}(t) - \dot{c}_{B(k)}(t)^T \rho_x^{(n_k)}(t)\right| : t \in [r_{h-1}^{(n)} + \delta_n, r_h^{(n)} - \delta_n] \text{ for } \right.$$
$$\left. \text{some } h \text{ and } \dot{c}_{B(j)}^T \rho_x^{(n_j)} \not\equiv \dot{c}_{B(k)}^T \rho_x^{(n_k)} \right\}.$$

Note that by the construction of R_n we have $\varepsilon_n > 0$. Again by the uniform continuity of $\rho^{(n_j)}$ and $\dot{c}$ there exists $\zeta_n > 0$ such that

$$\left|\dot{c}_{B(j)}(\tau_1)^T \rho_x^{(n_j)}(\tau_2) - \dot{c}_{B(k)}(\tau_1)^T \rho_x^{(n_k)}(\tau_3)\right| > \frac{\varepsilon_n}{2} \tag{11}$$

and

$$\left|\dot{c}_{B(k)}(\tau_1)^T \rho_x^{(n_k)}(\tau_2) - \dot{c}_{B(k)}(\tau_1)^T \rho_x^{(n_k)}(\tau_3)\right| < \frac{\varepsilon_n}{2} \tag{12}$$

for all $\tau_1, \tau_2, \tau_3 \in [r_{h-1}^{(n)} + \delta_n, r_h^{(n)} - \delta_n]$ for some h such that $|\tau_1 - \tau_2| \leq \zeta_n$ and $|\tau_1 - \tau_3| \leq \zeta_n$ and for all n_j and n_k such that $\dot{c}_{B(j)}^T \rho_x^{(n_j)} \not\equiv \dot{c}_{B(k)}^T \rho_x^{(n_k)}$.

We now define a partition P_n of $[0, T]$ with the following properties.

1. $P_n \cap [r_{h-1}^{(n)}, r_{h-1}^{(n)} + \delta_n) = \{r_{h-1}^{(n)}\}$ and $P_n \cap [r_h^{(n)} - \delta_n, r_h^{(n)}) = \{r_h^{(n)} - \delta_n\}$ for $h = 1, \ldots, N_n$.
2. $P_n \cap [r_{h-1}^{(n)} + \delta_n, r_h^{(n)} - \delta_n]$ is a partition of $[r_{h-1}^{(n)} + \delta_n, r_h^{(n)} - \delta_n]$ with norm no greater than $\min\{\delta_n, \zeta_n\}$ for $h = 1, \ldots, N_n$.

We are now ready to define $a^{(n)}$ and $b^{(n)}$. Suppose that $P_n = \{t_0^{(n)}, t_1^{(n)}, \ldots, t_{m_n}^{(n)}\}$. Define $a^{(n)}(0) = a(0) + 1/n\, e$ for $j = 1, \ldots, n_2$ and then define $d^{(n)}(t)$ by

$$d^{(n)}(t) \equiv \begin{bmatrix} \dot{a}^{(n)}(t) \\ b^{(n)}(t) \end{bmatrix} = \begin{bmatrix} \dot{a}(t_{i-1}^{(n)}) \\ b(t_{i-1}^{(n)}) \end{bmatrix} + \frac{1}{n} e \equiv d(t_{i-1}^{(n)}) + \frac{1}{n} e$$

for $t \in [t_{i-1}^{(n)}, t_i^{(n)})$. Thus

$$a^{(n)}(t) = a(0) + \frac{1}{n} e + \int_0^t \dot{a}^{(n)}(s)\, ds$$

for $j = 1, \ldots, n_2$. This defines a piecewise linear and continuous function $a^{(n)}$ and a piecewise constant function $b^{(n)}$ for each n. Moreover, by the construction of $a^{(n)}$ and $b^{(n)}$, in particular by (10), we have $a^{(n)}(t) \geq a(t)$ and $b^{(n)}(t) \geq b(t)$ on $[0, T]$ for each n with $a^{(n)} \to a$ and $b^{(n)} \to b$ uniformly on $[0, T]$, and hence the problem SCLP$_n$ with right-hand sides of $a^{(n)}$ and $b^{(n)}$ is feasible. We finally note that due to the structure of the matrix H, the feasible region for SCLP$_n$ is bounded.

We now claim that a piecewise constant optimal extreme point solution for SCLP$_n$ exists with at most $3N_n L$ changes of basis, where L is the number of basis matrices for K (as opposed to $\bar{L}$ for $\bar{K}$). As $N_n \leq N$ for each n, we can then set $M = 3NL$, and this will then establish the result.

Fix n and let $x^{(1)}(t), \ldots, x^{(L)}(t)$ be the possible values of an extreme point solution for SCLP$_n$ as given by Lemma 2.1. Strictly speaking, $x^{(1)}(t), \ldots, x^{(L)}(t)$ depend on n, but as we will now only be considering the one problem, SCLP$_n$, we choose to omit this dependence in the notation for simplicity. Now for each j there exists i such that the nonzero components of $x^{(j)}(t)$ are contained in $x^{(j)}_{B^{(i)}}(t)$ and $x^{(j)}_{B^{(i)}}(t) = \rho_x^{(n_i)}(t^{(n)}_{l-1})$ for $t \in [t^{(n)}_{l-1}, t^{(n)}_l)$, $l = 1, \ldots, m_n$. Hence $x^{(j)}(t)$ is piecewise constant with breakpoints in P_n for each j.

Consider the intervals $[r^{(n)}_{h-1} - \delta_n, r^{(n)}_{h-1})$ and $[r^{(n)}_h, r^{(n)}_h + \delta_n)$ for some h. Now $\dot{a}^{(n)}$ and $b^{(n)}$ are constant on these intervals, and so by Theorem 3.3 an optimal solution can be chosen for SCLP$_n$ such that the number of breakpoints, and hence the number of changes of basis, in each of $[r^{(n)}_{h-1} - \delta_n, r^{(n)}_{h-1})$ and $[r^{(n)}_h, r^{(n)}_h + \delta_n)$ is at most L. We now establish that such an optimal solution can be chosen so that for each h, the number of changes of basis in $[r^{(n)}_{h-1} + \delta_n, r^{(n)}_h - \delta_n)$ is at most L. This will establish the result. Note that we cannot expect $\omega(t)$ to have only L breakpoints on $[r^{(n)}_{h-1} + \delta_n, r^{(n)}_h - \delta_n)$ as the problem data $a^{(n)}$ and $b^{(n)}$ themselves may have more than L breakpoints on this interval.

Now by the assumption on the partition R_n, for each k and j we have either $\dot{c}(t)^T x^{(k)}(t) \neq \dot{c}(t)^T x^{(j)}(t)$ for all $t \in (r^{(n)}_{h-1}, r^{(n)}_h)$ or $\dot{c}(t)^T x^{(k)}(t) = \dot{c}(t)^T x^{(j)}(t)$ for all $t \in (r^{(n)}_{h-1}, r^{(n)}_h)$. We assume that $\dot{c}(t)^T x^{(k)}(t) \neq \dot{c}(t)^T x^{(j)}(t)$ for all $t \in (r^{(n)}_{h-1}, r^{(n)}_h)$ and for all $k \neq j$. The case when $\dot{c}(t)^T x^{(k)}(t) = \dot{c}(t)^T x^{(j)}(t)$ for all $t \in (r^{(n)}_{h-1}, r^{(n)}_h)$ for some $k \neq j$ can be treated in a way identical to Lemma 3.2 by taking a sequence $c^{(m)} \to c$ uniformly so that $\dot{c}^{(m)}(t)^T x^{(k)}(t) \neq \dot{c}^{(m)}(t)^T x^{(j)}(t)$ for all $t \in (r^{(n)}_{h-1}, r^{(n)}_h)$ for all $k \neq j$ and for all m. Having made this assumption we now note that from (11) and the definition of P_n we have

$$\left|\dot{c}(s)^T (x^{(k)}(s-) - x^{(j)}(s+))\right| > \frac{\varepsilon_n}{2}$$

for all $k \neq j$ and $s \in (r^{(n)}_{h-1} + \delta_n, r^{(n)}_h - \delta_n)$.

Suppose that $P_n \cap [r^{(n)}_{h-1} + \delta_n, r^{(n)}_h - \delta_n] = \{s_0, s_1, \ldots, s_\tau\}$. Again by Theorem 3.3 we know that an optimal solution, $\omega(t)$, for SCLP$_n$ can be chosen such that the number of breakpoints in $[s_{i-1}, s_i)$ is at most L (and with the conditions on the number of breakpoints in $[r^{(n)}_h - \delta_n, r^{(n)}_h + \delta_n)$ specified above). This is not sufficient for our purposes since bounding the number of breakpoints in $[s_{i-1}, s_i)$ will not give a bound independent of n. For such a bound we prove the following claim. If $[u, v] \subseteq [r^{(n)}_{h-1} + \delta_n, r^{(n)}_h - \delta_n]$ and

$$x(t) = \begin{cases} x^{(k)}(t), & t \in [u, s), \\ x^{(j)}(t), & t \in [s, v), \end{cases}$$

then $\dot{c}(s)^T x^{(k)}(s-) > \dot{c}(s)^T x^{(j)}(s+)$. This will show that $\omega(t)$ must have at most L changes of basis in $[r^{(n)}_{h-1} + \delta_n, r^{(n)}_h - \delta_n)$ (note that $x(t)$ will still have breakpoints at each s_i).

Suppose that the result is not true; i.e., for some $u, v, s \in [r^{(n)}_{h-1} + \delta_n, r^{(n)}_h - \delta_n]$ we have x defined as above but $\dot{c}(s)^T x^{(k)}(s-) < \dot{c}(s)^T x^{(j)}(s+)$. By the above, we now have

$$\dot{c}(s)^T (x^{(k)}(s-) - x^{(j)}(s+)) < -\frac{\varepsilon_n}{2}. \tag{13}$$

We will construct a new feasible solution with strictly improved objective function, thus contradicting the optimality of $\omega(t)$. We assume without loss of generality that $[u, v]$ is chosen small enough so that either $s = s_i$ for some i or $P_n \cap [u, v] = \emptyset$.

Let γ be such that $0 < \gamma \leq \min\{s-u, v-s\}$. Define

$$\hat{x}_\gamma = \frac{x^{(k)}(\tau_1) + x^{(j)}(\tau_2)}{2}$$

for $\tau_1 \in [s-\gamma, s)$, $\tau_2 \in [s, s+\gamma]$. Note that $\hat{x}_\gamma$ does not depend on τ_1 or τ_2. Let $S = [s-\gamma, s+\gamma]$ and let $\hat{a}_S$ and $\hat{b}_S$ be given by

$$\hat{a}_S = \frac{1}{2\gamma}(a^{(n)}(s+\gamma) - a^{(n)}(s-\gamma)),$$
$$\hat{b}_S = \frac{1}{2\gamma}\int_{s-\gamma}^{s+\gamma} b^{(n)}(t)\,dt.$$

Let $\pi = \operatorname{sgn}(\hat{a}_S - G\hat{x}_\gamma)$ and define $\hat{E}_S$ and $E_S(t)$ for $t \in S$ by

$$\hat{E}_S = \{\,\hat{\psi} : H\hat{\psi} \leq \hat{b}_S,\ \hat{\psi} \geq 0,\ \operatorname{sgn}(\hat{a}_S - G\hat{\psi}) = \pi\,\},$$
$$E_S(t) = \{\,\psi : H\psi \leq b^{(n)}(t),\ \psi \geq 0,\ \operatorname{sgn}(\dot{a}^{(n)}(t) - G\psi) = \pi\,\}.$$

Note that $\hat{E}_S$ is a convex set and, due to the structure of the matrix H, also a compact set. Let $\{\,\hat{\psi}^{(j)} : j = 1, \ldots, \bar{L}\,\}$ be the extreme points of $\hat{E}_S$; then by Minkowski's theorem (see, for example, Rockafellar [20]), there exists $\mu_j \geq 0$, $j = 1, \ldots, \bar{L}$ with

$$\sum_{j=1}^{\bar{L}} \mu_j = 1$$

and

$$\hat{x}_\gamma = \sum_{j=1}^{\bar{L}} \mu_j \hat{\psi}^{(j)}.$$

Now $S \subseteq [r_{h-1}^{(n)}, r_h^{(n)}]$, and so by a remark above, S satisfies the requirements for an interval in Lemma 4.1. Hence by this lemma, there exists $\psi^{(j)}(t)$, extreme points for $E_S(t)$, $t \in S$, that are piecewise constant (with a possible breakpoint at s only) such that

$$\int_{s-\gamma}^{s+\gamma} \psi^{(j)}(t)\,dt = \int_{s-\gamma}^{s+\gamma} \hat{\psi}^{(j)}\,dt,$$

or in this case,

$$\gamma(\psi^{(j)}(s-) + \psi^{(j)}(s+)) = 2\gamma\hat{\psi}^{(j)}. \tag{14}$$

Moreover, there exists k such that the nonzero components of $\psi^{(j)}(t)$ are contained in $\psi_{B(k)}^{(j)}(t)$ and either $\psi_{B(k)}^{(j)}(t) = \rho_x^{(n_k)}(s_{i-1})$ for $t \in [s-\gamma, s)$ and $\psi_{B(k)}^{(j)}(t) = \rho_x^{(n_k)}(s)$ for $t \in [s, s+\gamma]$ if $s = s_i$ for some i, or $\psi_{B(k)}^{(j)}(t) = \rho_x^{(n_k)}(s_i)$ for $t \in [s-\gamma, s+\gamma]$ for some i, if $P_n \cap [u,v] = \emptyset$. Define

$$\bar{x}_\gamma(t) = \begin{cases} \displaystyle\sum_{j=1}^{\bar{L}} \mu_j \psi^{(j)}(t), & t \in [s-\gamma, s+\gamma), \\ x(t) & \text{otherwise.} \end{cases}$$

Note that $\bar{x}_\gamma(t)$ is no more than a vector of components of a convex combination of $\rho^{(n_j)}(t)$ for $j = 1, \ldots, \bar{L}$. Hence $\bar{x}_\gamma(t)$ is piecewise constant on $[s-\gamma, s+\gamma)$, with a possible breakpoint at s only. Also from (12) and the definition of P_n we have

$$\dot{c}(s)^T(\bar{x}_\gamma(s+) - \bar{x}_\gamma(s-)) < \frac{\varepsilon_n}{2}. \tag{15}$$

Define $\bar{\omega}_\gamma(t)$ from the constraints of SCLP$_n$, i.e., so that $\bar{x}_\gamma(t)$, $\bar{y}_\gamma(t)$, and $\bar{z}_\gamma(t)$ satisfy (1) and (2) for SCLP$_n$ in place of $x(t)$, $y(t)$, and $z(t)$, respectively. Then $\bar{\omega}_\gamma(t)$ is feasible for SCLP$_n$. The proof of this argument is given in Anderson and Philpott [3], of which we present a brief description. Clearly we have $\bar{x}_\gamma(t), \bar{z}_\gamma(t) \geq 0$. By the construction of $\bar{x}_\gamma(t)$ we have $\bar{y}_\gamma(t) = y(t) \geq 0$ for $t \in [0, s-\gamma] \cup [s+\gamma, T]$. However, by the construction of $E_S(t)$, $\operatorname{sgn}(\dot{a}(t) - G\bar{x}_\gamma(t))$ is constant for $t \in (u, v)$. Hence each component of $\bar{y}_\gamma(t)$ is monotonic on $[s-\gamma, s+\gamma]$ and thus we have $(\bar{y}_\gamma)_i(t)$ between $(\bar{y}_\gamma)_i(s-\gamma)$ and $(\bar{y}_\gamma)_i(s+\gamma)$ for each i and $t \in (s-\gamma, s+\gamma)$. Thus $\bar{y}_\gamma(t) \geq 0$ for $t \in (s-\gamma, s+\gamma)$ as well and so $\bar{\omega}_\gamma(t)$ is feasible for SCLP.

We now claim that for some $\gamma > 0$ an improvement in the objective function can be made. For this purpose we use the standard notation $o(h^n)$ for $n \in \mathbb{N}$ to mean a function defined on an interval containing zero such that $\lim_{h \downarrow 0} o(h^n)/h^n = 0$. Now

$$\begin{aligned}
&\int_{s-\gamma}^{s} c(t)^T(\psi^{(j)}(t) - \hat{\psi}^{(j)})\,dt \\
&= \int_{s-\gamma}^{s} c(s)^T(\psi^{(j)}(s-) - \hat{\psi}^{(j)})\,dt + \int_{s-\gamma}^{s} (t-s)\dot{c}(s)^T(\psi^{(j)}(s-) - \hat{\psi}^{(j)})\,dt \\
&\quad + \int_{s-\gamma}^{s} o((t-s))\,dt \\
&= \gamma c(s)^T(\psi^{(j)}(s-) - \hat{\psi}^{(j)}) - \frac{\gamma^2}{2}\dot{c}(s)^T(\psi^{(j)}(s-) - \hat{\psi}^{(j)}) + o(\gamma^2).
\end{aligned}$$

Similarly,

$$\begin{aligned}
\int_{s}^{s+\gamma} c(t)^T(\psi^{(j)}(t) - \hat{\psi})\,dt = \gamma c(s)^T(\psi^{(j)}(s+) - \hat{\psi}^{(j)}) + \frac{\gamma^2}{2}\dot{c}(s)^T(\psi^{(j)}(s+) - \hat{\psi}^{(j)}) \\
+ o(\gamma^2).
\end{aligned}$$

Hence by (14) we obtain

$$\int_{s-\gamma}^{s+\gamma} c(t)^T(\psi^{(j)}(t) - \hat{\psi})\,dt = \frac{\gamma^2}{2}\dot{c}(s)^T(\psi^{(j)}(s+) - \psi^{(j)}(s-)) + o(\gamma^2),$$

thus giving

$$\int_0^T c(t)^T(\bar{x}_\gamma(t) - \hat{x}_\gamma)\,dt = \frac{\gamma^2}{2}\dot{c}(s)^T(\bar{x}_\gamma(s+) - \bar{x}_\gamma(s-)) + o(\gamma^2).$$

By a similar argument we now obtain

$$\begin{aligned}
&\int_0^T c(t)^T(\bar{x}_\gamma(t) - x(t))\,dt \\
&= \int_0^T c(t)^T(\bar{x}_\gamma(t) - \hat{x}_\gamma)\,dt + \int_0^T c(t)^T(\hat{x}_\gamma - x(t))\,dt \\
&= \frac{\gamma^2}{2}\dot{c}(s)^T\left(\bar{x}_\gamma(s+) - \bar{x}_\gamma(s-) + x^{(k)}(s-) - x^{(j)}(s+)\right) + o(\gamma^2).
\end{aligned}$$

Hence

$$\begin{aligned}
&\lim_{\gamma\downarrow 0}\frac{1}{\gamma^2}\int_0^T c(t)^T(\bar{x}_\gamma(t)-x(t))\,dt\\
&=\frac{1}{2}\dot{c}(s)^T\left(\bar{x}_\gamma(s+)-\bar{x}_\gamma(s-)+x^{(k)}(s-)-x^{(j)}(s+)\right)\\
&<\frac{\varepsilon_n}{4}-\frac{\varepsilon_n}{4}\\
&=0
\end{aligned}$$

by (13) and (15). Hence for some $0<\gamma\le\min\{s-u,v-s\}$ we have

$$\int_0^T c(t)^T\bar{x}_\gamma(t)\,dt<\int_0^T c(t)^T x(t)\,dt,$$

contradicting the optimality of $\omega(t)$. □

It is now a simple matter to construct a piecewise analytic optimal solution for SCLP under Assumption 4.1. Again if we repeat the above argument a finite number of times, it is now possible to extend this result to the case where the problem data are piecewise analytic. We may also weaken the assumption that there are upper bounds on $x(t)$ to the assumption that just the feasible region for SCLP is bounded to arrive at the main result of this paper.

THEOREM 4.3. *Suppose that the costs, $c(t)$, right-hand sides, $a(t)$ and $b(t)$, are piecewise analytic on $[0,T]$ (but with $a(t)$ continuous) and that the feasible region for SCLP is nonempty and bounded; then there exists an optimal extreme point solution for SCLP with $x(t)$ piecewise analytic on $[0,T]$.*

Proof. We first prove the result under Assumption 4.1. Let $\{a^{(n)}\}_{n=1}^{\infty}$ and $\{b^{(n)}\}_{n=1}^{\infty}$ be sequences of functions given by Lemma 4.2. Let SCLP$_n$ be the problem SCLP with right-hand sides $a^{(n)}$ and $b^{(n)}$, M be the uniform bound on the number of changes of basis in optimal solutions for SCLP$_n$, and $\omega^{(n)}$ be the optimal solution for SCLP$_n$ with no more than M changes of basis. Finally, let $P_n=\{t_0^{(n)},t_1^{(n)},\dots,t_{m_n}^{(n)}\}$ be the partition of $[0,T]$ containing the changes of basis for $\omega^{(n)}$ and $B(t_i^{(n)})$ be a basis matrix for K corresponding to the support of $\omega^{(n)}$ on $[t_{i-1}^{(n)},t_i^{(n)})$.

Now by introducing artificial changes of basis, if necessary, we can assume that $m_n=M$, i.e., that each $\omega^{(n)}$ has exactly M changes of basis. Now as $[0,T]$ is a bounded interval, there exists a subsequence $\{n_k\}_{k=1}^{\infty}$ and $t_0,t_1,\dots,t_M$ such that $t_i^{(n_k)}\to t_i$. Moreover, as there are only a finite number of choices for basis matrices for K, we may also assume that the subsequence $\{n_k\}_{k=1}^{\infty}$ is chosen so that for each i, $B(t_i^{(n)})=B^{(i)}$ for some basis matrix $B^{(i)}$ of K for $t\in[t_{i-1}^{(n_k)},t_i^{(n_k)})$. We now define the nonzero components of $x(t)$ by

$$x_{B^{(i)}}(t)=\rho_x^{(i)}(t),\quad t\in[t_{i-1},t_i),$$

for $i=1,\dots,M$, where

$$\rho^{(i)}(t)=B^{(i)^{-1}}\begin{bmatrix}\dot{a}(t)\\ b(t)\end{bmatrix}.$$

We now have $x^{(n_k)}(t)\to x(t)$ a.e. on $[0,T]$. Hence by Lemma 2.2, $\omega(t)$, defined from $x(t)$ by the constraints of SCLP, is an optimal extreme point for SCLP.

As mentioned above, we may now repeat the above argument a finite number of times to extend the result to the case where the problem data are piecewise analytic. It is therefore now only required to prove the result when the feasible region for SCLP bounded, given that it is true for SCLP with upper bounds on $x(t)$. Suppose then that the feasible region is bounded. Let N be such that $\|x\|_\infty \leq N$ for all feasible solutions for SCLP. Let SCLP(N) be the problem SCLP with the extra constraints

$$\begin{aligned} x(t) + v(t) &= Ne, \\ v(t) &\geq 0, \end{aligned}$$

where e is the vector of all ones. Now SCLP(N) has an optimal extreme point solution, $\omega(t)^T = (x(t)^T, y(t)^T, z(t)^T, v(t)^T)$ that is piecewise analytic on $[0, T]$. Clearly $(x(t)^T, y(t)^T, z(t)^T)$ forms a piecewise analytic optimal extreme point solution for the original SCLP. □

As a closing remark we state a simple condition that must hold at a breakpoint. As its proof is essentially contained in Lemma 4.2, we omit the details.

THEOREM 4.4. *Suppose that the costs, $c(t)$, right-hand sides, $a(t)$ and $b(t)$, are piecewise analytic on $[0, T]$ (but with $a(t)$ continuous) and that the feasible region for SCLP is nonempty and bounded. Let $\omega(t)$ be a piecewise analytic optimal solution for SCLP. Let $[\alpha, \beta]$ be a subinterval of the canonical partition. Suppose that $\omega(t)$ has a breakpoint at $s \in [\alpha, \beta]$; i.e., for some functions $x^{(k)}(t)$ and $x^{(j)}(t)$ analytic on neighbourhoods of $[\alpha, \beta]$ we have*

$$x(t) = \begin{cases} x^{(k)}(t) & a.e. \text{ on } [u, s), \\ x^{(j)}(t) & a.e. \text{ on } [s, v) \end{cases}$$

for some u and v. Then $\dot{c}(s)^T x^{(k)}(s) \geq \dot{c}(s)^T x^{(j)}(s)$.

Note that under the restrictions of piecewise linear $a(t)$ and piecewise constant $b(t)$ of the previous section, the canonical partition is just the set of breakpoints of $a(t)$ and $b(t)$ and the points 0 and Tn.

5. Extensions and remarks. It is interesting to speculate on possible extensions to the results contained in the previous sections. One possible extension could be that the costs need not be analytic but only n-times differentiable for some n, or even just continuous. However, as the counterexample below shows, the result given in Theorem 3.3 cannot be extended beyond analytic costs.

Example 5.1. Let $c(t)$ be any continuous real-valued function defined on $[0, 1]$ such that if

$$S = \{ t \in [0, 1] : c(t) < 0 \},$$

then S is an open disconnected set containing an infinite number of components. For example we could take

$$c(t) = \begin{cases} t^{n+1} \sin 1/t, & t \in (0, 1], \\ 0, & t = 0, \end{cases}$$

for some n. Consider the following SCLP problem:

$$\text{minimize} \quad \int_0^1 c(t)x(t)\, dt$$

$$\begin{aligned} \text{subject to} \quad & \int_0^t x(s)\,ds + y(t) = 1, \\ & x(t) + z(t) = 1, \\ & x(t), y(t), z(t) \geq 0, \qquad t \in [0,1]. \end{aligned}$$

Clearly this has the following optimal solution:

$$x(t) = \begin{cases} 1 & \text{a.e. on } S, \\ 0 & \text{a.e. on } [0,1] - S. \end{cases}$$

By the definition of $c(t)$, $x(t)$ has an infinite number of breakpoints.

In a similar way we may not extend Theorem 4.3 to allow $a(t)$ or $b(t)$ chosen from classes of more general functions, provided that a breakpoint is understood to be a point where the optimal solution is not as well behaved as the problem data rather than just a discontinuity (i.e., if the problem data are n-times differentiable, then a breakpoint is a point where the solution is not n-times differentiable).

Example 5.2. Define $b_1(t)$ and $b_2(t)$ on $[0,1]$ as follows.

$$b_1(t) = \begin{cases} t^{n+1}\sin 1/t + 1, & t \in (0,1], \\ 1, & t = 0, \end{cases} \qquad b_2(t) = \begin{cases} t^{n+1}\cos 1/t + 1, & t \in (0,1], \\ 1, & t = 0, \end{cases}$$

for some n. Consider the following SCLP problem:

$$\begin{aligned} \text{minimize} \quad & -\int_0^1 x(t)\,dt \\ \text{subject to} \quad & \int_0^t x(s)\,ds + y(t) = 2, \\ & x(t) + z_1(t) = b_1(t), \\ & x(t) + z_2(t) = b_2(t), \\ & x(t), y(t), z_1(t), z_2(t) \geq 0, \qquad t \in [0,1]. \end{aligned}$$

Let $S = \{t \in [0,1] : t^{n+1}\cos 1/t > t^{n+1}\sin 1/t\}$. Clearly the given problem has the following optimal solution:

$$x(t) = \begin{cases} b_1(t) & \text{a.e. on } S, \\ b_2(t) & \text{a.e. on } [0,1] - S. \end{cases}$$

Using the definition of breakpoints above, $x(t)$ has an infinite number of breakpoints.

Note, however, that the optimal solution for the above problem is continuous. It is an open question whether continuously differentiable $a(t)$ and continuous $b(t)$ (and sufficiently well behaved costs) give rise to an optimal solution that is piecewise continuous.

It is also worth noting that the optimal solutions to the above two problems both have a *countable* number of breakpoints. It is possible that extensions to Theorem 4.3 (and Theorem 3.3) could be made if we allow an optimal solution to have a countable number instead of just a finite number of breakpoints.

A simple extension to the results in this paper comes from recalling that each of the functions $x^{(1)}(t), \ldots, x^{(L)}(t)$ in Lemma 2.1 lie in the linear space spanned by the components of $\dot{a}(t)$ and $b(t)$. However, for $\omega(t)$ to be a piecewise analytic extreme point solution for SCLP with piecewise analytic problem data, on each interval where $x(t)$

is analytic we must have $x(t) \equiv x^{(i)}(t)$ for some i. (This was seen in the construction of the optimal solution in Theorem 4.3.) Thus, if $\dot{a}(t)$ and $b(t)$ are in some linear subspace S of the vector space of functions piecewise analytic on $[0, T]$, then under the remaining conditions of Theorem 4.3, we obtain an optimal extreme point solution for SCLP that is piecewise analytic, and on intervals where $x(t)$ is analytic, $x(t)$ is in the linear subspace S. In particular the following result holds.

THEOREM 5.1. *Suppose that $c(t)$ is piecewise analytic on $[0, T]$ and that $a(t)$ and $b(t)$ are piecewise polynomials of degrees $n+1$ and n, respectively (with $a(t)$ continuous). Suppose furthermore that the feasible region for SCLP is nonempty and bounded, then there exists an optimal extreme point solution for SCLP with $x(t)$ piecewise polynomial of degree n on $[0, T]$.*

Acknowledgments. The author wishes to thank Eddie Anderson for his help in correcting an error in an earlier version of this paper. In particular, the general nature of the method used in Theorem 3.3 to transform a problem with general costs into one where the costs have nonpositive second and third derivatives was his idea.

REFERENCES

[1] E. J. ANDERSON, *A Continuous Model for Job-Shop Scheduling*, Ph.D. thesis, University of Cambridge, U.K., 1978.

[2] E. J. ANDERSON, P. NASH, AND A. F. PEROLD, *Some properties of a class of continuous linear programs*, SIAM J. Control Optim., 21 (1983), pp. 758–765.

[3] E. J. ANDERSON AND A. B. PHILPOTT, *On the solutions of a class of continuous linear programs*, SIAM J. Control Optim., 32 (1994), pp. 1289–1296.

[4] K. M. ANSTREICHER, *Generation of feasible descent directions in continuous time linear programming*, Tech. report SOL 83-18, Department of Operations Research, Stanford University, Stanford, CA, 1983.

[5] T. M. APOSTOL, *Mathematical Analysis*, 2nd ed., Addison-Wesley, Reading, MA, 1974.

[6] R. E. BELLMAN, *Bottleneck problems and dynamic programming*, Proc. Nat. Acad. Sci. U.S.A., 39 (1953), pp. 947–951.

[7] R. N. BUIE AND J. ABRHAM, *Numerical solutions to continuous linear programming problems*, Z. Oper. Res., 17 (1973), pp. 107–117.

[8] W. P. DREWS, *A simplex-like algorithm for continuous-time linear optimal control problems*, in Optimization Methods for Resource Allocation, R. W. Cottle and J. Krarup, eds., Crane Russak, New York, 1974, pp. 309–322.

[9] R. J. HARTBERGER, *Representation extended to continuous time*, in Optimization Methods for Resource Allocation, R. W. Cottle and J. Krarup, eds., Crane Russak, New York, 1974, pp. 297–307.

[10] A. E. ILYOTOVICH, *Piecewise-continuous solutions of linear dynamic problems in economic planning*. I, Automat. Remote Control, 41 (1980), pp. 501–508.

[11] J. JASIULEK, *Structural properties of solutions to continuous linear programs*, Technical report, Department of Mathematics, Simon Fraser University, Burnaby, B.C., 1981.

[12] B. JÓHANNESSON AND M. A. HANSON, *On the form of solutions to the linear continuous time programming problem and a conjecture by Tyndall*, J. Math. Anal. Appl., 111 (1985), pp. 236–242.

[13] M. KÖHLER, *Pointwise maximum principle for convex optimal control problems with mixed control-phase variable inequality constraints*, J. Optim. Theory Appl., 30 (1980), pp. 269–291.

[14] E. B. LEE AND L. MARKUS, *Foundations of Optimal Control Theory*, Wiley-Interscience, New York, 1967.

[15] R. S. LEHMAN, *On the continuous simplex method*, RM-1386, Rand Corporation, Santa Monica, CA, 1954.

[16] H. MAURER, *On optimal control problems with bounded state variables and control appearing linearly*, SIAM J. Control Optim., 15 (1977), pp. 345–362.

[17] A. F. PEROLD, *Fundamentals of a continuous time simplex method*, Tech. report SOL 78-26, Department of Operations Research, Stanford University, Stanford, CA, 1978.

[18] M. C. PULLAN, *Separated Continuous Linear Programs: Theory and Algorithms*, Ph.D. thesis, University of Cambridge, U.K., 1992.

[19] ———, *An algorithm for a class of continuous linear programs*, SIAM J. Control Optim., 31 (1993), pp. 1558–1577.

[20] R. T. ROCKAFELLAR, *Convex Analysis*, Princeton University Press, Princeton, NJ, 1970.

[21] W. RUDIN, *Real and Complex Analysis*, 3rd ed., McGraw-Hill, Singapore, 1986.

[22] R. G. SEGERS, *A generalised function setting for dynamic optimal control problems*, in Optimization Methods for Resource Allocation, R. W. Cottle and J. Krarup, eds., Crane Russak, New York, 1974, pp. 279–296.

[23] W. F. TYNDALL, *A duality theorem for a class of continuous linear programming problems*, SIAM J. Appl. Math., 13 (1965), pp. 644–666.